



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

**TF-IDF PARA LA OBTENCIÓN AUTOMÁTICA DE
TÉRMINOS Y SU VALIDACIÓN MEDIANTE
WIKIPEDIA**

TESIS

QUE PARA OBTENER EL TÍTULO DE:

INGENIERO EN COMPUTACIÓN

PRESENTA:

LUIS ADRIÁN CABRERA DIEGO



DIRECTOR DE TESIS:

Dr. Gerardo Sierra Martínez

Ciudad Universitaria, México D.F.

2011

“Elige un trabajo que te guste y no tendrás que trabajar ni un día de tu vida”

Confucio

AGRADECIMIENTOS

Quiero agradecer a todas las personas que han estado conmigo y que me han apoyado para llegar a este punto de mi vida, en especial:

A mi padre, Ismael Cabrera Mancilla, quien siempre me ha apoyado en todos los sentidos para llegar a ser la persona que ahora soy. A él siempre le agradeceré, creciendo como persona y superando siempre mis metas.

A mi madre, Gabriela Diego Casimiro, porque fue quien me mostró el fascinante mundo de la computación a muy temprana edad y siempre estuvo ahí para escucharme, ayudarme y darme ideas.

A mi hermana, Gaby, que aunque no me diera cuenta siempre estuvo conmigo.

Al Dr. Gerardo Sierra, mi director de tesis, por su ayuda, paciencia y apoyo a lo largo de la realización de la misma. Él siempre tuvo tiempo de escucharme y me dio grandes ideas para que este trabajo quedara lo mejor posible.

Al Dr. Jorge Vivaldi de la Universidad Pompeu Fabra, por apoyarme en todo momento y ayudarme a desarrollar una parte de este trabajo, compartiendo un poco de su conocimiento.

A mis sinodales, el Dr. Juan-Manuel Torres-Moreno, el Dr. Alfonso Medina Urrea, el M. en A. Miguel Eduardo Gonzáles Cárdenas y el M.L. Carlos Francisco Méndez Cruz quienes me apoyaron en todo momento en el trámite de titulación a pesar del tiempo tan corto para realizarlo.

A mis profesores de la Facultad de Ingeniería, quienes me permitieron aprender de ellos y siempre me apoyaron, no solamente para convertirme en un ingeniero en computación, sino también para crecer y ser una mejor persona.

A mis amigos, Juan Miguel Rolland Bartilotti y Miguel Mazkiarán Ramírez, a quienes conozco desde el primer día que entré a la Facultad de Ingeniería y quienes siempre han estado cerca de mí, para apoyarme y ayudarme sin pedir nada a cambio.

A mi amiga, Alejandra L. Soto Osorio, quien fue mi mano derecha durante gran parte de la carrera y que además me mostró y me encaminó al módulo de Tecnologías del Lenguaje.

A mis demás amigos y compañeros de la carrera, con quienes conviví gran parte de mi tiempo en la facultad ya sea tomando clases, esperando a algún profesor o platicando en la biblioteca.

A mis compañeros del Centro de Enseñanza de Lenguas Extranjeras (CELE), en especial mis amigas Sofía Hernández Gonzáles, Marysol Segovia Oropeza, Ana Mercedes Martínez Aguirre y Beatriz López Portillo, quienes convertían las clases de francés (y las salidas de éstas) en un momento donde uno se podía relajar, reír y olvidar de los problemas de la facultad al menos por unos minutos.

A todos los integrantes del Grupo de Ingeniería Lingüística (GIL) quienes siempre me han apoyado y han hecho ameno el tiempo que trabajamos juntos.

A Teresita Adriana Reyes Careaga y Alejandro Rosas Gonzáles quienes revisaron una parte de esta tesis e hicieron que no hubiera errores ortográficos y de redacción y al mismo tiempo permitieron que aprendiera algunos consejos para escribir mejor.

A mi amiga Sofía Fuentes Rodríguez, quien me ayudó a revisar las últimas partes escritas de esta tesis. Espero que el tú, que empiezas tu carrera universitaria, superes todas tus metas y encuentres excelentes amigos así como yo lo hice.

Finalmente, quiero agradecer los apoyos recibidos del Consejo Nacional de Ciencia y Tecnología (CONACyT) provenientes de los proyectos “Extracción de relaciones léxicas para dominios restringidos a partir de contextos definitorios en español” registro 82050 y “El vocabulario básico científico en México: Una investigación de sus características, componentes y difusión” registro 58923.

TABLA DE CONTENIDO

Introducción	1
Objetivo.....	2
Estructura de la tesis	2
1 Procesamiento de Lenguaje Natural	4
1.1 Recursos y herramientas empleadas en PLN	5
1.1.1 Corpus lingüísticos.....	5
1.1.2 Tokenizadores	7
1.1.3 N-gramas.....	9
1.1.4 Etiquetadores de partes de la oración.....	10
1.1.5 Lematizadores	12
1.1.6 Palabras funcionales.....	14
1.2 Recuperación de información.....	15
1.2.1 Term frequency – Inverse document frequency (TF-IDF)	16
1.2.2 Normalización de la longitud del documento	20
1.2.2.1 Normalización de coseno.....	21
1.2.2.2 Normalización por pivote	24
1.2.2.3 Normalización por máximo TF	26
1.2.3 Evaluación de sistemas de recuperación de información	27
1.2.4 Extracción y recuperación de información	29
2 Terminología.....	31

2.1	Terminología y terminografía	31
2.1.1	Los términos.....	32
2.1.2	La terminografía.....	33
2.1.3	Extracción de información terminológica.....	35
2.2	Sistemas actuales de extracción terminológica	36
2.2.1	Sistemas basados en conocimiento lingüístico	37
2.2.1.1	LEXTER.....	38
2.2.1.2	HEID.....	41
2.2.2	Sistemas basados en conocimiento estadístico	42
2.2.2.1	ANA.....	43
2.2.2.2	Extractor de términos estadístico basado en corpus	45
2.2.3	Sistemas basados en conocimiento híbrido.....	46
2.2.3.1	Termext.....	46
2.2.3.2	YATE.....	47
2.3	Evaluación de los extractores terminológicos	49
2.3.1	Lista de referencia.....	49
2.3.2	Validación.....	50
2.4	Recursos electrónicos para la validación	50
2.4.1	WordNet y EuroWordNet.....	51
2.4.2	Lexicón Specialist UMLS.....	52
2.4.3	Wikipedia.....	52

3	Obtención automática de términos y su validación	54
3.1	Corpus de textos científicos en español de México (COCIAM)	54
3.1.1	Estructura del COCIEM.....	55
3.2	Preprocesamiento del COCIEM.....	56
3.2.1	Revisión, limpieza y adecuación de los documentos.....	57
3.2.2	Lematización usando FreeLing.....	57
3.2.3	Tokenización del COCIEM	61
3.2.4	Creación de n-gramas	62
3.3	Extracción de candidatos a término	64
3.3.1	Cálculo de TF.....	65
3.3.2	Limpieza de los n-gramas generados.....	66
3.3.3	Cálculo de IDF, TF-IDF y su normalización.....	68
3.4	Validación de los candidatos a término.....	71
3.4.1	Wikipedia para la validación	72
3.4.1.1	Conversión a una base de datos.....	75
3.4.1.2	Lematización de Wikipedia.....	78
3.4.2	Cálculo del coeficiente de dominio.....	79
3.5	Arquitectura del sistema.....	84
4	Resultados y evaluación.....	87
4.1	Extracción de candidatos a término del COCIEM.....	87
4.2	Selección de los candidatos a término a validar.....	89

4.3	Obtención de términos validados por Wikipedia	93
4.4	Evaluación de resultados	101
4.4.1	Observaciones de la evaluación de resultados	105
5	Conclusión y trabajo a futuro.....	110
	Bibliografía	113
	Anexos	119
	Anexo A: Lista de palabras funcionales	120
	Anexo B: Gráficas de precisión contra cobertura de matemáticas de bachillerato.....	125
	Anexo C: Gráficas de precisión contra cobertura de ecología de bachillerato	131
	Anexo D: Gráficas de precisión contra cobertura de matemáticas de primaria.....	137
	Anexo E: Gráficas de precisión contra cobertura de matemáticas de bachillerato evaluada con la segunda lista de términos	143
	Anexo F: Lista de términos validados de matemáticas de bachillerato	145
	Anexo G: Lista de términos validados de ecología de bachillerato.....	157
	Anexo H: Lista de términos validados de matemáticas de primaria.....	160

INTRODUCCIÓN

En la actualidad la cantidad de información a la que se puede tener acceso es enorme, ésta puede provenir de libros, revistas, recursos electrónicos, entre otros. Estos medios de información ciertas veces incluyen términos nuevos o desconocidos que deben ser buscados para comprender de manera óptima la información. Sin embargo, a pesar de la importancia que tienen en la vida académica y cotidiana los términos, su obtención es una tarea complicada y costosa, tanto en recursos como en tiempo, que tradicionalmente se lleva a cabo de manera manual con la ayuda de expertos en terminología como del área de donde se quieran obtener los términos.

Por lo anterior, es que desde la década de 1990, y más recientemente desde el año 2000, se han desarrollado numerosos sistemas extractores de términos que emplean diversos conocimientos, ya sean lingüísticos, estadísticos o híbridos, para la obtención de los términos de un conjunto de documentos de manera automática. Esto para emplear los términos en la creación de diccionarios y glosarios, verbigracia, pero igualmente en el enriquecimiento de sistemas de traducción automática, sistemas de generación de texto, bases de conocimiento, etcétera.

Pero aún con el desarrollo de diversos sistemas de extracción terminológica, pocos de ellos emplean recursos léxicos, como los son Wikipedia o EuroWordNet, para llevar a cabo una validación de los candidatos a término que se obtienen de los extractores terminológicos. Siendo que estos tienen como ventaja el poder obtener listas de unidades terminológicas mucho más fiables y acordes al área de análisis; así como la inclusión de información extra que sólo se encuentra en los recursos léxicos, como la relación entre diversos términos, sinónimos, abreviaturas, etcétera.



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Objetivo

El objetivo general de esta tesis es desarrollar un sistema extractor de términos que extraiga la mayor cantidad de candidatos a término y que a su vez los resultados que se obtengan de éste se validen empleando la enciclopedia gratuita Wikipedia.

Asimismo, como un objetivo particular de esta tesis es la obtención de listas de términos validadas, es decir, un conjunto de unidades terminológicas propias de un corpus determinado o de una selección de él.

Estructura de la tesis

Esta tesis está conformada por cinco capítulos; los dos primeros capítulos son los antecedentes necesarios para la comprensión de la tesis. De manera más específica, el primer capítulo es sobre el procesamiento de lenguaje natural, donde se hablará sobre lo que es, algunos recursos y herramientas que emplea y una de sus aplicaciones, la recuperación de información; en cambio, en el segundo capítulo se abordará lo que es la terminología, no sólo como disciplina sino también como conjunto de términos, asimismo se darán a conocer algunos de los extractores terminológicos desarrollados en los últimos años, los métodos de evaluación de estos y algunos de los recursos léxicos más representativos que existen.

En el tercer capítulo se dará a conocer la metodología empleada en el desarrollo de esta tesis, es decir, el corpus que se empleó, el método de extracción terminológica seleccionado, el método de validación que se usó para validar los resultados del extractor de términos y la arquitectura del sistema desarrollado para la obtención de términos y su validación empleando Wikipedia.

A lo largo del cuarto capítulo se describirán los resultados obtenidos tanto del extractor terminológico como los obtenidos de la validación en donde se empleó la enciclopedia Wikipedia. Asimismo, se hablará sobre el método de evaluación de los resultados empleado y cómo es que se llevó a cabo éste. De igual manera, se darán a conocer algunas observaciones que se vieron durante el proceso de validación de términos y de evaluación del sistema desarrollado en la tesis.

Finalmente, en el quinto capítulo se mostrarán las conclusiones a las que se llegaron al finalizar este trabajo de tesis, así como el trabajo a futuro que se puede llevar a cabo.

1 PROCESAMIENTO DE LENGUAJE NATURAL

El *procesamiento de lenguaje natural (PLN)* es la función de componentes de software o hardware en un sistema de cómputo que analiza o sintetiza el lenguaje hablado o escrito (Jackson y Moulinier, 2002). Para tener una mejor comprensión del término procesamiento de lenguaje natural es necesario saber qué es un lenguaje natural. Un *lenguaje natural* es aquel propiamente usado por los humanos para comunicarse entre ellos, como el español, el inglés, el francés, entre otros. Se diferencia de los *lenguajes artificiales*, los cuales son creados para que haya una comunicación humano-máquina, como los lenguajes de programación.

Las aplicaciones y usos de PLN se encuentran en muchas de las cosas que empleamos en nuestra vida diaria. Casos en los cuales se usa el PLN son los motores de búsqueda en línea como Google o Yahoo!, los traductores y resumidores automáticos, los correctores de estilo y ortografía de los procesadores de texto, los reconocedores de voz, entre otros.

Además, las capacidades del procesamiento de lenguaje natural son muy grandes: permiten disminuir y/o facilitar tareas que anteriormente se realizaban de manera manual. Pero también han logrado que se lleven a cabo en menor tiempo o que sean realizables; por ejemplo, una persona jamás podría analizar toda la cantidad de información que existe en la Biblioteca del Congreso de los Estados Unidos de América, al menos de manera manual. Con el PLN, esto es posible.

Pero detrás de las aplicaciones basadas en el PLN hay varios procesos en los que es necesario realizar un tratamiento de la lengua, escrita o hablada, para que pueda ser analizada por una computadora. Los procesos realizados tratan de simular el proceso que lleva a cabo el ser humano para comprender y analizar la lengua. La diferencia entre este proceso y el efectuado por los seres humanos, es que una computadora puede analizar enormes masas de datos a velocidades muy rápidas, aunque no de manera tan exacta y precisa como lo hacemos los humanos.

En este capítulo se abordará el PLN, también conocido como tratamiento de lenguaje natural. Para ello en una primera instancia se darán a conocer algunas de las herramientas básicas que son empleadas dentro del PLN; posteriormente, dado que el objetivo es el uso del PLN para la extracción de la terminología, se dará a conocer lo que es la recuperación de información como una de sus aplicaciones, así como las diversas técnicas de la recuperación de información que son utilizadas en esta tesis.

1.1 Recursos y herramientas empleadas en PLN

A continuación se darán a conocer algunas de las herramientas y recursos más importantes empleados en PLN. Si bien son muchas las herramientas y los recursos existentes, los principales son los corpus lingüísticos, los tokenizadores, las listas de palabras vacías, entre otros.

1.1.1 Corpus lingüísticos

Uno de los recursos básicos empleados en el PLN son los *corpus lingüísticos*. El nombre de corpus proviene del latín y según el Diccionario de la Real Academia Española (DRAE), en su edición 22, indica que un corpus “es un conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación”. Aunque esta definición da a conocer, de manera general, lo que es un corpus, la definición de corpus lingüístico es un poco más específica.

Un *corpus lingüístico*, según Sierra (2008), consiste en la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos. Su función principal es establecer la relación entre la teoría y los datos (Torruella y Llisterra, 1999); es decir, un corpus debe cumplir los modelos teóricos con datos reales.

Todo corpus lingüístico es creado con base en determinados requerimientos, según nuestras necesidades y el tipo de análisis que se vaya a llevar a cabo con él. Aun así, existen una serie de parámetros, o de criterios mínimos, que se deben seguir para que pueda cumplir su función principal de manera exitosa, las cuales se explican a continuación:

Variedad: Este parámetro indica que los recursos que conforman el corpus deben ser diversos. En este sentido, los documentos pueden provenir de distintas fuentes, épocas, hablantes, lugares, entre otros. Por ejemplo, si se busca crear un corpus de ciencia ficción, no se puede basar solamente en la serie de libros de “Dune” de Frank Herbert, se debe de incluir de igual manera libros como “Yo robot” de Isaac Asimov, “La guerra de los mundos” de Herbert George Wells, hasta “20 mil leguas de viaje submarino” de Julio Verne.

Representatividad: Esto se refiere a que el corpus abarque, de la manera más amplia posible, todas las formas y variedades que existen en la lengua en determinada área, tema o tiempo. Es como en probabilidad, cuando se busca analizar una población sumamente grande se toma una muestra, la cual debe representar de manera general a todo el conjunto poblacional. En el caso de corpus lingüísticos, la población es la lengua a analizar mientras que el conjunto de documentos es la muestra a emplear.

Equilibrio: El equilibrio de un corpus está relacionado con la existencia de una neutralidad en todos los aspectos que se buscan analizar. Y aunque este parámetro es difícil de cumplir, se debe mantener lo más posible, de lo contrario los análisis que se lleven a cabo con el corpus podrían estar sesgados o limitar el uso del corpus en otras investigaciones.

Tamaño: Un corpus debe tener un tamaño que permita obtener resultados significativos. Según MacMullen (2003) los corpus muy grandes no son necesariamente representativos; de hecho pueden causar problemas para encontrar elementos menos abundantes pero más importantes.

Manejable por la computadora: En la actualidad muchos de los corpus lingüísticos se encuentran de manera digital; esto se recomienda ya que se pueden obtener beneficios de ello, por ejemplo se pueden realizar análisis o búsquedas de manera más rápida, ya que antes se realizaban estos procesos de forma manual. Esto ha dado lugar al término de corpus informatizado, es decir, un conjunto de textos elegidos y anotados con ciertas normas y criterios para el análisis lingüístico, de forma que se sirve de la tecnología y de las herramientas computacionales para generar resultados más exactos (Sierra, 2008).

Derechos de autor: Este se debe de tener en cuenta cuando se realiza un corpus. La razón de ello se debe a que es necesario no violar las leyes o normatividades con respecto al uso o reproducción de documentos, ya que un corpus está conformado por documentos creados por distintas personas.

Existe una gran cantidad de tipos de corpus y según Torruella y Llisterri (1999), se pueden clasificar de la siguiente manera: por el porcentaje y la distribución de los diferentes tipos de textos que lo conforman, por la especificidad de los textos que lo componen, por la cantidad de texto que se recoge en cada documento, por la codificación y las anotaciones añadidas a los textos y por la documentación que lo acompañe.

Algunos ejemplos de corpus lingüísticos son el Brown Corpus (Francis y Kučera, 1979), el Corpus Técnico del IULA (Vivaldi, 1995), el Corpus Histórico del Español en México (Medina y Méndez, 2006), el Corpus de Referencia del Español Actual (CREA) y el Corpus Diacrónico del Español (CORDE), estos dos últimos creados por la Real Academia Española.

1.1.2 Tokenizadores

Antes de realizar cualquier análisis lingüístico o de procesar un documento es necesario encontrar y separar cada uno de los elementos que lo conforman, para ello se emplean los tokenizadores. Los *tokenizadores* (también conocidos como analizadores léxicos o segmentadores de palabras) segmentan un conjunto de caracteres en unidades con significado llamados *tokens* (Jackson y Moulinier, 2002), que se pueden llamar igualmente *casos*. A continuación se presenta un ejemplo:

Frase: La niña juega con la pelota

Tokens:

La	niña	juega	con	la	pelota
----	------	-------	-----	----	--------

Como se puede observar en el ejemplo anterior, la frase es segmentada en seis tokens, empleando como delimitador de cada token el espacio en blanco. De igual manera podemos ver que en los tokens se cuentan las repeticiones, en el caso anterior, la palabra “la” aparece dos veces y se contabilizan ambas. En el caso en que nos refiramos a la clase de todos los casos contenidos con la misma secuencia de caracteres (Manning et al., 2008), en otras palabras, el número de distintos tokens, entonces estamos dando a conocer lo que se define como *type* o *tipo*. Por tanto, en el ejemplo anterior existen 5 tipos, es decir, “la” se contabiliza una sola vez.

Este proceso, aunque aparenta ser fácil de llevar a cabo, es una tarea compleja. Existen casos en el cual la regla de espacios en blanco como delimitador no es suficiente, algunos de ellos son los siguientes (Ananiadou y McNaught, 2006):

Fronteras ambiguas: Existen lenguas, como el alemán, que son aglutinantes, donde un conjunto de palabras se unen y no existe algún espacio en blanco entre ellas. Por ejemplo, estación de autobús en alemán es “busbahnhof” que es la unión entre “bus” que significa autobús y “bahnhof” que significa estación. Por tanto, en este tipo de casos es difícil encontrar la frontera entre las dos palabras ya que no existe un espacio de por medio. De igual forma pasa en lenguas como el japonés y el chino, donde no se escribe ningún tipo de espacio entre los caracteres.

Formatos: Fechas, números, teléfonos, entre otros, se escriben de distinta manera dependiendo del lugar de donde sea un documento, por tanto se debe prevenir esto para realizar la correcta tokenización. Por ejemplo, la notación que se emplea para separar los números, en algunos países se emplea la coma decimal en lugar del punto decimal; de igual manera algunos usan el punto, la coma o un espacio en blanco para separar los millares.

Guiónes: Otro de los casos son los elementos unidos por guiones, donde estos se puede considerar como un solo token o como varios. Algunos ejemplos son los siguientes, compra-venta, México-68, entre otros. De igual manera, en otros idiomas se emplea de manera distinta el guión, por ejemplo en el francés se emplea el guión para indicar una posición inversa a la normal de los elementos como “dis-moi!”. En portugués peninsular se unen con guión los clíticos “parece-me”.

Abreviaturas y siglas: Además del espacio en blanco como delimitador, se podría considerar a los signos de puntuación como límite entre palabras gráficas, pero el caso de las abreviaturas y de las siglas complica esto. Por consiguiente hay que diferenciar entre los puntos del final de una oración y los de una abreviatura o sigla, de lo contrario habría problemas en casos como F. C. (ferrocarril) o S.C.T. (Secretaría de Comunicaciones y Transportes).

Apóstrofes: Aunque en el español no se emplean los apóstrofes (‘), en otras lenguas sí se emplean, como en el inglés y en el francés. Ejemplo: “L’hôpital” (en francés el hospital) o “isn’t” (en inglés, contracción de is not).

Según Jackson y Moulinier (2002) los tokenizadores usualmente dependen de reglas, máquinas de estados finitos, modelos estadísticos, y lexicones para identificar abreviaturas o palabras de varios elementos.

1.1.3 N-gramas

La definición más simple de un *n-grama* es la unión de uno o varios tokens o caracteres. La construcción de los n-gramas se lleva por medio de combinaciones entre tokens o caracteres vecinos. Para llevar esto a cabo se crea una ventana del tamaño del n-grama deseado: si se quiere un unigrama el tamaño de la ventana será de 1, si son bigramas (o digramas) la ventana será de tamaño 2 y así sucesivamente hasta llegar al tamaño de n-grama final deseado. Esta ventana se mueve a través del texto y abarca la cantidad de tokens o caracteres que el tamaño de la ventana indica. Un ejemplo de formación de n-gramas de tokens es el siguiente:

Frase: El termómetro de mercurio se rompió

Tokens

El	termómetro	de	mercurio	se	rompió
----	------------	----	----------	----	--------

Unigramas

El	termómetro	de	mercurio	se	rompió
----	------------	----	----------	----	--------

Bigramas

El termómetro	termómetro de	de mercurio	mercurio se	se rompió
---------------	---------------	-------------	-------------	-----------

Trigramas

El termómetro de	termómetro de mercurio	de mercurio se	mercurio se rompió
------------------	------------------------	----------------	--------------------

Se emplean los n-gramas de tokens para abarcar la mayor cantidad de construcciones que se encuentran en un texto y no emplear en el análisis solamente construcciones de un elemento. Como se puede observar en el ejemplo anterior, en trigramas obtenemos una construcción que es importante en la frase y expresa claramente un objeto, esta es “termómetro de mercurio”.

Los n-gramas de caracteres se emplean para obtener construcciones que frecuentemente se emplean dentro de los tokens en una determinada lengua. Por ejemplo, con el uso de los n-gramas de caracteres se puede conocer el idioma de un texto sin haberlo leído (Cavnar y Trenkle, 1994).

1.1.4 Etiquetadores de partes de la oración

Los etiquetadores de partes de la oración, también conocidos como etiquetadores POS por sus siglas en inglés “Part-of-Speech”, son herramientas que realizan el proceso de asignar partes de la oración u otra clase de marcador léxico a cada palabra en un corpus (Jurafsky y Martin, 2008), en otras palabras, son sistemas que ayudan en la determinación de la categoría gramatical (por ejemplo, si es verbo, sustantivo, preposición, etcétera) de cada una de las palabras de un texto o conjunto de ellos. Asimismo, el etiquetado que realizan estas herramientas se aplica de igual forma a signos de puntuación, números, cantidades, entre otros.

Para llevar el proceso del etiquetado POS es necesario primeramente realizar un análisis morfológico. Un *análisis morfológico* o *análisis estructural* es el proceso de descomponer palabras complejas en sus componentes morfológicos (partes significantes de las palabras) (Bellomo, 2009), este análisis provee información sobre la semántica de la palabra y el papel sintáctico que juega en una oración (Goyal y Singh Lehal, 2008). En otras palabras, un análisis morfológico permite conocer las posibles categorías gramaticales de cada una de las palabras que se encuentran en una oración. A continuación se muestran los casos que se obtendrían al analizar morfológicamente la frase “El gato come pescado”:

Frase:		El	gato	come	pescado
Categoría gramatical	Caso 1:	Artículo masculino singular	Nombre común masculino singular	Verbo principal indicativo presente tercera persona del singular	Sustantivo común masculino singular
	Caso 2:	-	-	Verbo principal imperativo segunda persona del singular	Verbo principal participio singular masculino

Como se puede observar en el ejemplo anterior, para “come” y “pescado” existen dos posibles casos; “come” puede representar un verbo en imperativo o en presente en tercera persona del singular, la razón es porque para los verbos terminados en –er estas dos formas terminan en –e. En cambio, la palabra “pescado” puede representar un sustantivo o un verbo en participio por su terminación –ado.

El segundo paso para hacer un etiquetado POS consiste en desambiguar los casos otorgados por el análisis morfológico y poner la etiqueta de la parte de la oración más probable. Para llevar a cabo esto es necesario utilizar etiquetadores que emplean algoritmos que pueden ser de dos tipos (Jurafsky y Martin, 2008):

Basados en reglas: Este tipo de algoritmo se basa en un conjunto de reglas escritas a mano que desambiguan los casos. Por ejemplo, una palabra será un sustantivo y no un verbo si está antecedido de un artículo.

Basado en aprendizaje: Este tipo de algoritmo se basa en las probabilidades que tiene una etiqueta en un determinado contexto. Para calcular estos valores es necesario utilizar previamente un corpus de aprendizaje, es decir, un corpus que fue etiquetado a mano lo suficientemente grande para que un programa computacional puede calcular las probabilidades de las etiquetas según el contexto.

Existen diversos formatos de etiquetas, incluso para la misma lengua; esto se debe a que no todos clasifican o anotan de la misma manera las partes de la oración. Sin embargo, existen diversos estándares, siendo el más conocido el desarrollado por el grupo EAGLES¹

¹ <http://www.ilc.cnr.it/EAGLES96/home.html>

(Expert Advisory Group on Language Engineering Standards) ya que trata de establecer un mismo formato de etiquetas para las lenguas de la Unión Europea (español², inglés, francés, etcétera). Las etiquetas de EAGLES tienen una estructura variable según la categoría gramatical y el idioma, ya que no en todas las lenguas europeas las categorías gramaticales tienen los mismos atributos; en la Tabla 1 se muestra un ejemplo de esta estructura para el caso de adverbios.

Adverbios			
Posición	Atributo	Valor	Código
1	<i>Categoría</i>	<i>Adverbio</i>	<i>R</i>
2	<i>Tipo</i>	<i>General</i>	<i>G</i>
		<i>Negativo</i>	<i>N</i>

Tabla 1. Estructura de la etiqueta de adverbios para el formato EAGLES

Con base en la Tabla 1, adverbios como “rápido” o “siempre” obtendrían una etiqueta “RG” mientras que “jamás” o “no” tendrían una etiqueta “RN”.

1.1.5 Lematizadores

En todo documento escrito en una lengua flexiva, como el español y el italiano, existen variaciones léxicas de las palabras, es decir, se pueden tener casos como “escribimos”, “democracias”, “industrialización”, etcétera. Pero dentro del procesamiento de lenguaje natural es necesario disminuir la cantidad de variaciones léxicas que existan en los documentos a analizar. Para ello se debe obtener el *lema* o *forma canónica*, es decir, la base o la forma de diccionario de una palabra (Manning et al., 2008).

El proceso de lematización se lleva a cabo de manera automática por parte de los humanos; cuando queremos buscar las palabras “encontramos” o “niñas” en un diccionario las pasamos a “encontrar” y “niño”, para ello empleamos nuestro conocimiento del mundo.

² Para información sobre las etiquetas EAGLES con ejemplos en español se puede consultar <http://nlp.lsi.upc.edu/freeling/doc/userman/parole-es.pdf>

Pero para las computadoras realizar este procedimiento es más complicado, ya que no tienen acceso a este conocimiento, por tanto, es necesario darle una serie de reglas y de recursos.

Para reducir el número de variaciones léxicas, ya sea por flexiones (*caminar* → *caminamos*) o por derivaciones (*activar* → *activación*), existen dos herramientas que se emplean en PLN, que son los *lematizadores* y los *truncadores* o *stemmers*. Aunque frecuentemente se confunden ambos términos, cabe aclarar que son dos métodos distintos.

Los lematizadores son herramientas que emplean diccionarios o tesauros, al igual que reglas, que buscan obtener el lema de las palabras; además estas herramientas realizan un etiquetado POS (sección 1.1.4) para conocer la categoría gramatical de las palabras.

A continuación se muestra un ejemplo de los lemas obtenidos por la lematización en dos frases:

Frase 1: El cuidado de las obras de arte es un trabajo arduo

Lemas:

el	cuidado	de	el	obra	de	arte	ser	un	trabajo	arduo
----	---------	----	----	------	----	------	-----	----	---------	-------

Frase 2: El museo fue cuidado por los policías

Lemas:

el	museo	ser	cuidar	por	el	policía
----	-------	-----	--------	-----	----	---------

Se puede observar que en la frase 1 tanto la palabra “cuidado” como “trabajo”, al lematizarse quedan sin cambios, la razón de ello es que ambas palabras hacen referencia a un sustantivo, y no a un verbo en participio y a un verbo en presente en primera persona, respectivamente. En cambio, en la frase 2, la palabra “cuidado” al ser un verbo conjugado en participio, su lema es “cuidar”.

Los truncadores, en cambio, son herramientas que emplean la heurística, es decir, ciertas reglas, para cortar las partes finales de las palabras, con el fin de llegar a su lema; estas herramientas no usan ni analizadores morfológicos o etiquetadores POS, aunque hay truncadores que incluyen reglas para eliminar afijos derivacionales, como “-ción”, “-ía”. Al no realizar un análisis morfológico o un etiquetado POS los stemmers no encuentran la diferencia entre “cuidado” que proviene del verbo y “cuidado” que es un sustantivo, como lo hacen los lematizadores. De igual forma, el problema de los stemmers es que el hecho de

cortar las partes finales de las palabras no siempre implica que se obtendrá la forma canónica correcta. A continuación en la Tabla 2 se muestran algunos ejemplos de la truncación:

Palabra	Lema por un stemmer ³
chicharrones	chicharron
torres	torr
torreón	torreon

Tabla 2. Ejemplos de lemas obtenidos por un stemmer

Como se puede observar en la tabla anterior, existen casos en el cual el lema es el correcto (siempre y cuando se omita la falta de acentos), pero en otros casos, como el de “torres”, su lema dado es incorrecto, la razón de este caso es que la regla empleada en el truncador indica que cuando una palabra termina en “-es” es un plural y por tanto se debe de quitar; en casos como “meses” o “celulares” sí funciona la regla anterior. La mayor ventaja que tienen los stemmers sobre los lematizadores es la velocidad de procesamiento.

Algunas herramientas lematizadoras son el FreeLing (Padró et al., 2010) y el TreeTagger (Schmid, 1994); mientras que para truncar existen sistemas basados en el algoritmo de Porter (Porter, 1980), que comenzó para el idioma inglés, pero se ha llevado a otros idiomas.

1.1.6 Palabras funcionales

Las *palabras funcionales*, *palabras vacías* o *stop words*, son las palabras que “carecen” de significado. Estas palabras son las de mayor frecuencia y las que aportan la menor cantidad de información, entre ellas se encuentran los artículos, las preposiciones, las conjunciones, entre otras. De las palabras funcionales se crean *listas de paro* o *stoplists*.

Según Pazienza et al. (2005), las palabras funcionales son automáticamente extraídas de un corpus genérico como aquellas con la más alta frecuencia, y posteriormente son

³ Los ejemplos fueron obtenidos de la página del proyecto “Snowball” el cual es un conjunto stemmers basados en algoritmo de Porter. La dirección web es la siguiente:
<http://snowball.tartarus.org/algorithms/spanish/stemmer.html>

validadas por expertos humanos. De igual manera, se pueden agregar algunas otras palabras que se desean eliminar en PLN.

El objetivo de emplear stoplists en PLN es reducir la cantidad de datos a analizar. De igual manera disminuye el espacio en memoria o en disco empleado por las herramientas que analizan lenguaje natural. En el siguiente ejemplo se puede observar que se quitan las palabras funcionales:

Frase 1: El monitor de esa computadora se descompuso

Tokens:

el	monitor	de	esa	computadora	se	descompuso
----	---------	----	-----	-------------	----	------------

Eliminando las palabras funcionales

Resultado:

monitor	computadora	descompuso
---------	-------------	------------

1.2 Recuperación de información

Existen múltiples tareas dentro de PLN, una de las más importantes y más utilizadas es la recuperación de información. La *recuperación de información*, también conocida como *búsqueda de información* o *information retrieval (IR)*, es el proceso por el cual se otorgan documentos relevantes (o información sobre ellos) a un usuario, según la consulta que se haya realizado (Kageura y Umino, 1998). Esta tarea es una de las más obvias que existe en PLN porque es una con la que se tiene mayor contacto; por ejemplo, se emplea la recuperación de información en buscadores en línea, en bibliotecas digitales, en el programa para buscar dentro de la computadora y prácticamente en cualquier medio electrónico.

Como se vio en el párrafo anterior, para llevar a cabo una recuperación de información, es necesaria una *consulta* o *query*, que puede ser la correcta o no para buscar la información, de igual manera puede estar mal escrita, con faltas de ortografía, con exceso de conectores o de datos inútiles, entre otros; lamentablemente, lo que se encuentre en esa consulta es la única pista que da el usuario para llevar a cabo la búsqueda necesaria. Por tanto, es necesario que todo sistema de recuperación de información tenga en cuenta las consideraciones anteriores.

Según Frakes y Baeza-Yates (1992) casi todos los sistemas de recuperación de información utilizan operadores booleanos o patrones de texto. Los primeros son empleados en sistemas de búsqueda donde existe una gran colección de documentos, como en el internet o en una biblioteca digital; en estos sistemas, cada documento es representado por una lista de palabras claves o de identificadores. De igual manera, en los sistemas booleanos, el usuario puede conectar los elementos de la consulta por medio de conectores lógicos. En cambio, en los sistemas basados en patrones, las búsquedas se basan en cadenas de texto o en expresiones regulares, estos sistemas se emplean dentro de documentos, o en colecciones pequeñas de archivos.

Existen diversos métodos para la creación de las listas de palabras claves que se emplean en los sistemas de recuperación de información booleanos. Uno de ellos es TF-IDF o term frequency-inverse document frequency, el cual se explicará en el siguiente apartado.

1.2.1 Term frequency – Inverse document frequency (TF-IDF)

Uno de los métodos empleados para la creación de listas de palabras clave para los sistemas de búsqueda de información es *TF-IDF (Term Frequency – Inverse Document Frequency)*, el cual es la unión de dos métricas, la primera de ellas es la de *frecuencia del término, term frequency* o *TF* y la segunda que es la *frecuencia inversa de los documentos, inverse document frequency* o *IDF*.

El método de TF-IDF genera listas de palabras clave con una calificación o peso que indica qué tan relevante es la palabra con respecto al documento seleccionado y al corpus en general. Además, estas listas permiten calificar a los documentos del corpus con base en estas palabras clave, es decir, si las palabras clave tienen un gran peso, entonces el documento está más relacionado con ellas que uno con las mismas palabras clave pero con menor peso. Por tanto, cuando un usuario ingrese una consulta, los documentos que tengan las palabras de esa consulta con mayor peso serán los que muestre el sistema de búsqueda de información.

La primera medida, TF, es un sistema de pesos basado en la idea de que construcciones (palabras, frases, grupos de palabras) que frecuentemente ocurren en el texto

de documentos tienen alguna relación con el contenido de los textos (Salton y McGill, 1986). El cálculo de estos pesos se lleva a cabo calculando la frecuencia relativa⁴ de cada una de las construcciones (también llamados términos) en cada uno de los documentos a analizar; esto se puede representar por medio de la siguiente fórmula:

$$TF_{i,j} = f_i^j \quad (1)$$

Donde f es el número de ocurrencias del término i en el documento j .

La segunda métrica, llamada Inverse Document Frequency o IDF, está basada en contar el número de documentos de la colección en donde se busca, que contienen (o están indizadas por) el término en cuestión (Robertson, 2004), en otras palabras, es saber el número de documentos de un corpus en donde las construcciones de palabras se encuentran. El nombre inicial de IDF fue *term specificity* y su fórmula fue la siguiente (Spärck-Jones, 1972):

$$term\ specificity = f(N) - f(n) + 1 \quad (2)$$

Donde la función $f(x) = y$ tal que $2^{y-1} < x \leq 2^y$, en la Figura 1 se puede observar los valores que podría tomar y ; N es el número de documentos que existen en el corpus y n es el número de documentos donde el término analizado se encuentra.

⁴ La frecuencia relativa es el número de ocurrencias de un caso en determinado evento. Se diferencia de la frecuencia absoluta, que es la frecuencia relativa entre la suma de las ocurrencias de todos los casos en el evento.

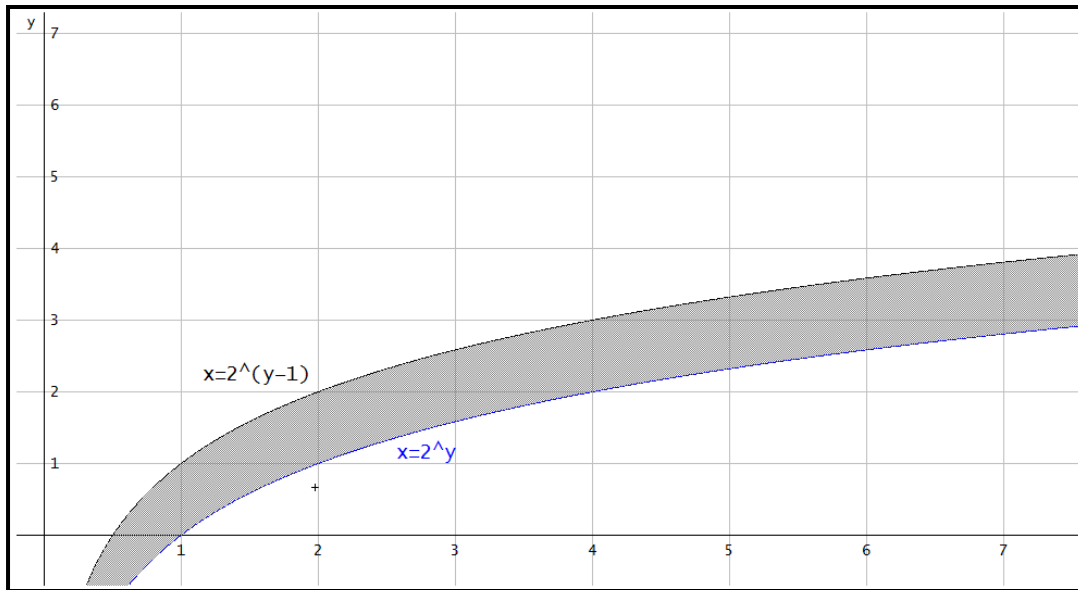


Figura 1. Gráfica de los valores de la función $f(x) = y$ tal que $2^{y-1} < x \leq 2^y$

Posteriormente la fórmula de *term specificity* fue adecuada por Robertson (1972) debido a que la fórmula $f(x)$ se podía aproximar a $f(x) \approx \log_2 x$ (esta aproximación es la función inversa de $x = 2^y$) y el 1 que se encontraba en la fórmula original era para evitar valores iguales a 0 en valores de n muy cercanos al valor de N , por lo tanto se podía prescindir de él. De igual forma le fue cambiado el nombre a inverse document frequency (IDF) y se observó que el logaritmo no tenía que ser forzosamente base 2, esto debido a la siguiente propiedad de los logaritmos:

$$\log_b a = \frac{\log_c a}{\log_c b} \quad (3)$$

Donde a es el número al cual se le calculará el logaritmo; b es la base del logaritmo original y c es la base de un logaritmo distinto a b . En el caso del IDF, se cambió el logaritmo de base 2 a base 10 debido a que es uno de los más usados.

La fórmula empleada en la actualidad para el cálculo de la métrica Inverse Document Frequency es la siguiente:

$$IDF_i = \log \frac{N}{n_i} \quad (4)$$

Donde N es el número de documentos que existen en el corpus, n es el número de documentos donde el término i se encuentra.

La creación del sistema de pesos basado en IDF se fundó en la teoría de Spärk-Jones que indicaba que los términos con alta frecuencia pueden ser útiles para aumentar la cobertura, pero las correspondencias entre la consulta y los términos del documento que ocurren raramente en una colección de documentos deberían ser tomadas como más importantes que aquellas que ocurren frecuentemente (Salton y Yang, 1973).

A partir de estos dos sistemas de pesos, se desarrolló el TF-IDF (Salton y Yang, 1973), el cual trabaja determinando la frecuencia relativa de las palabras en un documento específico comparado con la proporción inversa de esa palabra en todo el corpus (Ramos, 2003). Su fórmula es la siguiente:

$$TF - IDF_{i,j} = f_i^j * \log \frac{N}{n_i} \quad (5)$$

Donde f_i^j es la frecuencia del término i en el documento j ; N es el número total de documentos que conforman el corpus y n_i es el número de documentos donde se encuentra el término i .

En el método de TF-IDF existen, a grandes rasgos, cuatro tipos de pesos que son los siguientes:

Peso grande: Este ocurre cuando un término tiene una alta frecuencia de aparición pero no se encuentra en la mayoría de los documentos analizados.

Peso mediano: En este caso, el número de apariciones tanto dentro de un documento como en los archivos que conforman el corpus no es baja ni alta.

Peso bajo: Su aparición sucede cuando la frecuencia del término es baja y la construcción se encuentra en la mayoría de los documentos.

Peso nulo: Acontece cuando la frecuencia de aparición de un término dentro de un documento es nula o cuando el término aparece en cada uno de los documentos que pertenecen al corpus analizado.

En la Tabla 3 se muestra un ejemplo de los pesos calculados por el método de TF-IDF para una serie de términos en tres documentos:

Término	Documento 1			Documento 2			Documento 3		
	TF	IDF	TF-IDF	TF	IDF	TF-IDF	TF	IDF	TF-IDF
la	10	$\log \frac{3}{3} = 0$	0	12	$\log \frac{3}{3} = 0$	0	9	$\log \frac{3}{3} = 0$	0
geometría	5	$\log \frac{3}{2} = 0.17$	0.85	7	$\log \frac{3}{2} = 0.17$	1.19	0	$\log \frac{3}{2} = 0.17$	0
analítica	5	$\log \frac{3}{1} = 0.47$	2.35	0	$\log \frac{3}{1} = 0.47$	0	0	$\log \frac{3}{1} = 0.47$	0
descriptiva	0	$\log \frac{3}{1} = 0.47$	0	6	$\log \frac{3}{1} = 0.47$	2.82	0	$\log \frac{3}{1} = 0.47$	0
casa	0	$\log \frac{3}{1} = 0.47$	0	0	$\log \frac{3}{1} = 0.47$	0	8	$\log \frac{3}{1} = 0.47$	3.76
de	11	$\log \frac{3}{3} = 0$	0	8	$\log \frac{3}{3} = 0$	0	10	$\log \frac{3}{3} = 0$	0
José	0	$\log \frac{3}{2} = 0.17$	0	2	$\log \frac{3}{2} = 0.17$	0.34	5	$\log \frac{3}{2} = 0.17$	0.85

Tabla 3. Ejemplo de la aplicación de TF-IDF

Como se puede observar, el método de TF-IDF permite la discriminación de palabras frecuentes, como lo son las palabras funcionales, asignándoles pesos nulos o bajos. Por tanto, estas no suelen quedar en las listas de palabras clave que se emplean en los sistemas de búsqueda de información. Pero además califica con pesos altos las palabras con alto valor de importancia para cada uno de los documentos analizados.

Hasta la fecha se han realizado diversas variaciones al método de TF-IDF y se han empleado métodos de normalización, los cuales serán expuestos en el siguiente apartado. De igual manera el uso de TF-IDF ha pasado a otras áreas de PLN.

1.2.2 Normalización de la longitud del documento

Una de las modificaciones que han sido agregadas al método de TF-IDF es el empleo de un factor de normalización que permita equilibrar casos en los que se empleen documentos de distintos tamaños. Es decir, un documento muy extenso tiene una ventaja más alta sobre los

documentos más pequeños a la hora de recuperar información, esto debido a que el tamaño del documento es un parámetro que hasta cierto punto afecta el cálculo de los pesos. Esta ventaja se observa cuando el usuario hace una consulta al sistema de recuperación de información, el cual traerá los documentos que contengan los términos de la consulta con mayor peso.

Según Singha et al. (1996) existen dos razones principales por las cuales es necesario emplear la normalización:

Frecuencias de términos más altas: Los documentos grandes usualmente emplean los mismos términos repetidamente. Como resultado, los factores de frecuencia de los términos pueden ser grandes para documentos largos, aumentando la contribución promedio de sus términos a la similitud de la consulta de documentos.

Más términos: Los documentos largos también tienen una gran cantidad de términos diferentes. Esto aumenta el número de coincidencias entre la consulta y un documento largo, aumentando la similitud de la consulta de documentos, y los casos de recuperar documentos largos en vez de documentos cortos.

La normalización, a grandes rasgos, permite tratar de la misma manera a todos los documentos sin importar su longitud.

Existen diversos métodos para llevar a cabo la normalización, algunos de ellos atacan las dos razones principales vistas anteriormente, algunos otros métodos solamente una de ellas.

1.2.2.1 Normalización de coseno

La normalización de coseno es uno de los métodos más empleados para normalizar los pesos de TF-IDF, se basa en el *modelo de espacio vectorial* o *vector space model*. Este modelo es la representación de un conjunto de documentos como vectores en el espacio vectorial común (Manning et al., 2008). Un *vector* es un segmento de línea dirigido⁵ con dirección, sentido y

⁵ Según Castañeda De Isla Puga (2000) un *segmento de línea dirigido* o *segmento dirigido* es un segmento de recta en el que se ha asignado un punto origen y un punto extremo.

magnitud. Es decir, un documento se puede representar como un vector que sería de la siguiente manera:

$$\bar{D}_i = (w_1, w_2, w_3, \dots, w_{n-2}, w_{n-1}, w_n)$$

Donde D_i es el documento i -ésimo del corpus y w_j es cada uno de los pesos de los términos que conforman el documento, en el modelo de espacio vectorial, los pesos representan cada uno de los componentes escalares⁶ de un vector de n dimensiones. De manera gráfica, un documento con tres términos podría representarse como se muestra en la Figura 2.

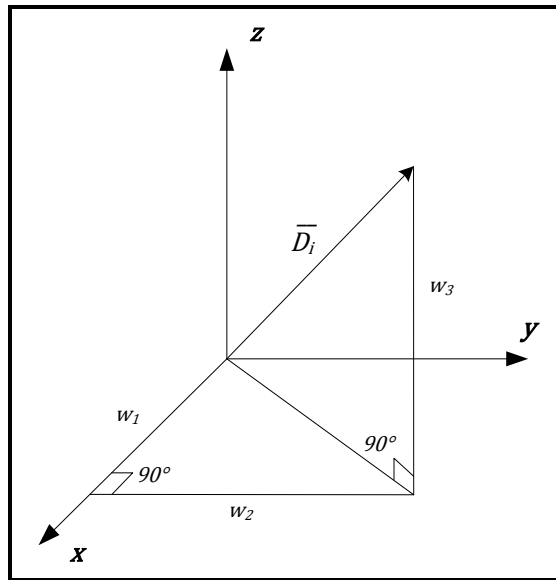


Figura 2. Representación gráfica de un documento de 3 términos en el modelo de espacio vectorial

Al igual que a cualquier otro vector, al vector \bar{D} , se le puede sacar su módulo; el *módulo*, *norma*, o *magnitud* de un vector es el tamaño de cualquier segmento dirigido que lo representa (Castañeda De Isla Puga, 2000); este valor es un escalar siempre positivo. Para un vector de la forma $\bar{V} = (a_1, a_2, a_3, \dots, a_{n-1}, a_n)$ el módulo se calcula de la siguiente manera:

⁶ Un *escalar* o *cantidad escalar* es un número o medida en que la dirección no interviene o carece de significado (Daintith, 2001). En otras palabras, un escalar sólo indica una magnitud.

$$|\bar{V}| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \dots + a_{n-1}^2 + a_n^2} = \sqrt{\sum_{i=0}^n a_i^2} \quad (6)$$

Donde $|\bar{V}|$ es el módulo del vector y a_i es cada uno de los componentes escalares que conforman al vector V . Tomando lo anterior en cuenta, para el vector de un documento \bar{D} , su módulo sería la raíz cuadrada de la suma de los pesos, es decir:

$$|\bar{D}| = \sqrt{\sum_{i=0}^n w_i^2} \quad (7)$$

De la misma manera en que se puede obtener el módulo de un vector, también se puede convertir en un vector unitario. Un *vector unitario* es aquel en el que su módulo es igual a la unidad (Castañeda De Isla Puga, 2000). Para ello se emplea la normalización de vectores que para un vector de la forma $\bar{V} = (a_1, a_2, a_3, \dots, a_{n-1}, a_n)$ consiste en la siguiente:

$$\bar{v} = \frac{\bar{V}}{|\bar{V}|} \quad (8)$$

Donde \bar{v} es el vector unitario resultante, \bar{V} es el vector original y $|\bar{V}|$ es el módulo del vector \bar{V} .

A partir de la normalización anterior, se desarrolló la *normalización de coseno* (Salton y Buckley, 1988), el cual consiste en multiplicar el vector del documento por el *factor normalizador*, como se muestra en la siguiente fórmula:

$$\bar{d}_i = \bar{D}_i * \frac{1}{|\bar{D}_i|} = \frac{\bar{D}_i}{|\bar{D}_i|} \quad (9)$$

La normalización de coseno permite eliminar los dos casos por los cuales se normaliza, mismos que fueron vistos en la sección 1.2.2. De igual manera cuando se emplea la normalización de coseno, los pesos de cada uno de los términos tienen una escala que va del cero al uno, debido a que este proceso crea vectores unitarios.

1.2.2.2 Normalización por pivote

Uno de los problemas que ocurren al normalizar es que el factor de normalización penaliza en exceso a los pesos de los documentos grandes, otorgándoles una desventaja a la hora de realizar una búsqueda. Por tanto, es necesario equilibrar el factor de normalización para aumentar la posibilidad de recuperar un documento de cierta longitud.

Para enfrentar este problema, se desarrolló la normalización por pivote (Singha et al., 1996). Este método está conformado a grandes rasgos en dos pasos, el primero de ellos es el cálculo de un factor de normalización por medio de un método de normalización como el de por coseno, y posteriormente el cálculo del nuevo factor de normalización.

El método se basa en la idea de que después de normalizar hay un punto en el cual la probabilidad de recuperar documentos de cierta longitud disminuye cuando su probabilidad de relevancia aumenta. Por ello es necesario rectificar el factor de normalización. En la Figura 3 se puede observar lo explicado antes.

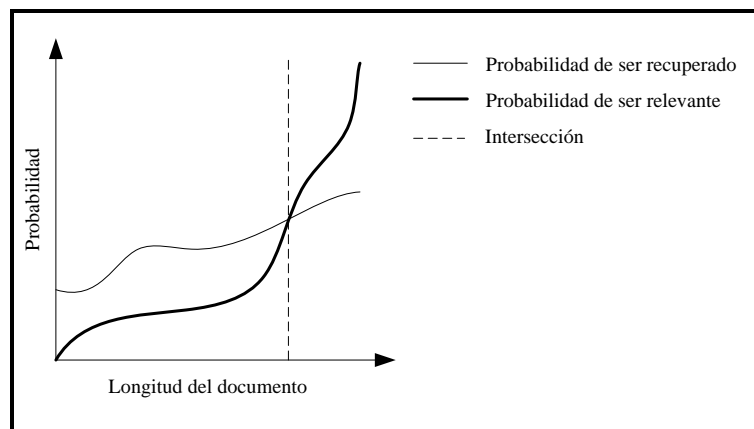


Figura 3. La probabilidad de recuperar un documento normalizado disminuye mientras más relevante sea

La normalización por pivote consiste en rectificar el factor de normalización, disminuyendo su valor en documentos largos, pero aumentándolo en documentos cortos. Para llevar a cabo esta rectificación es necesario considerar lo siguiente:

$$N_1 = n \quad (10)$$

$$N_2 = m(N_1) + b \quad (11)$$

Donde N_1 es la recta que representa la normalización original (una recta identidad) y N_2 es la recta que representa la normalización rectificada; n son los valores de la normalización original, m es la pendiente de la recta y b es la ordenada al origen. En un plano estas dos rectas se verían como se muestra en la Figura 4, el punto donde se intersecan es el pivote (p).

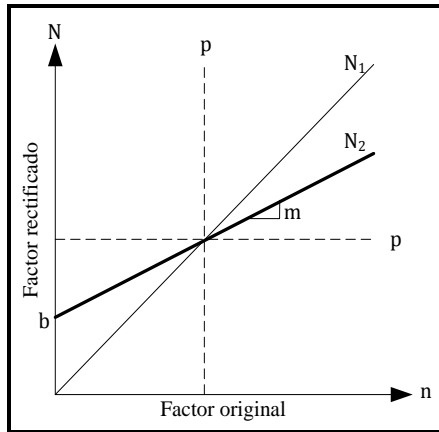


Figura 4. Representación gráfica de las rectas de normalización

La ecuación de la recta normalizada rectificada tiene dos parámetros los cuales se desconocen por defecto. En el caso de la pendiente m , si su valor es igual a 1 entonces se genera una recta paralela a la recta N_1 ; si su valor es 0 entonces N_2 es una recta horizontal y normalizaría todos los documentos de la misma manera; por tanto, es necesario que el valor de m esté en el intervalo $(0,1)$. Con respecto a b este debe tener un valor que permita a la recta N_2 cruzar por el mismo punto por el cual la recta N_1 cruza el pivote. Por tanto, para obtener el valor de b se sustituye la Ecuación (10) en la Ecuación (11) y queda de la siguiente manera:

$$N_2 = (m * n) + b \quad (12)$$

Si se considera el punto donde se intersecan las dos rectas de normalización, tiene por coordenadas (p,p) , entonces la Ecuación (12) se podría escribir de la siguiente manera para $n = p$:

$$p = (m * p) + b \quad (13)$$

Simplificando la Ecuación (13) y despejando b :

$$b = (1 - m)p \quad (14)$$

Finalmente sustituyendo la Ecuación (14) en la Ecuación (12) se tiene:

$$N_2 = (m * n) + (1 - m)p \quad (15)$$

La ecuación anterior es la que se emplea para obtener el factor de normalización por pivote, el cual se aplica dividiendo cada uno de los pesos de los términos sin normalizar entre el factor de normalización rectificado. El valor de p , es decir, del pivote que se emplea, no tiene una fórmula definida, simplemente los autores del método (Singha et al., 1996) recomiendan emplear como p el promedio de todos los pesos del documento a normalizar. Con respecto a m es un valor que se calcula empíricamente dependiendo del documento y que debe encontrarse entre cero y uno.

1.2.2.3 Normalización por máximo TF

La normalización por la máxima frecuencia de un término (TF) es otro método empleado en los sistemas que emplean TF-IDF. Esta normalización ha sido empleada en sistemas como SMART e INQUERY (Singha et al., 1996).

El método consiste en normalizar la frecuencia de un término por medio de la siguiente fórmula:

$$nTF_{i,j} = \alpha + \left((1 - \alpha) * \left(\frac{TF_{i,j}}{TF_{máx}(j)} \right) \right) \quad (16)$$

Donde α es el valor de suavidad (smoothing), el cual va de 0 a 1, aunque en la práctica se emplea un valor de $\alpha=0.4$; $TF_{i,j}$ es la frecuencia del término i en el documento j ; mientras que $TF_{máx}(j)$ es la frecuencia máxima de los términos que se encuentran en el documento j . Mientras que $nTF_{i,j}$ es la frecuencia del término normalizada.

Este método limita la frecuencia del término (TF) a un valor máximo de 1, por tanto, sólo compensa la primera razón por la cual se realiza la normalización (frecuencia alta de términos). En la Tabla 4 se muestra un pequeño ejemplo de lo anterior para $\alpha = 0.4$.

<i>TF</i>	<i>nTF</i>
10	$0.4 + (0.6 * (10/10)) = 0.4 + (0.6 * 1) = 1.00$
8	$0.4 + (0.6 * (8/10)) = 0.4 + (0.6 * 0.8) = 0.88$
5	$0.4 + (0.6 * (5/10)) = 0.4 + (0.6 * 0.5) = 0.7$
7	$0.4 + (0.6 * (7/10)) = 0.4 + (0.6 * 0.7) = 0.82$

Tabla 4. Ejemplo sobre la normalización por máxima frecuencia

1.2.3 Evaluación de sistemas de recuperación de información

En todo sistema, esté relacionado con el PLN o con cualquier otra área, es necesario siempre realizar una evaluación en donde se indique qué tan bueno es el sistema, de esta manera se puede saber cuáles son los puntos fuertes y/o débiles, y en qué se debe mejorar. Pero para llevar a cabo lo anterior es necesario contar con medidas que sean precisas, exactas y reflejen poca subjetividad.

En el caso de la recuperación de información existen dos medidas básicas que se emplean para calificar todo sistema que emplee esta tarea del procesamiento de lenguaje natural. Estas dos medidas son *precisión* o *precision*, por su término en inglés, y *cobertura* o *recall*, aunque de este último término existen otras traducciones al español como, *exhaustividad*, *especificidad* y *recuerdo*. Según Gelbukh y Sidorov (2006) la precisión es la relación entre los resultados correctos sobre los resultados obtenidos en total, mientras que la cobertura es la relación entre los resultados correctos sobre los resultados que deberían haber sido obtenidos. En otras palabras, la precisión indica qué tan preciso fue la recuperación de información, mientras que la cobertura da a conocer si se trajeron todos los resultados que debían ser traídos.

La fórmula para calcular la precisión de un sistema es la siguiente:

$$P = \frac{\text{Número de elementos relevantes recuperados}}{\text{Número de elementos recuperados}} \quad (17)$$

Mientras que para el cálculo de la cobertura se emplea la siguiente fórmula:

$$R = \frac{\text{Número de elementos relevantes recuperados}}{\text{Número de elementos relevantes}} \quad (18)$$

Aunque los desarrolladores de los sistemas de recuperación de información buscan que se cumplan con los mejores niveles de precisión y cobertura, estas dos medidas son hasta cierto punto inversamente proporcionales, es decir, mientras se busca mayor precisión en un sistema, claramente se dejarán pasar los casos que salgan de la norma y por tanto disminuirá la cobertura; en cambio, si se busca que el sistema obtenga todos los casos posibles, es decir, mayor cobertura, la precisión disminuirá porque se deja pasar mayor cantidad de información. En la Figura 5 se muestran dos gráficas, la primera es la curva ideal de la cobertura contra la precisión y la segunda es el comportamiento típico de esta curva en los diversos sistemas relacionados con el PLN.

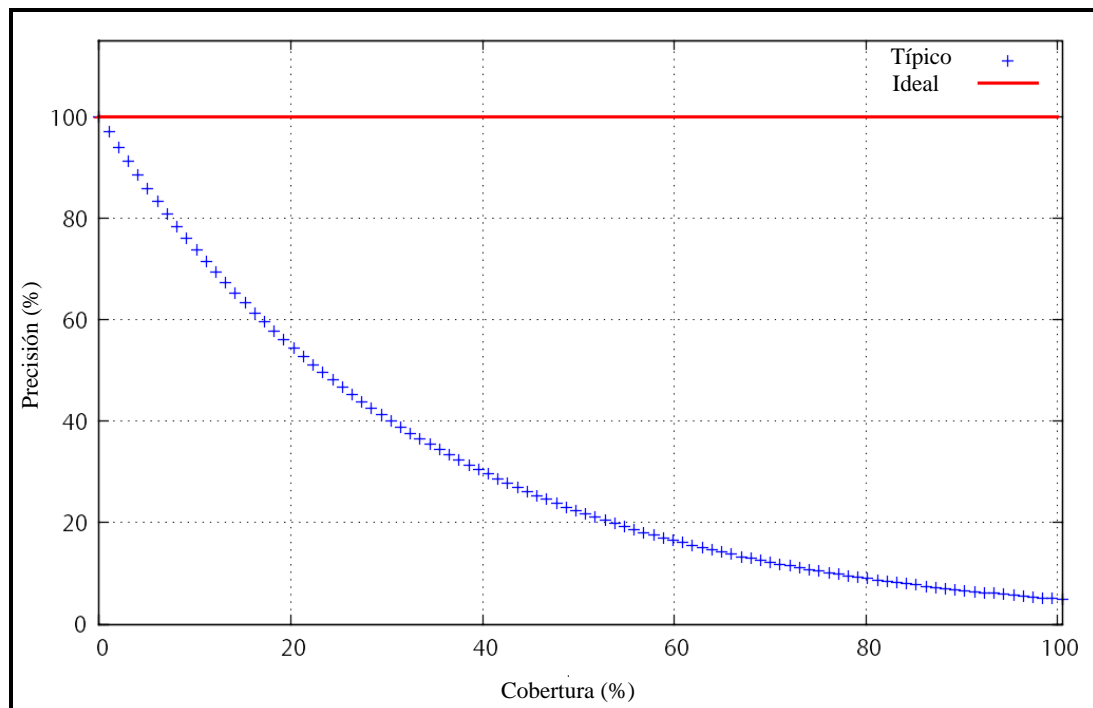


Figura 5. Gráficas típica e ideal del comportamiento de la cobertura contra la precisión

Si bien ambas medidas son las más utilizadas para evaluar los sistemas de recuperación de información, existe otra llamada *F-score* (van Rijsbergen, 1979). Esta medida también conocida como *F-measure* o F_1 es una combinación entre las medidas de precisión y cobertura, su fórmula es la siguiente:

$$F_1 = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad (19)$$

Donde β es el factor relativo al peso que se le dará a la precisión y a la cobertura, si su valor es igual a 1, ambas medidas reciben el mismo peso, si es mayor a 1 la precisión es favorecida, mientras que si es menor a 1 la cobertura tendrá mayor peso; P es la precisión y R es la cobertura del sistema.

1.2.4 Extracción y recuperación de información

Además de la recuperación de información existe otra tarea dentro de PLN que se llama *extracción de información* o *information extraction (IE)*. La extracción de información es el nombre dado a cualquier proceso que selecciona estructuras y combina datos que son encontrados, de manera explícita o implícita, en uno o más textos (Cowie y Wilks, 2000).

A grandes rasgos la extracción de información se caracteriza por lo siguiente (Ananiadou y McNaught, 2006):

- Toma un texto en lenguaje natural de un documento fuente, y extrae los hechos esenciales acerca de uno o más tipos de hechos predefinidos.
- Representa cada uno de los hechos como una plantilla donde los espacios son llenados con base en lo encontrado en el texto. Según Jurafsky y Martin (2008) sólo una pequeña parte de la información encontrada es relevante para llenar la plantilla; el resto puede ser ignorado.

El propósito de la extracción de información es estructurar el texto posiblemente no estructurado (Pudota et al., 2008). Con esto se pueden obtener resultados que pueden ser otorgados de manera directa al usuario o ser modificados para que puedan ser insertados en una base de datos.

Algunos ejemplos de las tareas de la extracción de información son identificar al hablante en un comunicado, encontrar las proteínas mencionadas en un artículo de una revista de biomedicina y extraer los nombres de las tarjetas de crédito aceptadas por un restaurante desde una reseña en línea (Kauchak et al., 2002).

La recuperación de información y la extracción de información son tareas similares, debido a que traen la información más relevante. De igual manera emplean las mismas métricas para evaluar los sistemas, las cuáles son, precisión, cobertura y F_1 . Asimismo, existen métodos que se emplean en ambas tareas, como el TF-IDF que se vio en el apartado 1.2.1. Aunque la recuperación y la extracción de información comparten características existen algunas diferencias: la extracción de información regresa hechos, datos o estructuras de ellos de manera automática, sin basarse en una consulta en concreto, por tanto, permite obtener datos sin hacer un análisis manual de los documentos fuente. En cambio, la recuperación de información busca información con base en una consulta, para ello los documentos fuente se deben encontrar indizados y organizados de tal manera que permitan la búsqueda. Frecuentemente se emplea la extracción de información para alimentar los datos que se emplean en la recuperación de información.

2 TERMINOLOGÍA

La palabra *terminología* en una primera instancia se puede considerar como la materia de intersección que se ocupa de la designación de los conceptos de las lenguas de especialidad (Cabré, 1992). Un *lenguaje especializado* es un lenguaje que se usa en un campo del conocimiento y que se caracteriza por el uso de medios específicos de expresión lingüística (ISO 1087-1:2000, 2000). Por tanto, en otras palabras, la terminología, como disciplina, es una materia interdisciplinaria que se encarga de designar conceptos del lenguaje que se emplean en los campos del conocimiento y que tienen características específicas que las diferencian de la lengua general o cotidiana.

A lo largo de este tercer capítulo se abordará la terminología no solamente como disciplina sino también sus aplicaciones, su relación con el procesamiento de lenguaje y la aplicación de ambas materias en diversos sistemas y herramientas.

2.1 Terminología y terminografía

La *Terminología*, no sólo designa a una disciplina, sino también define el conjunto de unidades léxicas usadas con un valor preciso en los ámbitos de especialidad (Cabré, 1992). Es decir, todo el grupo de conceptos que la terminología, como disciplina, designa. Tomando en cuenta lo anterior, para Cabré (1992), existen cuatro puntos que muestran los distintos enfoques sobre el estudio y la práctica de la terminología:

- Para los lingüistas, la terminología es una parte del léxico delimitada por criterios temáticos y pragmáticos.
- Para los especialistas, la terminología es el reflejo formal de la organización conceptual de una especialidad, y un medio inevitable de expresión y de comunicación profesional.
- Para los usuarios (directos e intermediarios), la terminología es un conjunto de unidades de comunicación, útiles y prácticas, cuyo valor se mide en función de criterios de economía, de precisión y de adecuación.

- Para los planificadores lingüísticos, la terminología es un ámbito del lenguaje donde se debe intervenir para reafirmar la existencia, la utilidad y la pervivencia de una lengua, y para garantizar, mediante su modernización, su continuidad como medio de expresión.

2.1.1 Los términos

Una *unidad terminológica*, o *término*, es un símbolo convencional que representanta una noción definida en un cierto dominio del saber (Lérat, 1989). La unión de varios términos, forman la terminología del dominio de especialidad.

Existen distintos tipos de términos, estos se suelen clasificar de distinta manera, en torno a cuatro aspectos que son forma, función, significado y procedencia (Cabré, 1992).

El aspecto de forma es un conjunto de criterios que no son necesariamente excluyentes y que expresan la manera en que un término puede estar conformado. Estos criterios son los siguientes:

- **Número de morfemas⁷:** Dependiendo del número de morfemas un término puede ser simple o complejo. Ejemplo: *cuadern-o*, *cuadern-os*, *en-cuadern-ado*.
- **Tipos de morfemas:** Los distintos tipos de morfemas existentes en un término complejo determinan si es derivado o compuesto. Ejemplos de derivados son *fruter-ía*, *libr-ero*, *verd-oso*. En cambio, algunos ejemplos de términos compuestos son *para-brisas*, *saca-corchos*, *balon-cesto*.
- **Estructura:** Existen términos complejos que son la combinación de palabras que siguen una determinada estructura sintáctica. Algunos ejemplos de estructuras que se emplean en el español son sustantivo-preposición-sustantivo (método de Newton-Raphson), sustantivo-adjetivo (cristal líquido).

⁷ Según el diccionario de la Real Academia Española un morfema es la unidad mínima analizable que posee sólo significado gramatical. En otras palabras es la parte variante de la palabra que otorga un significado y permite formar nuevas palabras. Ejemplo: *niñ-o*, *niñ-a*, *niñ-os*, *niñ-as*

- **Origen complejo:** En algunos casos los términos simples provienen de términos complejos; casos de este criterio son las abreviaturas (Del., av.), las siglas (SIDA, ONU), acrónimos (bit, sonar) o formas abreviadas (tele, cine).

El segundo aspecto existente es el de función, es decir, los términos siempre tienen una función determinada en las oraciones. Estas funciones pueden ser de nombres, adjetivos, verbos y adverbios. En el caso de las palabras funcionales, como las preposiciones, conjunciones, artículos, entre otros, Cabré (1992) indica que no tienen un carácter terminológico.

El aspecto siguiente es el de significado, el cual indica que un término denomina una determinada clase de conceptos. Para Cabré (1992) se pueden establecer cuatro grandes clases conceptuales que son las siguientes:

- Objetos o entidades: Nombres.
- Procesos, operaciones o acciones: Verbos, nominalizaciones de verbos⁸.
- Propiedades, estados, cualidades: Adjetivos.
- Relaciones: Adjetivos, verbos.

El último aspecto que es mencionado por Cabré es el de procedencia lingüística, es decir, los términos pueden ser creados o contruidos a partir de reglas del propio lenguaje o provenir de otras lenguas.

2.1.2 La terminografía

La *terminografía* es la rama aplicada de la terminología que se ocupa de la elaboración de diccionarios especializados o de glosarios terminológicos (Cabré, 1995). Esta tarea incluye además la compilación, la sistematización y la presentación de los términos de las áreas de especialización.

⁸ Es el proceso de convertir un verbo en un sustantivo, por ejemplo gotear goteo.

Aunque la tarea de la terminografía es similar al de la lexicografía⁹ (el de crear diccionarios y glosarios), estas dos tareas difieren en el método que emplean, la forma en que emplean los datos y la manera en que presentan los resultados.

Mientras que la lexicografía sigue un proceso semasiológico, es decir, a partir del término crea la definición; la terminografía parte de la definición o de una lista de conceptos para determinar su término (que corresponda a la forma en que se emplea en el área especializada), es decir, sigue un proceso onomasiológico.

De igual forma, dentro del proceso de la terminografía se lleva a cabo una normalización, esto quiere decir que se busca estandarizar los términos que se emplean dentro de un área especializada para conseguir una comunicación profesional precisa, moderna y unívoca (Cabré, 1995).

El proceso de la terminografía está conformado por seis fases que son las siguientes (Cabré, 1992):

- **Definición y delimitación del trabajo:** En esta primera fase se debe definir el tema a trabajar, cuál es el público al que va dirigido, cuál es la función que va a tener el trabajo y el alcance de la obra en función de las condiciones anteriores, pero también de las económicas, temporales, materiales, académicas, entre otras.
- **Preparación del trabajo:** Consiste en adquirir y reunir toda la información sobre el tema a trabajar, en la selección de asesores de trabajo, en la estructuración que se va emplear y en la propuesta del plan de trabajo.
- **Elaboración de la terminología:** En la tercera fase de la terminografía se localizan los términos en el corpus y se determina que pertenezcan al área analizada.
- **Presentación del trabajo:** En esta fase se crea la publicación que contendrá el trabajo realizado en las etapas anteriores.

⁹ Es la rama aplicada de la lexicología. Según la RAE la lexicología es el estudio de las unidades léxicas de una lengua y de las relaciones sistemáticas que se establecen entre ellas.

- **Supervisión del trabajo:** Durante esta fase se juntan los expertos en terminología y los del área determinada para supervisar que el trabajo realizado no tenga problemas y sea el adecuado.
- **Tratamiento y resolución de los casos problemáticos:** Si existen casos problemáticos es necesario resolverlos; para ello se emplean diversos caminos dependiendo del caso, como consultar bibliografía complementaria, consultar a especialistas en la materia, lexicógrafos, especialistas multilingües o consultar a organismos oficiales de normalización.

2.1.3 Extracción de información terminológica

El desarrollo de nuevas materias de investigación y aplicación, como la informática o las ciencias computacionales, y su incursión dentro de diversas áreas, han hecho que muchas materias de investigación cambien su metodología, planteamiento o rendimiento. La terminología no es la excepción, ya que en la actualidad existe la terminótica. Para Cabré (1992) la *terminótica* es la materia que se ocupa, en general, de las relaciones entre la informática y la terminología; y, en particular, que trata de la aplicación de la informática al trabajo terminológico.

Esta incursión de la informática en el área de la terminología, de manera más específica en la terminografía ha adquirido cierto protagonismo en algunas de las tareas que se llevan a cabo en la metodología, como la documentación previa, la constitución del corpus, la verificación de la información, entre otras tareas. Pero también la extracción de términos ha sido una de las tareas donde la informática, específicamente el PLN, participa activamente por medio de la extracción de información, esto ha desarrollado la *extracción de información terminológica*, *extracción terminológica* o *terminology extraction (TE)*.

La extracción de información terminológica es el uso de métodos propios de la extracción de información con el objetivo de extraer los términos de un corpus apoyándose en el poder de procesamiento de las computadoras.

Cabe destacar que la extracción terminológica está altamente relacionada con la recuperación de información, no solamente porque la extracción de información está relacionada con esa tarea, sino por que frecuentemente los términos (empleando su sentido de

la búsqueda de información) que indizan los documentos son los términos (en su sentido lingüístico) que conforman a un documento. La única diferencia es que la extracción terminológica busca obtener todas las unidades terminológicas y no sólo las más representativas de un documento. Por tanto, son constantemente empleadas técnicas que en un principio eran solamente de indización de documentos en sistemas de extracción de terminología.

2.2 Sistemas actuales de extracción terminológica

Según Cabré et al. (2001) desde el 2000 los lingüistas computacionales, los investigadores en lingüística aplicada, traductores, intérpretes, periodistas, científicos e ingenieros en computación han estado interesados en el aislamiento automático de la terminología de textos. La razón de ello es que la terminología no sólo sirve para crear diccionarios o glosarios, también es útil en la traducción automática, en el resumen automático, en bases de conocimiento, en sistemas expertos, entre otras tareas.

Por lo anterior se han desarrollado sistemas que extraigan de manera automática la terminología de grandes cantidades de texto, de una manera rápida. Sin embargo, con el paso del tiempo los desarrolladores de los sistemas de extracción terminológica han observado que existen diversas complicaciones la cuales, según Cabré et al. (2001), son las siguientes:

- Identificación de términos complejos, es decir, se necesita reconocer cuándo una unidad discursiva¹⁰ constituye una frase terminológica y dónde comienza y termina ésta.
- Identificación de la naturaleza terminológica de una unidad léxica¹¹, esto es, conocer cuando dentro de un texto especializado una unidad léxica tiene una naturaleza terminológica o pertenece al lenguaje general.
- La propiedad y conveniencia de una unidad terminológica en un vocabulario dado.

¹⁰ Una unidad discursiva es una estructura que puede ser identificable dentro de un texto (<http://linguistics-ontology.org/gold/DiscourseUnit>).

¹¹ Una unidad léxica es un elemento que es objeto de definición en un diccionario, vocabulario, glosario, etcétera (Luna Trail et al., 2005).

Los sistemas de extracción terminológica se basan en tres tipos de conocimientos que son los lingüísticos, los estadísticos y los híbridos. Cada uno de estos tipos de sistemas se explicará en los apartados siguientes, además de que se darán a conocer algunos sistemas de extracción terminológica.

2.2.1 Sistemas basados en conocimiento lingüístico

Como se indicó en el apartado anterior, los sistemas de extracción terminológica se basan en distintos tipos de conocimiento y uno de ellos es el lingüístico; su razón de uso es porque la terminología y los términos están ampliamente relacionados con la lingüística.

Para Pazienza et al. (2005) los sistemas con un acercamiento lingüístico tratan de identificar términos a través de sus propiedades sintácticas, esto se debe a que frecuentemente las unidades terminológicas tienen estructuras sintácticas definidas, como se vio en la sección 2.1.1. Estos sistemas se pueden basar en dos tipos de información (Cabré et al., 2001):

- **Término específico:** Este consiste en la detección de patrones recurrentes de unidades terminológicas complejas; en la Tabla 5 podemos ver algunas estructuras empleadas en el español que definen por lo general un término; en cambio en la Tabla 6 podemos observar algunas estructuras sintácticas que por lo general no forman un término. Los patrones que se buscan provienen de reglas que se obtienen de manera empírica a través del análisis de datos y se pueden programar a través de expresiones regulares o autómatas de estados finitos.
- **Lenguaje genérico:** Consiste en la detección de estructuras lingüísticas más básicas, como los sintagmas¹² nominales (por ejemplo: libro, campo de trigo), sintagmas preposicionales (de María, para cocinar), entre otros. Para ello se emplean herramientas de PLN complejas, como son los *analizadores sintácticos*, también conocidos como *parsers*, que son herramientas que analizan la estructura de un texto con base en una gramática.

¹² Un sintagma, según la Real Academia Española, es un conjunto de palabras. Por ejemplo: un sintagma nominal está construido en torno a un nombre o sustantivo. En cambio, uno preposicional, es el formado alrededor de una preposición.

Estructura sintáctica	Ejemplos
sustantivo	agua, planeta, protozooario, cimientto
sustantivo + adjetivo	plano inclinado, agua oxigenada
sustantivo + preposición + sustantivo	lámpara de halógeno, dióxido de carbono

Tabla 5. Ejemplos de estructuras sintácticas para términos en español

Estructura sintáctica	Ejemplos
artículo + sustantivo	la casa, el niño, los países
sustantivo + y/o + sustantivo	águila o sol, coseno y tangente

Tabla 6. Ejemplos de estructuras sintácticas que no forman por lo general términos en español

Los tipos de información explicados anteriormente se basan en el análisis morfológico.

Los sistemas terminológicos basados en conocimiento lingüístico tienen como ventaja que encuentran términos sin importar su frecuencia o importancia en el texto, pues se basan en su estructura. En cambio, su desventaja, es que son propensos al *ruido*, es decir, los sistemas son proclives a encontrar estructuras falsas debido a errores en la asignación de la categoría gramatical (análisis morfológico); de igual manera, los sistemas basados en conocimiento lingüístico son dependientes de la lengua, ya que las reglas generadas pueden no servir en otras lenguas.

2.2.1.1 LEXTER

El sistema de extracción de términos LEXTER (Bourigault, 1994) fue desarrollado para el francés basándose en conocimiento lingüístico; su objetivo principal era mejorar el sistema de indización de la compañía EDF (Electricité de France).

El principio básico de LEXTER es encontrar las fronteras de los sintagmas nominales, pero en lugar de hacerlo de manera “positiva”, es decir, encontrando las estructuras que emplean los términos frecuentemente en francés, se realizó de manera “negativa”, en otras palabras, era encontrar estructuras sintácticas que claramente no formaran un término.

La primera tarea que realiza LEXTER es un análisis morfológico y de desambiguación para cada uno de los textos que se va a analizar. Posteriormente, el sistema busca, dentro del texto preprocesado, patrones que no sean parte de un sintagma nominal y por tanto, de un término. Algunos casos según Bourigault et al. (1996) son verbos, pronombres, preposiciones unidos a artículos posesivos, entre otros. Este proceso deja secuencias de palabras que por lo general corresponden a sintagmas nominales y son candidatos a ser términos o partes de ellos son candidatos; a este conjunto de palabras le llamaron MLNP (Maximal-Length Noun Phrases).

La segunda tarea consiste en un analizador sintáctico que analiza los MLNP para dividir candidatos terminológicos complejos en partes más sencillas llamadas cabeza (head, H) y expansión (expansion, E). El módulo del analizador sintáctico se basa en reglas, que indican qué partes son la cabeza y qué partes son la expansión del MLNP; en caso de encontrar estructuras ambiguas, existe un algoritmo de desambiguación que ejecuta distintas formas de una regla si se hallan formaciones en la estructura ambigua que ya hubieran sido encontradas durante el análisis. A continuación, en la Tabla 7 se muestra una regla no ambigua, mientras que en la Tabla 8 se ejemplifica otra donde se presentan casos de ambigüedad.

Regla no ambigua
<i>sustantivo₁ + adj + prep + sustantivo₂</i>
→
Cabeza: <i>sustantivo₁ + adj</i>
Cabeza: <i>sustantivo₁</i>
Extensión: <i>adj</i>
Extensión: <i>sustantivo₂</i>

Tabla 7. Ejemplo de una regla no ambigua empleada en LEXTER

Regla ambigua	
<i>sustantivo₁ + prep + sustantivo₂ + adj</i>	
Caso 1	Caso 2
→	→
Cabeza: <i>sustantivo₁</i>	Cabeza: <i>sustantivo₁ + prep + sustantivo₂</i>
Extensión: <i>sustantivo₂ adj</i>	Cabeza: <i>sustantivo₁</i>
Cabeza: <i>sustantivo₂</i>	Extensión: : <i>sustantivo₂</i>
Extensión: <i>adj</i>	Extensión: <i>adj</i>

Tabla 8. Ejemplo de una regla ambigua empleada en LEXTER

La tercera parte del proceso es un módulo de estructuración que emplea la información dada por el paso anterior para crear una red terminológica. Este consiste en vincular las cabezas y extensiones de términos complejos con términos menos complejos, y estos, a su vez, vincularlos con términos todavía menos complejos hasta formar una red. En la Figura 6 se muestra un ejemplo¹³ de la red terminológica generada por LEXTER.

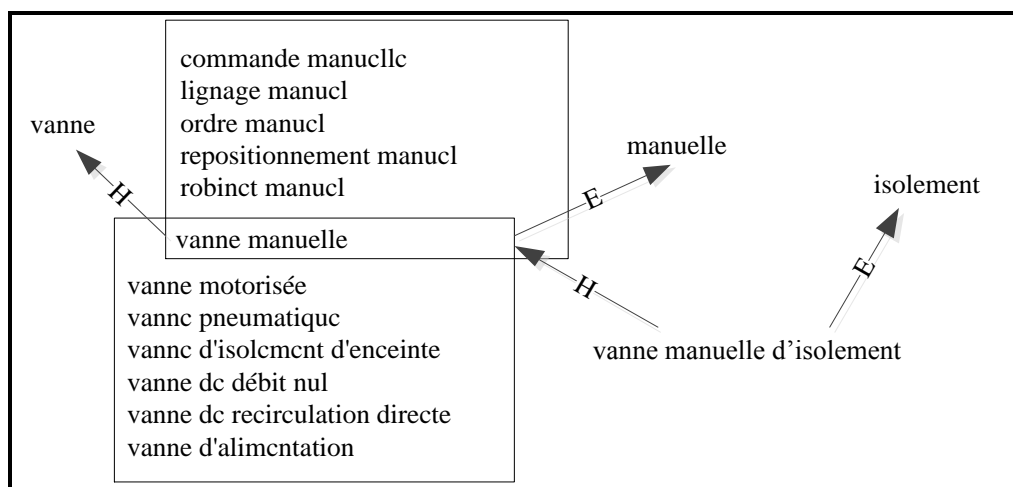


Figura 6. Ejemplo de una red terminológica creada por LEXTER

Al sistema extractor de términos LEXTER se le considera un sistema robusto, preciso e independiente del dominio desarrollado para el idioma francés. Sin embargo, LEXTER tiene algunos problemas de ruido por errores en el análisis morfológico, como ocurre en la mayoría de los sistemas basados en conocimiento lingüístico. Aun así, se le considera a este

¹³ Ejemplo extraído de Bourigault et al. (1996)

extractor de términos un buen sistema por su habilidad de aprender conforme se van obteniendo unidades terminológicas.

2.2.1.2 HEID

HEID (Heid et al., 1996) es un sistema de extracción terminológica que se basa en conocimiento lingüístico para el idioma alemán. Su objetivo es aumentar la eficiencia del proceso de creación de glosarios en tareas relacionadas con la traducción de textos técnicos, en este caso de ingeniería automovilística.

El sistema de extracción está compuesto de dos partes, la primera de ellas es el análisis lingüístico y la anotación de los textos; la segunda es la extracción de términos por medio de consultas en el corpus.

El análisis lingüístico consiste en un tokenizador, un analizador morfosintáctico¹⁴, un etiquetador POS¹⁵ y un lematizador que se ejecutan al inicio del análisis. Posteriormente se extraen construcciones características de los sintagmas nominales, esto se debe a que no existía en el momento del desarrollo del extractor terminológico un analizador sintáctico de cobertura amplia para el alemán que pudiera extraer de manera total sintagmas nominales.

La extracción de términos está conformada por tres componentes principales:

- **Procesador de consultas de corpus general (CPQ):** Es un procesador que puede soportar expresiones complejas de consultas, como expresiones regulares, etiquetas POS, lemas, entre otras.
- **Macroprocesador para el lenguaje de consulta CPQ:** La extracción de términos en HEID se basa en listas de afijos y en la verificación de los contextos típicos de los candidatos a término (Heid et al., 1996); para llevar a cabo este proceso, dado un parámetro en consulta, ejecuta este en un gran número de palabras mientras mantiene los demás parámetros de la consulta iguales.

¹⁴ Identifica las categorías gramaticales, morfosintácticas y características distribucionales (Heid et al., 1996)

¹⁵ Es un etiquetador de partes de la oración, el cual según Heid et al. (1996) desambigua los casos identificados en el proceso morfosintáctico.

- **XKWIC:** Esta herramienta gráfica muestra los términos y sus concordancias¹⁶; también permite ordenar de manera automática el material extraído según las necesidades del usuario.

El extractor terminológico HEID fue evaluado empleando manuales de mantenimiento en alemán. Se buscó extraer principalmente términos monopalabra, que frecuentemente representan sintagmas nominales en alemán; en este tipo de casos se obtuvieron algunos problemas por ruido los cuales, según los desarrolladores, pueden ser eliminados con el uso de filtros (por frecuencia, por categoría gramatical, entre otros). Asimismo, HEID permite extraer colocaciones¹⁷ combinando sustantivos y verbos, aunque, en este caso los resultados no son muy buenos.

2.2.2 Sistemas basados en conocimiento estadístico

Además de los sistemas basados en conocimiento lingüístico, existen aquellos que se basan en conocimiento estadístico, es decir, en el empleo de fórmulas matemáticas, modelos probabilísticos, modelos heurísticos, entre otros.

Estos sistemas, además de extraer términos, otorgan una calificación que permite clasificar los resultados en buenos o malos. Aunque lo anterior es algo ambiguo, lo que se busca es que los términos extraídos con una alta calificación expresen una mayor relevancia en el documento o corpus, mientras que uno con baja calificación indique lo opuesto.

Existen múltiples medidas estadísticas que se emplean en los extractores terminológicos, como el TF-IDF, el logaritmo de la verosimilitud (Log Likelihood), el T-score, entre otros.

La ventaja de estos sistemas de extracción es que son independientes de la lengua e indican una calificación para cada uno de los términos. El problema con este tipo de enfoque

¹⁶ Las concordancias, según la Real Academia Española (RAE), es el índice de todas las palabras de un libro o del conjunto de la obra de un autor, con todas las citas de los lugares en que se hallan.

¹⁷ Propiedad que tienen ciertos sustantivos y verbos, y algunos sustantivos y adjetivos de coincidir en estructuras sintagmáticas, gracias a su estructura semántica: *gato* y *ronronear*, *planta* y *marchita* (Luna Trail et al., 2005).

es que existen términos de baja frecuencia difíciles de manejar por los sistemas de extracción (Cabré et al., 2001), esto genera lo que se llama *silencio*.

2.2.2.1 ANA

El sistema ANA (Euguehard y Pantera, 1994), “Automatic Natural Acquisition”, es un extractor terminológico basado en conocimiento estadístico. Se basó en la idea de que este sistema debía poder extraer los términos de cualquier texto, sin importar si estaba bien escrito o no, si eran textos escritos o transcripciones de conversaciones y sin la utilización de conocimiento lingüístico. El extractor estaba diseñado para funcionar con cualquier lengua europea que no fuera aglutinante; sus pruebas se basaron en el inglés y el francés.

El sistema está formado por dos módulos: el de familiarización y el de descubrimiento. El primero de estos determina tres listas que emplea como conocimiento de la lengua a analizar; este conocimiento es extraído de manera estadística sin el uso de diccionarios o gramáticas. Las listas empleadas como conocimiento son las siguientes:

- **Palabras funcionales:** Es un conjunto de palabras que aportan poco o ninguna información (Sección 1.1.6). En esta lista entran artículos, pronombres y algunos verbos recurrentes.
- **Palabras esquemáticas:** Son las palabras que establecen una relación semántica entre otras palabras. Por ejemplo, Euguehard y Pantera (1994) indican que en el fragmento “box of nails”, la palabra “of” indica una cierta relación entre “box” y “nails”, por lo tanto “of” es una palabra esquemática.
- **Palabras base (bootstrap):** Es el conjunto de términos base con el que se inicia el sistema, es decir, este grupo de unidades terminológicas es el núcleo del extractor terminológico ANA.

El segundo módulo que conforma ANA es el de descubrimiento y se basa en la adquisición de nuevos términos a través del descubrimiento, como lo hace una persona que aprende un idioma. Este proceso se apoya en la co-ocurrencia de las palabras, esto puede tener tres interpretaciones:

- **Expresiones:** Una expresión se genera y se agrega a la lista de términos (bootstrap) cuando dos términos co-ocurren frecuentemente, es decir, aparecen en estructuras similares. Por ejemplo, en las frases “the *diesel engine* is”, “this *diesel engine* has”, los términos “diesel” y “engine”, que pertenecen al bootstrap, aparecen contiguos frecuentemente, por lo tanto es posible que “diesel engine” sea un término y se agrega a la lista de palabras base.
- **Candidato:** Cuando una palabra, llamémosla X, aparece seguidamente de una palabra esquemática y de términos pertenecientes al bootstrap, se le considera como un candidato a término y se agrega a la lista de palabras base. Ejemplo: en las frases “shade of wood”, “shade of color”, “shade of beech”, donde “of” es una palabra esquemática y las palabras “wood”, “color” y “beech” son términos, la palabra “shade” cumple con la interpretación de candidato.
- **Expansión:** Este caso es similar al anterior, la diferencia es que no existe ninguna palabra esquemática entre el término y la palabra X. Un ejemplo sería: “use any *soft woods* to”, “this *soft woods* or”, donde “wood”¹⁸ es un término, por tanto la palabra “soft wood” se agregaría al conjunto de términos.

El proceso del módulo 2 se realiza de manera recursiva hasta que no se encuentre ningún término nuevo en el documento. Además, durante el proceso de descubrimiento se genera una red semántica, en el cual se muestran algunas relaciones morfológicas y las co-ocurrencias de los términos.

Con respecto a los resultados, el sistema ANA fue evaluado para el inglés y para el francés. En el caso del inglés se empleó un corpus de 25,000 palabras el cual no fue ejecutado en el módulo de familiarización, sino que se indicaron de manera manual cada una de las listas que se crean en este módulo por el pequeño tamaño que tenía el corpus; del uso de ANA en este corpus se obtuvieron 200 nuevos términos. Para el francés, en cambio, se

¹⁸ Aunque la palabra en el ejemplo es “woods” y el término es “wood”, ANA reconoce que son la misma palabra debido a que emplea una herramienta que llama Reconocimiento Flexible de Cadenas. Esta herramienta emplea la distancia de edición; por ejemplo, si se tiene “casa” y “casas” su distancia de edición es 1 (adición de una s), en cambio para “caza” y “casa” es de 2 (eliminación de z y adición de s); por tanto dos palabras se parecen si su distancia de edición es muy pequeña.

usó un corpus de 120,000 palabras el cual sí pasó por el módulo de familiarización; del proceso de extracción se obtuvieron más de 3,000 nuevos términos.

A pesar de los resultados obtenidos, los desarrolladores de ANA consideran que este sistema es un extractor terminológico especializado en corpus de gran tamaño pero que sean de mala calidad, ya que aprende sobre la lengua empleada.

2.2.2.2 Extractor de términos estadístico basado en corpus

Este extractor terminológico fue desarrollado por Pantel y Lin (2001) y se basa únicamente en conocimiento estadístico.

El extractor terminológico consta de dos partes; la primera consta de la extracción de candidatos de términos. Para ello primero se recuperan todas los bigramas que se encuentren en el texto y su frecuencia; esta información se almacena en una base de datos de proximidad¹⁹. Posteriormente, se eliminan los bigramas que no cumplen con una serie de valores que están relacionados con la frecuencia del bigrama, con el valor de información mutua entre bigramas adyacentes²⁰ y el valor del logaritmo de la verosimilitud entre las palabras que pertenecen a un mismo bigrama²¹.

La segunda parte del extractor consiste en la extracción de términos multipalabra; en esta parte se realiza la extracción de todas las construcciones que puede tener un bigrama (extraído en el paso anterior) con sus palabras adyacentes, esto para obtener términos que sean más grandes que bigramas; de este proceso sólo se guardan las palabras adyacentes que aparecieron en una misma construcción con el bigrama en cuestión varias veces. En seguida, la base de datos de proximidad se actualiza con el bigrama formado por una palabra del término original y por la de la nueva palabra que se encontró en la construcción. Finalmente,

¹⁹ Una base de datos de proximidad es una base de datos con dos tablas; en la primera se almacena el objeto o el registro, mientras que en la segunda se guardan vínculos; cada tabla además tiene algunos atributos, como el nombre o valor (<http://c2.com/cgi/wiki/Wiki?ProximityDatabase>; <http://kdl.cs.umass.edu/software/about.html>).

²⁰ Esto se lleva a cabo para eliminar bigramas que tengan una palabra que no esté altamente relacionada con un posible término.

²¹ Esto se realiza para saber si las palabras dentro del bigrama están por casualidad o por una verdadera importancia.

el proceso se vuelve recursivo y se emplea la nueva información que se obtuvo en la base de datos de proximidad, para que se pueda extender un término y obtener sus variantes.

Este sistema de extracción terminológica se evaluó usando precisión y cobertura usando un corpus segmentado en el idioma chino, la razón de lo anterior es que dicen los desarrolladores del sistema que el detectar palabras en chino es similar a detectar frases en inglés. La precisión fue evaluada contra los valores que se obtuvieron del logaritmo de la verosimilitud, mientras que la cobertura contra la frecuencia mínima de las palabras. Este sistema de extracción terminológica obtuvo una precisión máxima de 74.4% y una cobertura del 62.3%

2.2.3 Sistemas basados en conocimiento híbrido

Los sistemas de extracción terminológica no sólo pueden estar basados en un tipo de conocimiento; pueden emplear tanto el lingüístico como el estadístico, de esta manera se forma un sistema con conocimiento híbrido. El objetivo de este tipo de extractores terminológicos es crear sistemas que aprovechen al máximo las ventajas tanto de la parte lingüística como de la estadística y disminuyan las desventajas que ambos tienen.

2.2.3.1 *Termext*

Termext (Barrón-Cedeño et al., 2009) es un extractor terminológico de tipo híbrido que se basa en una adaptación para el español del método de C-Value/NC-Value (Frantzi et al., 2000). Además el método fue modificado para que aceptara unigramas como términos.

Este extractor de términos está dividido a grandes rasgos en dos partes, la de C-Value, y la de NC-Value. La primera parte, a su vez, se divide en dos procesos, el lingüístico y el estadístico. El proceso lingüístico consiste en etiquetar con partes de la oración y lematizar cada uno de los textos a analizar por medio de la herramienta TreeTagger. Posteriormente, dentro de este mismo proceso, se aplica un filtro lingüístico que consiste en almacenar las estructuras que pueden formar un término en español; este filtro puede ser abierto o cerrado, si es abierto este es más flexible con los patrones de los términos, de lo contrario es estricto con los patrones encontrados. En el proceso estadístico se calcula cuál es la probabilidad de que una estructura extraída sea un término; es decir, el C-Value, y para tal fin se toma en

cuenta la frecuencia de la estructura, la frecuencia de la estructura en estructuras más grandes, el número de ocurrencias de las estructuras más grandes anteriores y la longitud de la estructura.

La segunda parte que conforma a Termext es la del cálculo de NC-value. Este valor considera el contexto en el cual se encontraban los términos obtenidos en el proceso anterior, esto con base en que un término, por lo general, está rodeado de palabras que están altamente relacionadas y pueden ser un indicio que exprese qué tan representativo el término es o no. Para ello se obtienen las palabras que en el contexto del término tengan cierta relevancia y se les calcula un peso. Posteriormente, se calcula el NC-Value, usando estos pesos y el valor C-Value del término. Finalmente, los términos con valores más altos de NC-Value son los términos que son más importantes en el documento, mientras que los de menor valor, son términos no tan representativos.

El extractor Termext fue evaluado con precisión y cobertura cuatro veces, la primera de ella con un filtro abierto sin una lista de paro obtuvo 23% de precisión y 82.6% en cobertura. La segunda de evaluación fue con un filtro lingüístico abierto y con lista de paro, la cual tuvo una precisión de 26.5% y una cobertura de 79.4%. La tercera evaluación se llevó a cabo con un filtro cerrado sin lista de paro y la cuarta de ellas con un filtro cerrado y lista de paro, en precisión se obtuvo un 24% y 30.8% respectivamente mientras que en cobertura se alcanzó un 46.3% y 50.3% de manera respectiva. Además, para su uso, se indica que Termext obtiene los mejores resultados de precisión y cobertura cuando se emplea un corpus de carácter técnico o científico de alto nivel de especialización, de lo contrario se genera una gran cantidad de ruido.

2.2.3.2 YATE

YATE (Vivaldi, 2001) es un extractor terminológico que emplea conocimiento tanto estadístico como lingüístico. Permite extraer términos tanto en español como en catalán, en los dominios de medicina, economía y genética. Las principales características de YATE son dos: la primera es que emplea una combinación de varias técnicas de extracción de términos y la segunda, que usa EuroWordNet como recurso léxico principal; de este recurso se hablará más adelante en la sección 2.4.1.

Grosso modo, existen 3 procesos que conforman YATE, los cuales se explican a continuación:

- **Proceso lingüístico:** Este es el primer proceso del extractor YATE. En él se lleva a cabo la segmentación, un análisis morfológico y, finalmente, un etiquetado de partes de la oración. En este proceso se emplean recursos léxicos como diccionarios, EuroWordNet y un corpus de referencia.
- **Filtro lingüístico:** Este proceso filtra las construcciones sintácticas que tienden a generar términos ya sea en español o en catalán, dependiendo del texto analizado. De este proceso se obtienen los candidatos a término que serán utilizados en el siguiente proceso.
- **Analizador de candidatos a término:** Este es el último proceso que forma parte de YATE. En él se calculan las diversas métricas y los datos que emplea YATE para determinar si un candidato a término pertenece o no al dominio seleccionado. Algunos de sus módulos son los siguientes (Vivaldi et al., 2001):
 - **Sistema de combinación:** En este módulo se unen todos los resultados para crear la lista final de candidatos.
 - **Extractor de contenido semántico:** Este módulo emplea EuroWordNet para determinar cuándo una palabra dada pertenece al dominio analizado, empleando identificadores de dominio.
 - **Formas griegas y latinas:** En el vocabulario médico se emplean muchas palabras que contienen formas griegas y latinas; por lo tanto, el conocer los términos que contienen estas formas puede dar información útil.
 - **Análisis colocacional:** En este módulo se emplean algunas medidas estadísticas para clasificar los candidatos a término, como la información mutua y la información mutua cúbica (MI^3).

Para llevar a cabo la evaluación de YATE se empleó un corpus de 10,000 palabras que consistía en resúmenes de artículos médicos. Este sistema de extracción terminológica fue evaluado con las medidas de precisión y cobertura, donde obtuvo un 97.2% de exactitud para una cobertura del 30%.

2.3 Evaluación de los extractores terminológicos

Los sistemas de extracción terminológica, al igual que muchos otros sistemas realizados por el hombre, necesitan que se les evalúe, ya que se necesita ver que el sistema cumpla con los objetivos, funcione con los estándares adecuados y sea lo suficientemente bueno como para realizar la tarea de forma automática y no manual. Sin embargo, aun cuando la extracción y el reconocimiento automático de términos han sido trabajados por largo tiempo y desde diferentes perspectivas, ningún *gold standard*²² de evaluación ha sido introducido para evaluar claramente y comparar distintos enfoques (Pazienza et al., 2005).

Aun así, se han desarrollado dos técnicas para la evaluación de los extractores terminológicos y se presentan a continuación.

2.3.1 Lista de referencia

Uno de los métodos utilizados para la evaluación de los sistemas de extracción terminológica es el empleo de una lista de referencia. En este caso, según Pazienza et al. (2005), una lista de referencia se toma como un *gold standard*; esta puede ser una lista de términos ya existente de un dominio o área específica, o puede ser construida por un experto analizando el corpus que se empleó para extraer los términos.

Con la lista de referencia, el extractor terminológico se evalúa mediante el empleo de las métricas de precisión y de cobertura que se vieron en el apartado 1.2.3.

Aunque la lista de referencia tiene sus ventajas, para Pazienza et al. (2005), en términos de eficiencia, la lista de referencia no es la mejor técnica para calcular la precisión. Esto se debe a que puede haber términos reales que no fueron colocados en la lista y, por tanto, se consideran como falsos, disminuyendo la precisión del sistema.

²² Un *gold standard* o una prueba estándar es una prueba o punto de referencia que califica, en este sentido, un sistema; puede que esta prueba no sea la mejor, pero no existe alguna otra y cumple con los estándares más básicos (http://en.wikipedia.org/wiki/Gold_standard_%28test%29).

2.3.2 Validación

Otro de los métodos empleados para la evaluación de los extractores terminológicos es la validación. Este método es preferido cuando ningún gold standard está disponible o cuando algunas características particulares del proceso de extracción de términos tienen que ser explícitas (Pazienza et al., 2005).

Este método consiste en validar los términos que se encuentran en la lista creada por el sistema en evaluación. Para poder llevar esto a cabo, Pazienza et al. (2005) indican que es necesario que se cumplan dos cosas. La primera de ellas, es que la validación de la lista debe ser realizada por varios expertos, esto para tener una lista de términos mucho más confiable. El segundo parámetro a cumplir es que cada experto que va a participar en el análisis debe recibir una introducción a lo que es un término. De todas maneras, cabe aclarar que aun siguiendo estos dos parámetros, es posible que las listas resultantes sean diferentes, esto puede ser debido a los distintos conocimientos de los expertos, al juicio del experto o a la ambigüedad de lo que es una unidad terminológica; por tanto, es necesario que se llegue a un acuerdo entre los expertos para obtener una lista validada.

Con la lista de términos validada se emplean las métricas de precisión y de cobertura de la misma forma que ocurre en los sistemas de recuperación de información.

Al igual que la lista de referencia, este método de evaluación tiene sus desventajas, una de ellas es que no es el mejor método para calcular la cobertura del sistema. La razón de ello es que, al enfocarse en una lista extraída por el mismo sistema, se cierra la posibilidad de conocer si existen otros términos que se debieran haber obtenido.

2.4 Recursos electrónicos para la validación

Actualmente, existen algunos extractores terminológicos que validan cada uno de los términos encontrados en el documento antes de presentárselos al usuario; además algunos de ellos agregan información que podría ser de utilidad. Para ello emplean recursos semánticos, en su mayoría creados por expertos, que otorgan información sobre el dominio al que

pertenecen, como sinónimos. Algunos extractores que emplean este tipo de validación, además de YATE, son MetaMap (Aronson y Lang, 2010) y TRUCKS (Maynard, 2000).

2.4.1 WordNet y EuroWordNet

WordNet es una base de datos léxica electrónica desarrollada por la Universidad de Princeton, la cual sirve como recurso para aplicaciones en PLN y recuperación de información (Fellbaum, 1998). Esta base de datos sólo maneja inglés y es de acceso libre por internet²³. Su extensión a otros idiomas, como el español, se realizó por medio de *EuroWordNet* (EWN), que es de paga y actualmente está en crecimiento en algunas lenguas.

Dentro de WordNet y, por consiguiente, de EuroWordNet, existen tres estructuras que se encargan de las diversas categorías lingüísticas que maneja, es decir, hay una para sustantivos, otra para verbos y una para adjetivos y adverbios.

Esta base de datos se basa principalmente en conjuntos de sinónimos, llamados *synset*, que representan todo un concepto. Por ejemplo, en el caso del inglés, cuando se busca “elevator” también se muestra su variante británica que es “lift”; en el caso del español si buscamos “tepalcate” nos muestra que tiene como *synset* “tejoleta”, “tiesto” y “casco”.

La estructura de sustantivos, de WordNet y EWN, además de manejarse a través de los *synset*, se maneja por medio de relaciones de hiponimia e hiperonimia. La hiponimia es una relación que denota un subconjunto o subclase de una palabra; por ejemplo, en EWN la palabra “automóvil” tiene como hipónimos las palabras “limosina”, “sedán”, “jeep”, entre otros. En cambio, la hiperonimia es una relación que expresa una superclase de una palabra; “vivienda”, por ejemplo, es un hiperónimo de “casa”, de “estudio” y de algunos otros más.

WordNet y EWN, además de contar con los *synset*, incluye definiciones tipo diccionario y ejemplos de uso.

²³ <http://wordnetweb.princeton.edu/perl/webwn>

2.4.2 Lexicón Specialist UMLS

Uno de los recursos léxicos electrónicos más importantes del área de la biomedicina es el lexicón *Specialist de UMLS*. Este lexicón es uno de los tres recursos que se generaron dentro del proyecto UMLS (Unified Medical Language System) creado por la Biblioteca Nacional de Medicina de los Estados Unidos de América (NLM).

Según Ananiadou y McNaught (2006), el lexicón Specialist es un diccionario general del inglés que contiene una gran cantidad de términos de biomedicina. Todos estos términos fueron extraídos de diversos recursos, como de los registros de MEDLINE/PubMed²⁴, del metatesauro UMLS²⁵ y de diccionarios médicos del inglés.

Cada una de las entradas del lexicón puede ser monopalabra o multipalabra; a su vez, estos términos tienen información como categoría gramatical, patrones complementarios permitidos, lema, variantes ortográficas y morfológicas.

2.4.3 Wikipedia

Otro de los recursos que se han estado empleando actualmente para la validación de extractores es Wikipedia²⁶. La *Wikipedia* es una enciclopedia gratuita, multilingüaje, creada para la red y construida de manera colaborativa por voluntarios (Zesch et al., 2008).

Esta enciclopedia está formada por artículos que crean una red interconectada de conocimiento, adicionada con categorías y subcategorías (se podría decir que es un tipo de hiperonimia e hiponimia, aunque no cumplan forzosamente con las relaciones) que los voluntarios crean y organizan, y que permiten hasta cierto punto dividir los conocimientos en áreas o dominios. El uso de categorías y subcategorías forma lo que se conoce como una *taxonomía*, es decir una ordenación jerárquica y sistemática; aunque hay autores como Peters

²⁴ MEDLINE es una base de datos que almacena bibliografía médica que provienen desde 1950. Su motor de búsqueda es la herramienta de PubMed.

²⁵ Es otro de los recursos del proyecto de UMLS que incluye conceptos del área de biomedicina, nombres de conceptos, sinónimos, así como las relaciones entre los conceptos.

²⁶ <http://www.wikipedia.org>

(2009), que consideran esto realmente como una folksonomía²⁷, ya que es la gente quien desarrolla la jerarquización y sistematización de la Wikipedia.

Además Wikipedia contiene una gran cantidad de información semántica y léxica que se complementa con el conocimiento de entidades nombradas y términos de dominio específico o especializado que incluye el sitio. De igual forma, incluye un sistema de redireccionamiento, que podría ser considerado un diccionario de sinónimos en el cual se toman en cuenta variaciones ortográficas, morfológicas y de abreviaturas; por ejemplo, si se busca en la Wikipedia “ajolote”, “axolote” o “axolotl” se redirecciona a “*Ambystoma mexicanum*”, el nombre científico del ajolote. También el sistema de redireccionamiento funciona, en un menor grado, como un sistema que pasa de un tema específico a uno general, o de un verbo a un sustantivo.

Entre las ventajas con las que cuenta Wikipedia se puede mencionar que es un recurso libre, que se actualiza y crece rápidamente, que maneja una gran cantidad de dominios y que está en diversas lenguas, no solamente en las principales. Algunas de sus desventajas es que no existe un control editorial o por expertos, y que no se siguen lineamientos específicos para su construcción.

De este recurso electrónico se hablará más adelante, en la sección 3.4, donde se abordará la estructura interna y la manera en que fue empleada en el proyecto de tesis.

²⁷ Una folksonomía es un sistema de clasificación de contenidos desarrollado de manera colaborativa (Peters, 2009).

3 OBTENCIÓN AUTOMÁTICA DE TÉRMINOS Y SU VALIDACIÓN

La *obtención automática de términos* es el uso de sistemas de extracción de información terminológica para la recuperación de unidades terminológicas de un determinado corpus. La *validación* de estas unidades terminológicas es el proceso que consiste en aprobar o desaprobar su estatus de término y obtener como resultado listas de términos con una alta probabilidad de pertenecer a una determinada área de conocimiento.

A través de los capítulos anteriores se ha dado a conocer el marco teórico que permite establecer la obtención automática de términos y su validación. Asimismo, se han dado a conocer algunos de los extractores terminológicos más representativos que por el tipo de conocimiento que emplean, por su restricción temática o por su falta de validación de términos no pueden ser empleados en todos los casos para llevar a cabo una extracción terminológica. Por lo anterior, es que se decidió desarrollar un extractor terminológico que pueda ser empleado para el español y que la validación de términos no esté acotada a una sola área específica.

A lo largo de este capítulo se dará a conocer el corpus que se empleó para la extracción terminológica. Se explicará el método que se usó para la obtención de términos. Posteriormente, se indicará el procedimiento para la validación de las unidades terminológicas. Finalmente se dará a conocer la arquitectura del sistema desarrollado para esta tesis.

3.1 Corpus de textos científicos en español de México (COCIEM)

Como se vio en la sección 1.1.1, uno de los recursos empleados dentro de PLN son los corpus lingüísticos. En este proyecto se empleará un corpus informatizado llamado el Corpus de textos científicos en español de México (COCIEM).

Este corpus fue creado por la Dra. María Pozzi en el proyecto de investigación “El vocabulario básico científico en México: Una investigación de sus características, componentes y difusión”, patrocinado por el Consejo Nacional de Ciencia y Tecnología (CONACyT) con número de proyecto 58923.

La construcción del COCIEM se llevó a cabo para poder identificar y obtener el vocabulario básico científico mexicano, es decir, el vocabulario científico elemental que debería ser conocido por un hablante promedio al final de su educación media superior. Posteriormente, a partir del vocabulario obtenido, será posible crear diccionarios o glosarios, así como también emplearlo como recurso lingüístico en diversas aplicaciones de la ingeniería lingüística, por ejemplo para la traducción automática, la generación automática de texto, etcétera.

El COCIEM está conformado por los libros de texto de mayor uso a nivel nacional por los estudiantes de educación básica y media superior. Dentro de estos libros de texto se encuentran libros de teoría, de prácticas de laboratorio, de ejercicios, etc., esto para tener una representatividad del conocimiento pre-universitario.

Cada uno de los libros que pertenecen al COCIEM fue digitalizado y se almacenó en un archivo de texto plano.

En esta tesis, se consideró al COCIEM como un buen corpus para llevar a cabo la extracción terminológica por su gran variedad de materias y porque por ser de carácter educativo tendría que contener una gran cantidad de términos.

3.1.1 Estructura del COCIEM

El Corpus de textos científicos en español de México está conformado por 92 libros de texto divididos en tres niveles educativos: primaria, secundaria y bachillerato. A su vez, cada uno de los niveles educativos se encuentra dividido por año escolar.

Los 92 libros que forman parte del COCIEM son de materias que tienen un enfoque meramente científico, es decir, no son libros de asignaturas que estén relacionadas con las humanidades, como lo es la historia, la ética, la literatura o geografía.

En la Tabla 9 se muestra la estructura del COCIEM por nivel educativo y por materia en general.

Nivel educativo	Materia	Número de libros de texto	Número de tokens	Número de tipos
Primaria	Ciencias naturales	6	175,240	11,437
	Matemáticas	9	125,723	9,377
Total		15	300,963	20,814
Secundaria	Biología	8	369,099	23,908
	Matemáticas	24	734,374	55,797
	Física	9	538,042	29,759
	Química	8	382,697	23,031
	Educación ambiental	1	73,247	7,922
Total		50	2,097,459	140,417
Bachillerato	Biología	3	133,262	8,307
	Matemáticas	11	499,552	32,566
	Física	3	219,795	14,251
	Química	5	156,192	11,162
	Educación para la salud	3	139,369	9,421
	Ecología	2	124,799	7,731
Total		27	1,272,969	83,438
Gran total		92	3,671,391	244,669

Tabla 9. Estructura del COCIEM por niveles educativos y materias

Como se puede observar en la tabla anterior, el número de libros que conforman el corpus es diferente para cada uno de los niveles educativos y materias. Esto se contrapone con lo dicho en la sección 1.1.1, en el cual se indicaba que un corpus lingüístico debía ser lo más equilibrado posible, pero como indica Sierra (2008), en los corpus especializados es de esperarse que existan áreas o categorías con una mayor cantidad de textos que otras.

3.2 Preprocesamiento del COCIEM

Antes de llevar a cabo la extracción de términos dentro del COCIEM, cada uno de los textos recibió diversos tratamientos. Este conjunto de tareas son la revisión y limpieza de los

documentos, la lematización, la tokenización y la creación de n-gramas; estas tareas forman lo que se conoce como preprocesamiento.

3.2.1 Revisión, limpieza y adecuación de los documentos

La primera tarea dentro del preprocesamiento del COCIEM fue la revisión, limpieza y adecuación de cada uno de los documentos para las tareas que vendrían posteriormente, como la extracción terminológica.

Debido a que los libros del COCIEM fueron escaneados y se empleó un *reconocedor óptico de caracteres (OCR)*, existían casos en los cuales el reconocimiento de los caracteres no era el adecuado y por tanto existían errores. Por ello era necesario que se llevara a cabo una revisión, al menos a grandes rasgos, de los documentos y se eliminaran los errores existentes. Por ejemplo, un error concurrente dentro de los documentos del COCIEM era la aparición del símbolo de negación (\neg) entre los caracteres de una palabra.

De igual manera, para tratar de obtener términos solamente del texto y eliminar la posibilidad de encontrar cadenas muy largas unidas por punto o diagonales, se eliminaron correos electrónicos y páginas web de los documentos.

Además, cada uno de los textos, fueron guardados con la codificación UTF-8 para tener un manejo estándar entre las diversas codificaciones existentes.

3.2.2 Lematización usando FreeLing

La lematización de documentos es una de las tareas, ciertas veces esenciales, para poder llevar a cabo un procesamiento de lenguaje natural. Dada esta razón se decidió emplear un lematizador para obtener la forma canónica de cada una de las palabras de los documentos que pertenecían al COCIEM.

El objetivo de la lematización del COCIEM es la reducción y agrupamiento de los candidatos a término que se obtendrán más adelante, pero también obtener formas canónicas que, como se dijo anteriormente (sección 1.1.5), es la forma que se emplea en un diccionario. Por ejemplo, con la lematización “ecosistema” y “ecosistemas” se convierten en sólo “ecosistema” y, por consiguiente, dos candidatos a término se convierten en uno solo.

Para llevar esta tarea a cabo se empleó el lematizador FreeLing en su versión 2.2, el cual es una biblioteca de procesamiento de lenguaje multilingüe de código abierto que provee un amplio conjunto de analizadores del lenguaje para varios idiomas (Padró et al., 2010). Actualmente FreeLing soporta los idiomas español, inglés, catalán, gallego, galés, italiano, portugués y asturiano.

FreeLing emplea diversos recursos y módulos para llevar a cabo el procesamiento de lenguaje natural. Uno de los recursos que emplea es un diccionario que, en su versión para el idioma español, tiene más de 550,000 formas que corresponden a más de 76,000 lemas²⁸. Algunos de los módulos que forman parte de FreeLing son los siguientes (Padró et al., 2010):

- **Tokenizador:** Es un herramienta que recibe un texto plano y crea un archivo con los tokens encontrados.
- **Morfo:** Esta herramienta recibe oraciones e indica las posibles anotaciones morfosintácticas de cada una de las palabras de la oración. Dentro de este procesamiento se encuentran sufijos, números, fechas, cantidades (como razones, porcentajes, monedas), símbolos de puntuación, nombres propios, entre otros.
- **Etiquetador POS:** Recibe la información del módulo Morfo y desambigua las posibles anotaciones morfosintácticas que se indicaron para cada una de las palabras de las oraciones. Esto se lleva a cabo para obtener la etiqueta POS más probable con base en toda la información otorgada por el módulo Morfo.

Para la lematización del COCIEM empleando FreeLing, se desarrolló un programa que está conformado de tres módulos, en la Figura 7 se muestra su arquitectura. El primer módulo extrae un documento del directorio del corpus a analizar. El segundo módulo llama al programa de FreeLing para que lleve a cabo la lematización. Y el tercero cambia el formato de salida que otorga FreeLing a uno más entendible para el humano. En la Figura 8 se muestra un texto sin lematizar, mientras que en la Figura 9 se muestra un fragmento del texto anterior ya lematizado por FreeLing.

²⁸ Esta información fue extraída de la página oficial de FreeLing:

http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=23&Itemid=58

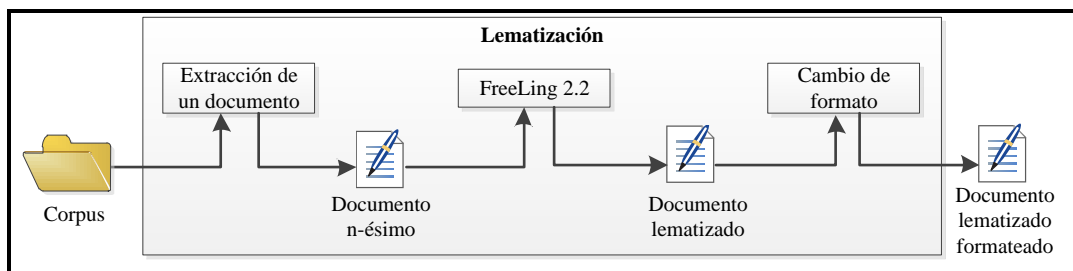


Figura 7. Arquitectura del programa de lematización

María Gaetana Agnesi (1718-1799) fue una matemática italiana que dejó varias contribuciones a la ciencia, entre ellas una curva muy famosa con un nombre sumamente peculiar. Este nombre tan peculiar fue por culpa de un traduttore traditore (traductor traidor) quien confundió la palabra versiera con avversiera que significa bruja. Es por ello que la curva de Agnesi se le conoce como Bruja de Agnesi, no sólo en el español sino en muchos idiomas más.

Figura 8. Un texto que servirá de ejemplo para llevar a cabo la lematización empleando FreeLing

<i>Este este</i> DDOMS0 0.956743	<i>traidor traidor</i> AQOMS0 0.509558
<i>nombre nombre</i> NCMS000 0.97973)) Fpt 1
<i>tan tan</i> RG 1	<i>quien quien</i> PROCS000 1
<i>peculiar peculiar</i> AQOCS0 1	<i>confundi3 confun3ir</i> VMIS3S0 1
<i>fue ser</i> VSIS3S0 0.932292	<i>la el</i> DA0FS0 0.972146
<i>por por</i> SPS00 1	<i>palabra palabra</i> NCFS000 1
<i>culpa culpa</i> NCFS000 0.866667	<i>versiera versiera</i> VMSI3S0 0.500172
<i>de de</i> SPS00 0.999919	<i>con con</i> SPS00 1
<i>un uno</i> DIOMS0 0.986987	<i>avversiera avversiera</i> VMSI3S0 0.500172
<i>traduttore traduttore</i> VMSP3S0 1	<i>que que</i> CS 0.4375
<i>traditore traditore</i> VMSP3S0 1	<i>significa significar</i> VMIP3S0 0.958333
((Fpa 1	<i>bruja brujo</i> NCFS000 0.6
<i>traductor traductor</i> NCMS000 0.490442	. . Fp 1

Figura 9. Extracto del archivo de salida generado por FreeLing después de lematizar el texto de la Figura 8

Como se puede observar en la Figura 9, es algo complicado revisar el texto lematizado en el formato de salida propio de FreeLing; es por ello que se desarrolló el tercer módulo, el cual se encarga de cambiar el formato de salida por uno que sea similar al formato original. Para ello, primeramente, se llevó a cabo un análisis de la presentación de salida que otorga FreeLing, el cual se puede observar en la Figura 10.

Palabra original	Lema	Etiqueta POS	Probabilidad
------------------	------	--------------	--------------

Figura 10. Formato de salida de la lematización realizada por FreeLing

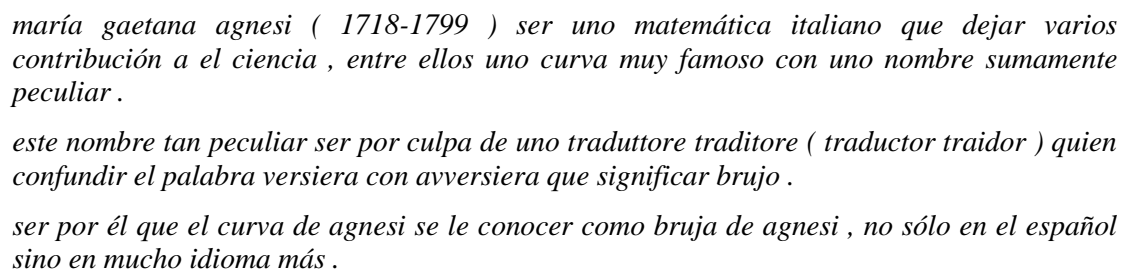
En el área de la palabra original se muestra la palabra o conjunto de ellas de la manera en que se encontraron en el texto. En el área de lema se muestra la forma canónica o un determinado formato en el caso de números, fechas, cantidades, entre otros; cabe aclarar que los lemas se presentan en letras minúsculas. En la etiqueta POS se emplea el formato propuesto por el grupo EAGLES (sección 1.1.4), el cual sirve para todas las lenguas europeas. Finalmente, en la probabilidad, se muestra la probabilidad de que la etiqueta POS elegida para la palabra original sea la correcta.

Para llevar a cabo el cambio de formato se desarrollaron las siguientes reglas:

- **Multipalabras:** Debido a que FreeLing concatena en algunos casos palabras para formar una multipalabra, como en los nombres propios, fue necesario crear una regla que cambiara los guiones bajos con los que une las palabras por espacios.
- **Diagonal (/):** La diagonal no solamente expresa un sentido matemático, en ocasiones permite unir palabras que se emplean de manera conjunta. Un ejemplo es el caso de VIH/SIDA el cual al obtener su forma canónica usando FreeLing se separa en tres elementos, “vih”, “/” y “sida”. Por tanto, se creó una regla que uniera estos tres elementos en el cambio de formato para obtener su estructura original.
- **Minúsculas:** A pesar de que el formato de salida del lema es en minúsculas, existen casos en el cual la primera letra del lema inicia con mayúsculas, esto debido a un error de FreeLing. Por ejemplo, en la frase “Lamentablemente lo que se pide en el examen no se puede realizar.”, la palabra “Lamentablemente” en su forma canónica aparece con mayúscula. Por ello fue necesario en el programa de cambio de formato reconvertir los lemas a minúscula por si ocurrían casos como éste y así mantener uniformemente el formato de salida.
- **Salto de línea:** Una de las características de FreeLing es que el análisis de cada una de las oraciones encontradas es separado por un salto de línea extra. Empleando lo anterior, se realizó una regla la cual indica cuándo colocar un salto de línea para indicar que una oración finalizó. En este caso, no se puede saber exactamente la

posición en la que se encontraba la oración en un párrafo, es por ello que no se pueden reconstruir estos como se encontraban en el documento original.

Empleando las reglas anteriores, se permite recuperar hasta cierto punto la estructura o ciertas construcciones que tenía el documento original. En la Figura 11 se muestra el cambio de formato realizado sobre el archivo de salida de FreeLing (Figura 9).



maría gaetana agnesi (1718-1799) ser uno matemática italiano que dejar varios contribución a el ciencia , entre ellos uno curva muy famoso con uno nombre sumamente peculiar .

este nombre tan peculiar ser por culpa de uno traduttore traditore (traductor traidor) quien confundir el palabra versiera con avversiera que significar brujo .

ser por él que el curva de agnesi se le conocer como bruja de agnesi , no sólo en el español sino en mucho idioma más .

Figura 11. Cambio de formato realizado a partir de la salida otorgada por FreeLing

Las razones de llevar el cambio de formato son permitir que se puedan revisar de una manera más sencilla los textos lematizados y se puedan corregir los errores si se desea.

3.2.3 Tokenización del COCIEM

La tokenización es el proceso que consiste en la segmentación de documentos en un conjunto de unidades con significado llamados tokens, como se pudo leer en la sección 1.1.2. Esta tarea forma parte del preprocesamiento que recibió el COCIEM.

Para realizar esta tarea se creó un tokenizador sencillo en Flex (Fast lexical analyzer)²⁹, que emplea la información otorgada por el programa desarrollado para lematizar el COCIEM (sección 3.2.2), más específicamente del módulo del cambio de formato. Para ello se crearon una serie de expresiones regulares con el objetivo de extraer las construcciones o tokens de los archivos lematizados del corpus y clasificarlos por su tipo (cadena, número, etcétera) al mismo tiempo; esto último se llevó a cabo para que se pueda realizar de forma paralela la creación de n-gramas, del cual se hablará en la siguiente sección (3.2.4).

²⁹ <http://flex.sourceforge.net/>

Entre las características que tiene este tokenizador están la detección de palabras unidas por guiones o de palabras unidas por una diagonal; por ejemplo óxido-reducción o ONU/UN. De igual forma, la detección de números, signos de puntuación o de cadenas compuestas por números como lo es H1N1, entre otros.

Asimismo, el tokenizador aprovecha la detección de abreviaturas por parte de FreeLing para tomarlas también como tokens.

3.2.4 Creación de n-gramas

Como se describió en la sección 1.1.3, los n-gramas son uniones de caracteres o de palabras. En el caso de esta tesis lo que se busca es obtener términos multipalabra y, por tanto, es necesario crear n-gramas de palabras y no de caracteres. Con base en la opinión de expertos en el área de terminología es posible obtener una base confiable para la obtención de unidades terminológicas con trigramas. La razón es que a partir de los trigramas ya se pueden visualizar términos comunes muy fácilmente sin emplear tantos recursos computacionales. Por lo tanto, el tamaño máximo de los n-gramas utilizados fue de 3 tokens.

Para llevar a cabo lo anterior, se desarrolló un programa en C que se basa en un ciclo de trabajo para la formación de n-gramas y que emplea la información otorgada por el tokenizador de la sección 3.2.3 para cada uno de los documentos del corpus. El ciclo de trabajo permite crear construcciones de n-gramas con base en los signos de puntuación, los saltos de línea, números y algunos otros marcadores tipográficos, como los paréntesis, las comillas y las llaves. En la Figura 12, se muestra un ejemplo en el cual se generan n-gramas a partir del texto de entrada “A B C D E F. G H I ¶ J K”; donde cada letra representa un token y el símbolo calderón (¶), el fin de un párrafo. Primeramente se genera un unigrama con A (t_1), luego al leer B (t_2) se crea el unigrama de B y el bigrama AB, posteriormente se genera el unigrama C (t_3) al mismo tiempo que el bigrama BC y el trigramas ABC; este proceso se lleva así sucesivamente hasta llegar a t_6 , donde se tiene el unigrama F, el bigrama EF y el trigramas DEF. Después, cuando se lee el punto se reinicia el ciclo de trabajo para comenzar desde el principio el ciclo de trabajo. La creación de n-gramas se lleva a cabo hasta que se termina de analizar todo el documento, de esta manera, se puede generar el número exacto de n-gramas siguiendo siempre la estructura del texto.

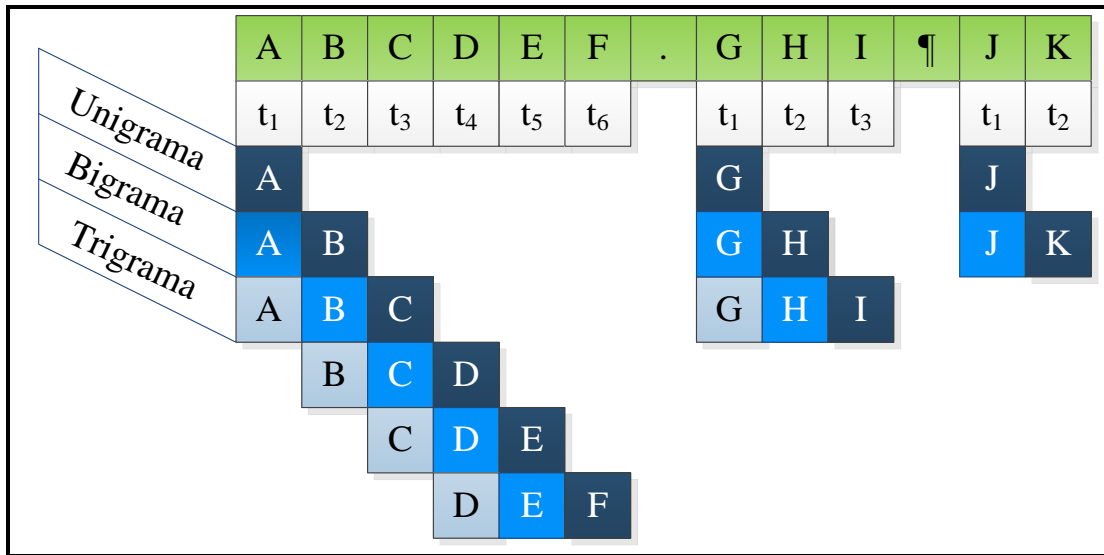


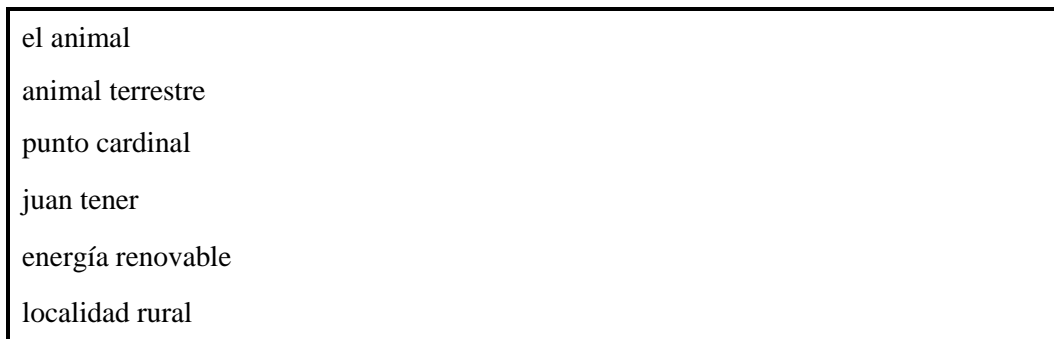
Figura 12. Ejemplo del ciclo de trabajo para la generación de n-gramas. Los cuadros superiores son el texto y el tiempo del ciclo.

El uso de los signos de puntuación, los saltos de línea y otros símbolos para reiniciar el ciclo de trabajo de la creación de n-gramas se basa en la idea de que en teoría ningún término, en el caso de uno multipalabra, tendrá en su interior un signo de puntuación que divida sus partes. De esta manera se pueden disminuir el número de n-gramas a analizar, ya que se reduce el número de formaciones sin sentido o que no existían en el texto original. Por ejemplo, en la frase lematizada “tapar el vaso de precipitado . llenar con agua el matraz erlenmeyer”, si no se reiniciara el ciclo con los signos de puntuación se tendría el bigrama “precipitado llenar”, aun cuando éste no existe.

Con respecto a los números, estos no son considerados términos, pues el sistema de numeración no es un término en sí, por tanto, con base en la clasificación que otorga el tokenizador desarrollado en la sección 3.2.3, estos se pueden omitir de manera sencilla en los n-gramas y reiniciar asimismo el ciclo de trabajo que los genera. Sin embargo, la única excepción de lo anterior es cuando los números se encuentran unidos por guiones a palabras o forman parte de una palabra, y por tanto de un posible término, por ejemplo el caso de carbono-14 o CH₄; en este caso, estos candidatos son considerados en la generación de n-gramas del ciclo de trabajo como si fueran una cadena de caracteres normal.

Además de la generación de n-gramas, una de las tareas de este módulo es dividir en archivos cada uno de los tipos de n-gramas (unigramas, bigramas y trigramas) de los

documentos del corpus analizado. Esto se realiza para que posteriormente se puedan analizar cada uno de los n-gramas y así extraer de ellos los candidatos a término. En la Figura 13 se muestra un ejemplo del formato de los archivos de salida del generador de n-gramas; en cada línea se encuentra un bigrama encontrado en un documento.



```
el animal
animal terrestre
punto cardinal
juan tener
energía renovable
localidad rural
```

Figura 13. Ejemplo de un archivo de n-gramas generado durante el preprocesamiento

3.3 Extracción de candidatos a término

La extracción de candidatos a término del COCIEM se lleva a cabo empleando el método de TF-IDF, el cual fue explicado en la sección 1.2.1. Este método consiste en la asignación de pesos para cada una de las palabras de un documento en corpus, empleando las métricas Term Frequency e Inverse Document Frequency.

Aunque en un principio el método de TF-IDF se empleaba para la creación de listas de palabras claves de sistemas de búsqueda de información, su uso se ha ido ampliando a áreas como la extracción terminológica automática. La razón de lo anterior es que las palabras clave de los sistemas de recuperación de información por lo general son candidatos a término del documento analizado. Por tanto, el método de TF-IDF puede emplearse como método para la extracción terminológica en corpus.

Las ventajas de emplear este método de extracción terminológica es su capacidad de ser independiente de la lengua, su rapidez y sencillez de implementación. De igual manera, TF-IDF tiene como ventaja que permite analizar a la vez una gran cantidad de información de manera paralela y obtener resultados separados, en otras palabras, todos el corpus del COCIEM puede ser analizado por libro al mismo tiempo sin ningún problema y generar una lista de términos por libro. En este caso, TF-IDF se diferencia de otros métodos de extracción

terminológica, como el logaritmo de la verosimilitud donde sólo se pueden comparar dos recursos a la vez, siendo uno de ellos el de referencia y otro el de análisis: un caso práctico es mostrado por Cabrera-Diego et al. (2011); otro caso es el del método C-Value/NC-value donde la extracción terminológica se lleva de manera general a todo un corpus o documento.

La desventaja del método de TF-IDF es que para llevarlo a cabo es necesario que existan al menos dos corpus o documentos, de lo contrario se obtienen pesos iguales a cero, aunque para este caso no es ningún problema por la gran cantidad de libros existentes en el COCIEM. Por ello este método es el adecuado para la extracción terminológica de términos del corpus seleccionado.

En el caso del COCIEM, el método de TF-IDF se aplica a unigramas, bigramas y trigramas, los cuales son generados empleando el método descrito en la sección 3.2.4.

A continuación se mostrarán los algoritmos empleados para llevar a cabo el análisis de TF-IDF y de algunas tareas extras que se realizaron para obtener los candidatos a término.

3.3.1 Cálculo de TF

El primer paso para realizar el método de TF-IDF es el cálculo de la frecuencia relativa de cada uno de los términos (TF) de todos los documentos a analizar. Para ello se creó un programa en C que emplea como entrada cada uno de los tipos de archivos de n-gramas (unigramas, bigramas, trigramas) que se generaron en la sección 3.2.4; su algoritmo se muestra en la Figura 14.

Cálculo de TF	
1.	Se abre el archivo de n-gramas
2.	Se ordena el archivo de n-gramas de manera alfabética
3.	Se crea un archivo de salida (S)
4.	Se lee la primera línea (L_1) del archivo de n-gramas
5.	Mientras $L_1 \neq \text{EOF}$
6.	Se imprime L_1 en S
7.	Se imprime en S un tabulador
8.	Se imprime en S el número del documento analizado
9.	Se imprime en S un tabulador
10.	$B=0$
11.	Mientras $B \neq 0$

12.	Se lee la siguiente línea (L_2) del archivo de n-gramas
13.	Si $L_1=L_2$
14.	Se aumenta en 1 la frecuencia (F) de L_1
15.	De lo contrario
16.	Se imprime F en S
17.	Se imprime en S un salto de línea
18.	$B=1$ y $L_1=L_2$

Figura 14. Algoritmo usado para el cálculo de TF

Este programa crea un archivo de salida por cada uno analizado; en la Tabla 10 se muestra un ejemplo de éste, la primera columna es el n-grama, la segunda el identificador (ID) del documento y la tercera el TF de cada uno de los n-gramas analizados, en este caso son bigramas.

N-grama	ID del documento	TF
el animal	1	15
animal terrestre	1	10
punto cardinal	1	12
juan tener	1	6
energía renovable	1	23
localidad rural	1	11

Tabla 10. Ejemplo de un archivo de n-gramas con el identificador de documento y el TF

3.3.2 Limpieza de los n-gramas generados

La limpieza de n-gramas consiste en la eliminación de las construcciones con alta posibilidad de no ser términos. Su objetivo es disminuir el número de candidatos a término obtenidos por el extractor terminológico.

Dado que las unidades terminológicas no empiezan o terminan con palabras vacías como en los ejemplos dados en la sección 2.2.1, la limpieza consiste en la eliminación de los n-gramas que empiecen o terminen con palabras funcionales o números romanos. Asimismo, se eliminan los unigramas que son totalmente palabras funcionales y que no se les haya asignado un peso nulo durante el método de TF-IDF. Con respecto a los números romanos, al ser letras con referente numérico, no se les considera como términos y por tanto es necesario eliminarlos de los archivos de n-gramas; para lo cual se generó una expresión regular que

permitiera la detección de los n-gramas que tuvieran los caracteres de los números romanos (I, V, X, L, C, D, M) o sus combinaciones en los extremos para posteriormente eliminarlos.

La lista de palabras funcionales es un conjunto de palabras comunes que pertenecen al vocabulario general, como preposiciones, conjunciones, artículos, pronombres relativos, entre otros. Asimismo se tienen algunos verbos, como decir, hacer, estar, tener y ser (en este último caso, “ser” como nombre/sustantivo fue diferenciada del verbo, de lo contrario se podrían eliminar candidatos posibles como “ser humano” o “ser vivo”). De igual manera, en la lista de palabras funcionales se encuentran nombres de secciones de libro como capítulo y página. Esta lista de palabras se encuentra en el Anexo A.

Para llevar a cabo esta tarea de limpieza de n-gramas se generó un programa escrito en Perl que permite el uso de distintas listas de palabras vacías; el algoritmo que se empleó para programarlo se muestra en la Figura 15. En la Tabla 11 se muestra el formato de salida de un archivo generado por este programa, donde la primera columna indica el identificador del documento y la segunda el TF del n-grama.

Eliminación de palabras vacías	
1.	Se abre el archivo que contiene la lista de palabras vacías
2.	Se carga la lista de palabras vacías en un hash
3.	Se abre el archivo de n-gramas con frecuencias
4.	Se crea el archivo de salida (S)
5.	Se lee la primera línea (L) del archivo de n-gramas con frecuencias
6.	Mientras L!=EOF
7.	Se guarda en una cadena (C) la parte del n-grama de L
8.	Se extrae de C la primera palabra y se guarda en P
9.	Si P no existe en el hash entonces
10.	Si P es distinto a un número romano entonces
11.	Si C es un unigrama entonces
12.	Se imprime L en S
13.	De lo contrario se extrae de C la última palabra y se guarda en P
14.	Si P no existe en el hash entonces
15.	Si P es distinto a un número romano entonces
16.	Se imprime L en S

Figura 15. Algoritmo empleado para la limpieza de n-gramas

N-grama	ID del documento	TF
animal terrestre	1	10
punto cardinal	1	12
energía renovable	1	23
localidad rural	1	11

Tabla 11. Ejemplo de un archivo de salida del módulo de limpieza de n-gramas

3.3.3 Cálculo de IDF, TF-IDF y su normalización

Teniendo las listas de n-gramas limpias y con sus valores de TF se procede a realizar el tercer paso para la extracción de candidatos a término. En este paso se realizan tres tareas al mismo tiempo, que es el cálculo de la frecuencia inversa de los documentos (IDF), la multiplicación de las métricas IDF y TF para la obtención de los pesos TF-IDF y la normalización de estos pesos.

Con respecto a la normalización de los pesos de TF-IDF, ésta es frecuente en los sistemas de búsqueda de información para equilibrar los pesos que reciben los documentos largos y cortos, como se observó en la sección 1.2.2. No obstante, en el caso de esta tesis, se emplea la normalización para acotar el valor de los pesos otorgados por el método de TF-IDF entre 1 y 0; esto permite el establecimiento de umbrales que indiquen el peso mínimo para que un término pueda ser considerado como un buen candidato, de manera estática, es decir, que no se tenga que estar calculando según el valor máximo TF-IDF obtenido en cada documento. Por tanto, para realizar la normalización, se eligió el método de normalización de coseno (sección 1.2.2.1), debido a que no tiene una alta complejidad para obtener el factor de normalización, es sencillo de programar y, además, permite obtener pesos de TF-IDF entre 1 y 0.

Para llevar a cabo las tareas dadas a conocer al principio de esta sección se creó un programa escrito en C el cual tiene como entrada los archivos generados por el programa de limpieza de n-gramas (sección 3.3.2). En la Figura 16 se muestra el algoritmo empleado para realizar estas tres tareas.

Cálculo del IDF, TF-IDF y factor de normalización

1. Se unen todos archivos de n-gramas del mismo tipo en uno solo (A).
2. Se ordena A de manera alfabética
3. Se crea un archivo de salida (S)
4. Se crea un arreglo de frecuencias (TF)
5. Se crea un arreglo de TF-IDF (TF_IDF)
6. Se crea un arreglo de columnas (P)
7. $P[0] = "B"$
8. Se crea un arreglo de factores de normalización (FN)
9. $IDF=0, D_1=0, D_2=0, D_3=0$, y T=total de documentos
10. Se imprime "GRAMA" en S
11. Para $D_3 < T; D_3++$
12. Se imprime en S un tabulador
13. Se imprime el nombre del libro
14. Se lee la primera línea (L_1) de A
15. Se extrae de L_1 el n-grama (N_1)
16. Se imprime N_1 en S
17. Se extrae de L_1 el número del documento (D_2) donde N_1 se encontraba
18. Se extrae de L_1 la frecuencia (F) de N_1
19. Mientras $L_1 \neq EOF$
20. Para $D_1 < D_2; D_1++$
21. Si $D_1 \neq (D_2-1)$ entonces
22. $TF[D_1] = 0.0$
23. De lo contrario
24. $TF[D_1] = F$
25. $IDF++$
26. $D_1 = D_2$
27. Se lee la siguiente línea (L_2) de A
28. Se extrae de L_2 el n-grama (N_2)
29. Se extrae de L_2 el número del documento (D_2) donde N_2 se encontraba
30. Se extrae de L_2 la frecuencia (F) de N_2
31. Si $N_1 \neq N_2$
32. Para $D_1 < T; D_1++$
33. $TF[D_1] = 0.0$
34. $IDF = \log_{10}(T/IDF)$
35. $D_3 = 0$
36. Para $D_3 < T; D_3++$
37. Se imprime en S un tabulador
38. $TF_IDF[D_3] = TF[D_3] * IDF$
39. $FN[D_3] += TF_IDF[D_3]^2$
40. Se imprime "="TF_IDF[D₃]/P"\$1" en S
41. Si $D_3 \% 26 == 0$ entonces
42. $P[0] = "A" + (D_3/26) - 1$

43.	P[1]="A"
44.	De lo contrario, si $D_3 > 26$
45.	P[1]+=1
46.	De lo contrario
47.	P[0]+=1
48.	Se imprime en S un salto de línea
49.	Se imprime N_2 en S
50.	IDF=0, $D_1=0$, $D_2=0$, $D_3=0$
51.	$L_1=L_2$
52.	Se crea un archivo para guardar los factores de normalización (C)
53.	Se imprime "Factor de normalización" en C
54.	Para $D_3 < T$; D_3++
55.	$FN[D_3] = 1/\text{SQRT}(FN[D_3])$
56.	Se imprime en C un tabulador
57.	Se imprime $FN[D_3]$ en C
58.	Se imprime en C dos saltos de línea
59.	Se concatena el archivo C y S en se guardan en un archivo con formato de hoja de cálculo

Figura 16. Algoritmo para el cálculo de IDF, TF-IDF y del factor de normalización

En la línea 41 de la Figura 16 se verifica si D_3 al dividirlo entre 26 se obtiene un residuo de cero, esto para modificar el valor de P cuando se hayan puesto todas las letras del alfabeto latino, las cuales son 26, en la variable.

Cuando se llega a la línea 52 del algoritmo mostrado en la Figura 16, el archivo de salida tiene el formato mostrado en la Tabla 12, donde la primera columna indica el n-grama analizado, las siguientes son los pesos de TF-IDF dividido por la casilla donde se colocará el factor de normalización para cada libro analizado.

N-GRAMA	LIBRO 1	LIBRO 2	LIBRO 3
animal doméstico	=0/B\$1	=0/C\$1	=7.63/D\$1
animal terrestre	=1.76/B\$1	=0/C\$1	=6.16/D\$1
ecosistema	=0/B\$1	=12.88/C\$1	=0/D\$1
energía renovable	=0/B\$1	=0/C\$1	=0/D\$1
localidad rural	=1.93/B\$1	=0/C\$1	=3.69/D\$1
punto cardinal	=2.11/B\$1	=0/C\$1	=0.52/D\$1

Tabla 12. Ejemplo de un archivo de salida intermedio de la extracción terminológica antes de la normalización

A partir de la línea 52 del algoritmo mostrado en la Figura 16 se procede a llenar las casillas establecidas para el factor de normalización y genera un archivo que está en un formato de hoja de cálculo; al abrir el archivo la normalización se lleva a cabo de manera automática, además de que este formato de archivo tiene la ventaja de que permite analizar las listas n-gramas mucho más fácilmente. En la Tabla 13 se muestra un extracto del archivo de salida de este módulo.

	A	B	C	D
1	FACTOR DE NORMALIZACIÓN	3.35	12.88	10.49
2				
3	N-GRAMA	LIBRO 1	LIBRO 2	LIBRO 3
4	animal doméstico	0	0	0.72
5	animal terrestre	0.52	0	0.58
7	ecosistema	0	1	0
6	energía renovable	0	0	0
8	localidad rural	0.57	0	0.35
9	punto cardinal	0.62	0	0.04

Tabla 13. Ejemplo de un archivo de salida de la parte de la extracción terminológica

3.4 Validación de los candidatos a término

Aunque frecuentemente las listas que se obtienen de los procesos de extracción terminológicos se consideran las listas de términos finales, en este proyecto de tesis se busca emplear recursos léxicos para la validación de los candidatos a término como lo hace YATE o MetaMap. La razón de llevar a cabo una validación es que ésta permitiría obtener listas de términos mucho más fiables, de menor tamaño y orientadas al área o categorías de análisis.

Para llevar a cabo la validación de los candidatos a término se empleó Wikipedia como recurso léxico por su amplia cobertura en distintas áreas científicas y porque se consideró que tiene la mayoría de los términos que se encuentran dentro del COCIEM. Este proceso consiste primeramente en seleccionar los candidatos a término a validar, a partir de las listas generadas por el análisis de TF-IDF (sección 3.3) con base en los pesos de TF-IDF.

Posteriormente, se determinan las categorías de Wikipedia que correspondan de la manera más cercana a las áreas de las listas de candidatos a término a validar. Finalmente se calcula el coeficiente de dominio, el cual indica qué tan relacionado está el candidato a término con las áreas seleccionadas.

A continuación, se hablará de la estructura de Wikipedia, su conversión a una base de datos, la manera en que se lematizó y del coeficiente de dominio.

3.4.1 Wikipedia para la validación

Uno de los grandes recursos léxicos digitales es Wikipedia, la cual se encuentra en más de 200 idiomas y es la enciclopedia en línea más grande del mundo. Aunque se había dado a conocer su estructura a grandes rasgos en la sección 2.4.3, en esta parte se explicará a mayor profundidad la organización y arquitectura de Wikipedia.

Según Zesch y Gurevych (2007a) la enciclopedia Wikipedia está formada por dos grafos interconectados, el primero de ellos es el de categorías, mientras que el segundo es el de páginas o artículos. Con base en estos autores, la Figura 17 muestra la arquitectura de la enciclopedia Wikipedia.

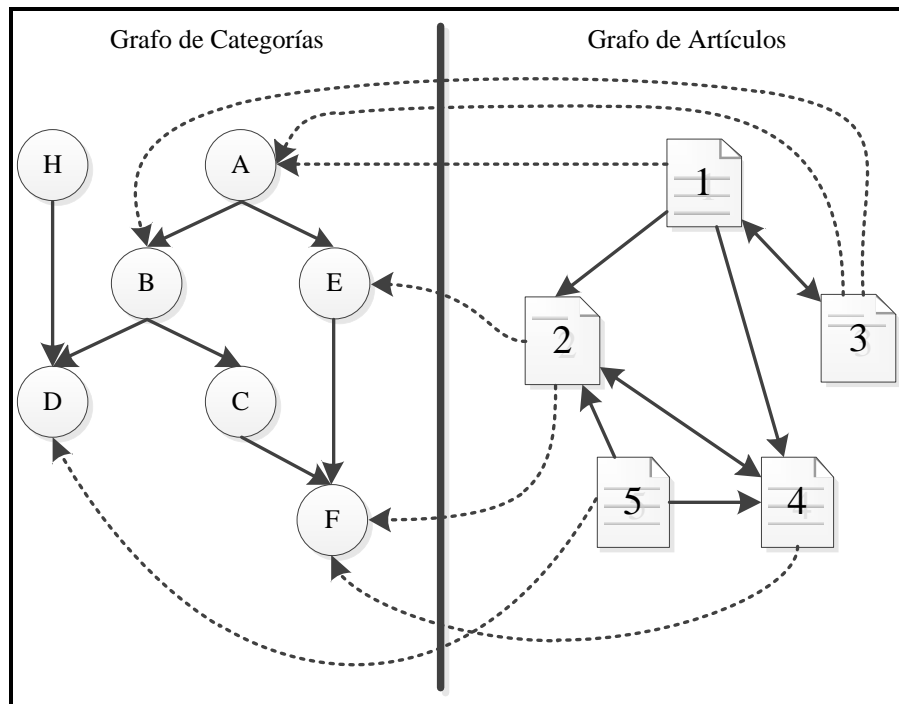


Figura 17. Arquitectura de Wikipedia

El *grafo de categorías* o *Wikipedia Category Graph* (WCG) es una estructura organizada a manera de taxonomía, la cual se maneja, al igual que las taxonomías, mediante una estructura jerárquica; este grafo almacena todas las categorías existentes en Wikipedia. Cada categoría tiene un número arbitrario de subcategorías, donde una subcategoría es típicamente establecida por relaciones de hiponimia³⁰ y meronimia³¹. Cabe aclarar que Wikipedia no forma, de manera estricta, una taxonomía, debido a que ciertas veces las categorías no cumplen con las relaciones mediante las cuales deberían estar unidas y se generan ciclos o categorías desconectadas de la taxonomía. Por lo anterior, Peters (2009) considera a la Wikipedia como una folksonomía, ya que la gente es quien desarrolla la jerarquización y no los expertos en las materias como ocurre, por ejemplo, con las taxonomías biológicas.

La Wikipedia en español se encuentra dividida en las siguientes partes: Anexos, Artículos, Ayuda, Categorías y Wikipedia. La categoría “Artículos” es la más grande de todas, pues no sólo incluye las categorías de Wikipedia, sino también las páginas de los artículos, y por tanto es la de mayor uso.

El *grafo de artículos* o *grafo de páginas* es una estructura no organizada la cual se genera automáticamente a través de los vínculos que contienen las páginas hacia otros artículos de Wikipedia. Por ejemplo: El artículo “Zeus” tiene un vínculo hacia “Mitología griega”, éste a su vez con “Religión de la Antigua Grecia” y finalmente, este artículo tiene un vínculo con la página “Zeus”. Como se puede observar en el ejemplo anterior y en la Figura 17, las relaciones entre las páginas pueden ser recíprocas y cíclicas.

Además de los artículos que contienen información, existen algunos artículos de Wikipedia que contienen solamente un redireccionamiento a otra página o forman lo que se conoce como una página desambiguación; estos dos tipos de página funcionan como información extra de la enciclopedia. El redireccionamiento, como se observó en la sección

³⁰ Las relaciones de hiponimia incluyen a su vez relaciones de hiperonimia, ya que es la relación inversa o que va en sentido contrario.

³¹ La *meronimia* es la relación semántica entre un elemento léxico que denota una parte y otro elemento léxico que denota al primer elemento y a otros a su vez (Cruse, 1986). Ejemplo: Bujía es un merónimo de motor, y motor es merónimo de automóvil.

2.4.3, permite incluir, principalmente, variaciones ortográficas, morfológicas o abreviaturas de los nombres de los artículos; de manera menos frecuente, el redireccionamiento se utiliza para pasar de una página que habla de un tema muy específico a una página que expresa un tema más general o de un verbo a un sustantivo, por ejemplo, el verbo “sumar” redirige a “suma”, mientras que “DVI-I” dirige a “Digital Visual Interface”. En cambio, las páginas de desambiguación son un repositorio de artículos polisémicos, es decir, artículos que representan varios temas pero que tienen un nombre igual. Ya que el nombre de las páginas de Wikipedia debe ser único, a la página de la acepción más común se le deja frecuentemente el nombre del tema, mientras que a las páginas de las acepciones menos comunes se les coloca el nombre del tema más un identificador entre paréntesis que expresa más específicamente la acepción (Zesch et al., 2007b); por ejemplo, en la Wikipedia en español para “Metro” existe una página de desambiguación que permite elegir entre la unidad de longitud (Metro_(medida)), el sistema de transporte metropolitano (Metro_(sistema_de_transporte)), un periódico (Metro_(periódico)) y un canal de televisión argentino (Metro_(canal_de_televisión)), entre otros.

El grafo de artículos y el de categorías están unidos, ya que la gran mayoría de los artículos de Wikipedia están asignados a una o más categorías. En la Figura 17 se muestran con líneas punteadas las uniones entre los artículos y las categorías.

La arquitectura de Wikipedia no es a prueba de errores. Existen casos en los cuales los artículos no están vinculados a la categoría correcta; por ejemplo, en la categoría “Almacenamiento informático” están vinculados de manera correcta los artículos “Unidades de disco”, “Caché”, entre otros, pero también aparecen las páginas “Quantum Corp.” que es una empresa, y “Robocopy” que es un comando de Windows para hacer copias de archivos, ambas que no debieran aparecer. Asimismo, otro de los problemas de Wikipedia es que existen algunos casos en los que los artículos no se encuentran unidos a ninguna categoría, ya que su creador no los unió a alguna de ellas. De igual manera, hay artículos que están unidos a categorías especiales para indicar que existen problemas de edición, de coherencia, de neutralidad, etcétera.

También existen problemas, algunas veces, con la unión entre las páginas de los artículos, es decir, los vínculos dentro de los artículos dirigen a un tema equivocado o a una

acepción errónea. De igual forma, las páginas de redireccionamiento o de desambiguación pueden no estar unidas al artículo correcto y por tanto no hay una coherencia entre lo buscado y el resultado obtenido.

A pesar de que tiene algunas desventajas la arquitectura de la enciclopedia, la Wikipedia ha sido empleada en diversas áreas de PLN, por ejemplo Toral et al. (2006) usan la enciclopedia para la creación de listas de entidades nombradas, Ponzetto y Strube (2008) crean taxonomías de manera automática a partir de Wikipedia. Mientras que Suchanek (2008) desarrolla ontologías usando la información de la enciclopedia y Gabrilovich y Markovitch (2009) emplean Wikipedia para la interpretación semántica de textos en lenguaje natural.

3.4.1.1 Conversión a una base de datos

Para llevar a cabo el proceso de validación de los candidatos a término es necesario tener acceso a Wikipedia, para realizar esto se pueden emplear diversos métodos como los siguientes:

Web crawler: También conocido como *araña web*, es un programa que inspecciona el internet o una determinada página, en este caso Wikipedia, de manera metódica y sistematizada. Estos programas permiten almacenar las páginas web visitadas. Aunque se puede emplear este método, Wikimedia³² pide que no se empleen arañas web para extraer información de Wikipedia por la sobrecarga que se genera en los servidores.

Bot: Otro de los métodos empleados en la extracción de información de Wikipedia es el uso de *bots*, los cuales son programas informáticos que actúan como si fueran un humano. En este caso sirven como web crawlers pero son mucho más lentos debido a que es necesario que actúen lo más natural posible y no saturaren los servidores.

APIs: Las *Interfaces de programación de aplicaciones* o *APIs* por su acrónimo en inglés son un conjunto de métodos o funciones incluidas en una biblioteca que permiten emplear un determinado software. En la actualidad existen muchos de estos APIs para el manejo de Wikipedia mediante el empleo de copias de las bases de datos que Wikimedia publica

³² Wikimedia es la organización encargada de manejar la Wikipedia.

frecuentemente para que servidores u otras computadoras tengan acceso a la enciclopedia sin tener que saturar los servidores originales empleando un web crawler.

En el caso de esta tesis se empleará el método descrito por Zesch et al. (2008), el cual se encuentra disponible como una API en el internet³³. Esta API se llama *Java-based Wikipedia Library (JWPL)*.

La estructura original de las bases de datos de Wikipedia está optimizada para la búsqueda de artículos a través de palabras claves que son hechas por millones de usuarios de Wikipedia cada día (Zesch et al., 2008). Sin embargo, esta estructura no es la adecuada para su uso en proyectos de PLN, debido a que es necesario que se soporten búsquedas iterativas, acceso a gran número de caminos de Wikipedia o a la información dentro de las páginas de los artículos como los vínculos o categorías. Por ello JWPL permite la conversión de la base de datos de Wikipedia en una base de datos optimizada para usarla en tareas del PLN.

La optimización de la base de datos de Wikipedia consiste en convertir la información de redireccionamiento y de otros recursos léxicos y semánticos que se encuentra de forma implícita a una forma explícita; la razón de ello es que una gran parte de la información que contiene Wikipedia está dentro de las páginas de los artículos y no en los archivos de la base de datos de la enciclopedia que se publican; por ejemplo, la página “Capacitor” contiene como información “[[Redirect: Condesador_eléctrico]]”, el cual indica que se debe redireccionar a la página “Condensador eléctrico”; por consiguiente, sin la optimización las páginas de los artículos deberían ser analizadas sintácticamente, cada vez que se consulte, para saber si son de redireccionamiento y así obtener la página a la que se dirige (Zesch et al., 2007b). Lo que lleva a cabo una de las herramientas de JWPL es analizar cada una de las páginas de la Wikipedia y separar las que sean de redireccionamiento o que contengan otra información léxica o semántica; estas últimas son analizadas y se les extrae la información implícita y se almacena de manera explícita en tablas de una base de datos.

La base de datos optimizada de Wikipedia está conformada por 11 tablas y son las siguientes:

³³ <http://code.google.com/p/jwpl/issues/list>

Category: En esta tabla se guardan todas las categorías de Wikipedia con su nombre e identificador.

Category_inlinks: Almacena los identificadores de las categorías superiores para una determinada categoría.

Category_outlinks: Guarda los identificadores de las subcategorías de una categoría específica.

Category_pages: La tabla tiene recopilados los identificadores de los artículos que pertenecen a cada categoría.

MetaData: Contiene la información básica de la base de datos de Wikipedia, como el idioma, número de páginas, nombre de la categoría de desambiguación, etcétera.

Page: Guarda todas las páginas de los artículos de Wikipedia; en esta tabla se encuentra el texto, si es de desambiguación, el nombre de la página, entre otros.

Page_categories: En esta tabla se encuentran los identificadores de todas las categorías que tiene asociado cada artículo de Wikipedia.

Page_outlinks: Esta tabla se genera en la optimización y almacena el identificador de las páginas que salen de un artículo.

Page_inlinks: Al igual que “Page_outlinks”, esta tabla se genera en la optimización y se encarga de guardar los identificadores de las páginas que se dirigen hacia un mismo artículo.

Page_redirects: Contiene los identificadores de las páginas que redireccionan a una página en específico. Esta tabla se genera al momento de optimizar la base de datos de Wikipedia.

PageMapLine: Esta tabla contiene el nombre de las páginas de los artículos, su identificador general y el identificador de la página a la que corresponde, que en caso de no ser igual que el identificador general indica que existe un redireccionamiento; en algunos casos aparece el lema y una truncación del nombre de la página.

Entre las ventajas de la optimización de las bases de datos de Wikipedia está tener una eficiencia computacional en tareas de lenguaje natural de gran escala y obtener resultados reproducibles (siempre y cuando se emplee la misma versión de Wikipedia).

Con respecto a la versión de Wikipedia que se empleará en la tesis es la copia de la base de datos de noviembre de 2010.

3.4.1.2 Lematización de Wikipedia

Aunque la lematización de los textos permite obtener algunas ventajas como la reducción y agrupamiento de términos, conlleva un problema con Wikipedia. Este problema está relacionado con los nombres de los artículos de la enciclopedia, pues estos no siempre concuerdan con su forma lematizada; por ejemplo, “Análisis de circuitos” y “Sexualidad humana”, tienen como lema “Análisis de circuito” y “Sexualidad humano”, respectivamente. Por tanto, no es posible emplear los nombres de los artículos de Wikipedia de manera exacta para realizar la validación de los candidatos a término, ya que los candidatos a término se encuentran lematizados.

Para resolver el problema anterior se decidió lematizar los nombres de los artículos de Wikipedia. Esta tarea se desarrolló con un programa creado en Perl que preparara la Wikipedia y emplea el lematizador FreeLing, este último se usa para que se obtengan resultados similares a los que se llevan a cabo con los textos.

Los pasos seguidos para realizar este proceso fueron, en primer lugar, la extracción de los nombres de los artículos de Wikipedia junto con su identificador³⁴ para su almacenamiento en un archivo de texto, información que fue obtenida de la tabla “PageMapLine” de la base de datos de Wikipedia. Posteriormente, el archivo fue preprocesado; esto consistió en agregar el símbolo de almohadilla (#) y un espacio en blanco antes del nombre de cada artículo y colocar en minúscula la primera letra del nombre. Lo anterior se llevó a cabo porque sin él la lematización no se realizaba de manera adecuada por falta de contexto; por ejemplo, “Derivada” quedaba como “derivada”, mientras que “Sistemas complejos” como “sistemas complejo”; esto porque FreeLing considera en algunos casos que las palabras que inician en mayúscula al principio de una oración son nombres propios y empleando la almohadilla y la conversión a minúscula de la primera letra se corrige este problema. El siguiente paso consistió en la ejecución del programa de FreeLing sobre el archivo de texto preprocesado. Finalmente, la última parte del proceso consistió en convertir

³⁴ Este identificador o ID es el número secuencial que se crea al agregar un artículo a Wikipedia de manera automática y permite diferenciar cada una de las entradas de la enciclopedia; este ID se encuentra en la tabla Page y PageMapLine (sección 3.4.1.1).

el archivo de texto lematizado a un archivo lo más similar al obtenido en el primer paso (usando un método similar al dado a conocer en la sección 3.2.2), esto para no afectar de manera sistemática la arquitectura de la Wikipedia. Además, dentro de este proceso se eliminó la almohadilla y el espacio en blanco extra y se colocó en mayúscula la primera letra del nombre del artículo.

Una vez teniendo los nombres lematizados, se procedió a actualizar la base de datos de Wikipedia. Para llevar esto a cabo se crearon dos tablas más en la base de datos de Wikipedia, una de ellas para almacenar los nombres de los artículos lematizados y otro para los nombres sin lematizar, ambas como respaldo. Los datos de la tabla con los nombres en su forma canónica actualizaron las tablas “PageMapLine” y “Page”, las cuales son las tablas en donde se puede buscar la información por nombre del artículo.

3.4.2 Cálculo del coeficiente de dominio

Para poder llevar a cabo la validación de los candidatos a término que se obtuvieron del extractor automático es necesario acordar la manera en que se determinará si pertenecen a una materia o a otra. Para ello es necesario emplear un método que permita establecer un *coeficiente de dominio*; es decir, una métrica que indique qué tan relacionado se encuentra un término con una determinada categoría o área de Wikipedia.

En esta tesis se empleará el método desarrollado por Vivaldi y Rodríguez (2010) para el cálculo del coeficiente de dominio, el cual se explica a continuación.

Como se había indicado en la sección 3.4.1, Wikipedia está conformada por una serie de categorías y subcategorías. En el cálculo del coeficiente de dominio las categorías y sus divisiones forman lo que llamaremos *fronteras de dominio*, las cuales definen las áreas o materias a las que puede pertenecer un término. En algunos casos la materia a analizar tiene su par exacto en Wikipedia, como “química” o “economía”; en algunos otros no, como “computación” que necesita las categorías de Wikipedia “informática” y “electrónica”.

A partir de lo anterior se procede a analizar Wikipedia para cada uno de los candidatos a validar. Este análisis comienza buscando la página del tema que sea igual a la del candidato a término dentro de la tabla “PageMapLine” de la base de datos de Wikipedia.

Luego, si el término es encontrado en la tabla “PageMapLine”, entonces éste se busca en la tabla “Page” para verificar si es una página de desambiguación o no. Si las páginas no son de desambiguación se buscan las categorías a las que está asociada la página del artículo dentro de la tabla “Page_categories”. En el caso de que la página sea de desambiguación, se almacenan los identificadores de las páginas relacionadas al término polisémico y se analizan, una por una, buscando sus categorías como se explicó en arriba. Por ejemplo, en el caso de la Figura 18, se muestra que el término “Oxidrilo”, el cual fue redireccionado a “Grupo hidroxilo”, está asociado a las categorías “Grupos funcionales”, “Compuestos de oxígeno” y “Compuestos de hidrógeno”. Posteriormente, para cada una de las categorías que se encontraron en el paso anterior, se realiza una búsqueda recursiva en sus categorías superiores hasta encontrar las fronteras de dominio o una categoría tope de Wikipedia, en este caso será “Artículos”. Además, esta búsqueda tiene un límite establecido, ya que en ocasiones se pueden encontrar ciclos infinitos y, por tanto, puede llegarse a no terminar el análisis si no existe un límite.

The image shows a screenshot of the Wikipedia article for "Grupo hidroxilo". The page title is "Grupo hidroxilo" with a note "(Redirigido desde Oxidrilo)". The main text explains that the hydroxyl group (OH) is a functional group consisting of one oxygen and one hydrogen atom, characteristic of alcohols. It also mentions the hydroxide ion (OH⁻). A 3D ball-and-stick model of the hydroxide ion is shown on the right. The page includes a sidebar with navigation options and a search bar at the top.

Figura 18. Ejemplo de las categorías a las que pertenece el término Grupo hidroxilo

Posteriormente, con la información obtenida del análisis anterior se calcula el coeficiente de dominio con base en las fórmulas mostradas por Vivaldi y Rodríguez (2010) con algunas modificaciones llevadas a cabo por comunicación expresa con el Dr. Jorge

Vivaldi. A continuación se muestran las fórmulas empleadas para el cálculo del coeficiente de dominio:

Basado en el número de caminos:

$$CDnc(t) = \frac{NCdominio(t)}{NCtotal(t)} \quad (20)$$

Donde $CDnc$ es el coeficiente de dominio, t es el candidato a término, $NCdominio$ es el número de caminos al dominio y $NCtotal$ es el número de caminos a la categoría máxima de Wikipedia.

Basado en la longitud de los caminos:

$$CDlc(t) = \frac{LCtotal(t) - LCdominio(t)}{LCtotal(t)} \quad (21)$$

Donde $CDlc$ es el coeficiente de dominio, t es el candidato a término, $LCdominio$ es la longitud (o número de saltos) de los caminos al dominio y $LCtotal$ es la longitud de los caminos a la categoría máxima de Wikipedia. Además hay que aclarar que $CDlc$ tendrá un valor de 1 cuando $LCtotal$ y $LCdominio$ sean iguales.

Basado en la longitud promedio de los caminos:

$$CDlmc(t) = \frac{LMCtotal(t) - LMCdominio(t)}{LMCtotal(t)} \quad (22)$$

Donde $CDlmc$ es el coeficiente de dominio, t es el candidato a término, $LMCdominio$ es la longitud media de los caminos al dominio y $LMCtotal$ es la longitud media a la categoría máxima de Wikipedia. El coeficiente $CDlmc$ será igual a 1 cuando $LMCtotal$ y $LMCdominio$ tengan el mismo valor.

En el caso de de los coeficientes $CDlc$ y $CDlmc$, la longitud de los caminos al tope que pasan por la frontera de dominio se miden hasta ella, es decir, si hay nodos más arriba de ésta no se toman en cuenta. Esto se realiza para que se obtenga un coeficiente de dominio igual a 1 cuando todos los caminos que salen del término se dirigen a la frontera de dominio, de esta manera, la distancia entre la frontera de dominio y la categoría tope no afecta el valor del coeficiente de dominio.

El coeficiente de dominio para los candidatos a término puede tener un valor de -1 si no se encuentran en Wikipedia, entre $(-1, 0]$ si se encuentran en la enciclopedia pero no tienen ninguna relación con las fronteras de dominio, entre $(0, 1)$ si tienen una cierta relación con la frontera de dominio y 1 si pertenecen totalmente a la frontera de dominio.

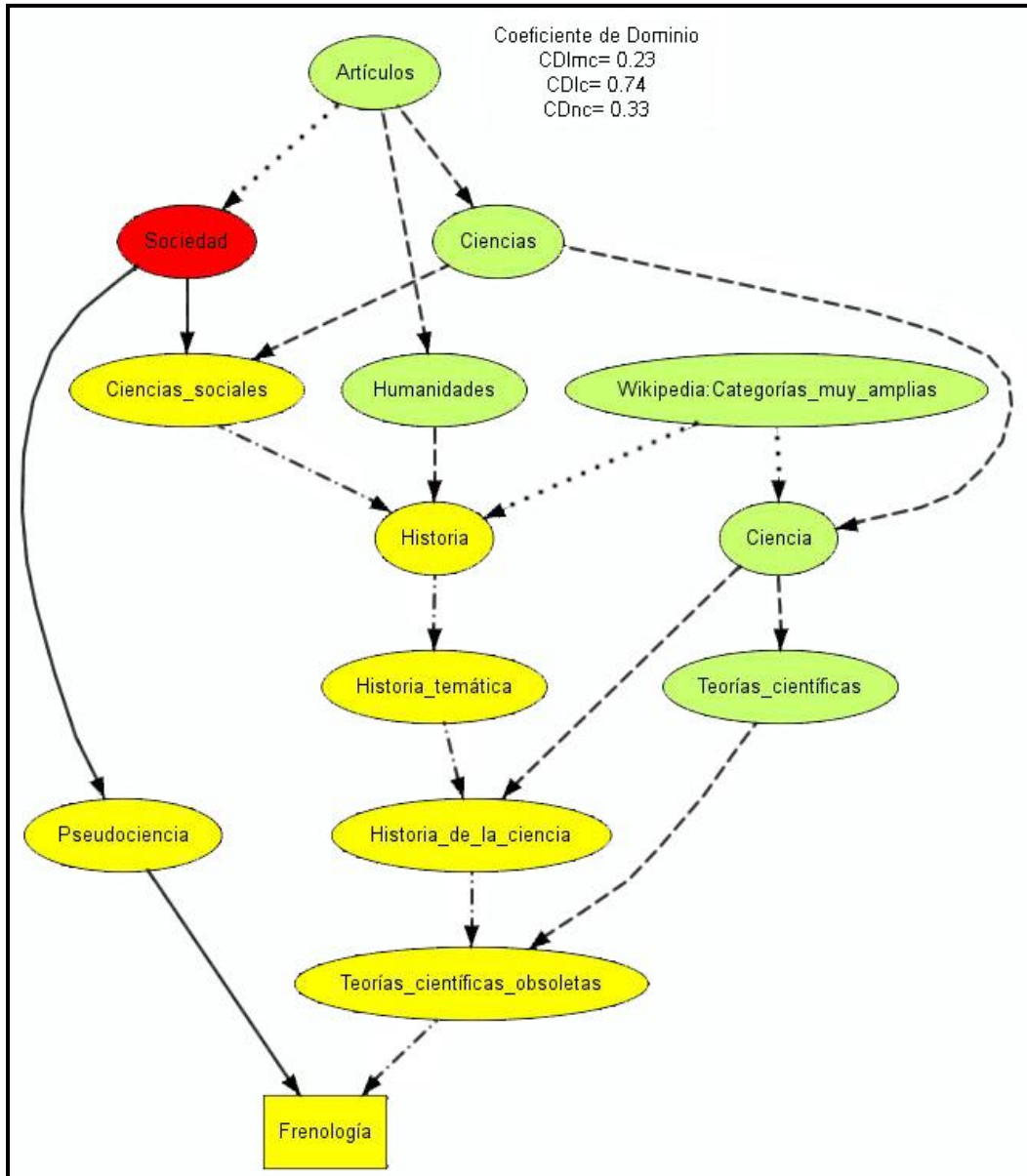


Figura 19. Grafo para el término “Frenología” y sus categorías

En la Figura 19 se muestra un grafo que muestra los datos otorgados por Wikipedia para el término “Frenología”. El rectángulo en amarillo indica el término a buscar; los óvalos amarillos indican las categorías que se encuentran entre el término y la frontera de dominio.

El óvalo rojo, es la frontera de dominio, que en este caso es “Sociedad”. Finalmente los óvalos verdes, son las categorías que van a la categoría tope pero no pasan por la frontera de dominio. En el caso que existiera en la Figura 19 un óvalo blanco, éste indicaría que es una categoría donde pasa un camino que va únicamente de la frontera de dominio a la categoría tope.

El óvalo verde llamado “Wikipedia: Categorías muy amplias” no se toma en cuenta para este análisis como una categoría tope y sus caminos no son contados en el análisis porque no todos los temas están conectados de alguna manera a esta categoría y se busca que en general todos los temas tengan el mismo tope, que en este caso es la categoría “Artículos”.

Asimismo, la Figura 19 indica en la parte superior derecha los coeficientes de dominio calculados a partir de las Ecuaciones (20), (21) y (22). La línea continua indica los caminos del término a la frontera de dominio; la línea punteada es un camino que no se cuenta, ya sea porque va de la frontera de dominio a la categoría tope o porque va a “Wikipedia: Categorías muy amplias”. La línea formada por puntos y guiones son los caminos que van tanto al tope como a la frontera de dominio. Y la línea intermitente son los caminos que van solamente a la categoría tope. A continuación se muestran los cálculos de los coeficientes de dominio con base en la Figura 19:

$$CDlmc (Variable dependiente) = \frac{\frac{31}{6} - \frac{8}{2}}{\frac{31}{6}} = \frac{5.1666 - 4}{5.1666} = \frac{1.1666}{5.1666} = 0.225 \approx 0.23$$

$$CDlc (Variable dependiente) = \frac{31 - 8}{31} = \frac{23}{31} = 0.74$$

$$CDnc (Variable dependiente) = \frac{2}{6} = 0.3\bar{3}$$

Para comprender mejor el coeficiente $CDlc$ y el coeficiente $CDlmc$, el cálculo de la distancia con valor igual a 31 se muestra en la Tabla 14.

Distancia del término a la frontera de dominio (suma de segmentos de la línea recta y de la punteada con guiones)	8
Distancia entre el término y la categoría tope sin pasar por la frontera de dominio (suma de la línea punteada con guiones y de la intermitente)	23
Distancia total	31

Tabla 14. Cálculo desglosado para la distancia entre el término y la categoría tope usando como ejemplo la Figura 19

Todo el proceso de validación, como se indicó anteriormente, se lleva a cabo para cada uno de los candidatos a término seleccionados. La validación genera archivos de salida con el formato indicado en la Tabla 15, donde la primera columna indica el coeficiente de dominio, la segunda el candidato a término analizado y la tercera la manera en que se encontró en Wikipedia: “page” si se encontró de manera exacta, “nil” si no se encontró en la enciclopedia, “pagedir” si hubo un redireccionamiento y “pagedesamb” si se encontró una página de desambiguación. En estas dos últimas formas se indica el nombre de la página del artículo a donde se dirigió o desambiguó.

- Bigramas		
- Coeficiente de dominio CDwp_lc		
1.00	corriente eléctrico	(page)
1.00	capacitor	(pagedir: Condensador_eléctrico)
0.96	resistencia	(pagedesamb: Resistecia_eléctrico)
0.00	caucho natural	(pagedir: caucho)
-1.00	Juan Cárdenas	(nil)

Tabla 15. Ejemplo del formato de salida de los archivos de los candidatos a término validados por Wikipedia

3.5 Arquitectura del sistema

Para la obtención y validación de términos se desarrolló un sistema que fue explicado a lo largo de este capítulo. Para finalizar, se mostrará a continuación la arquitectura del sistema creado y se puede observar en la Figura 20.

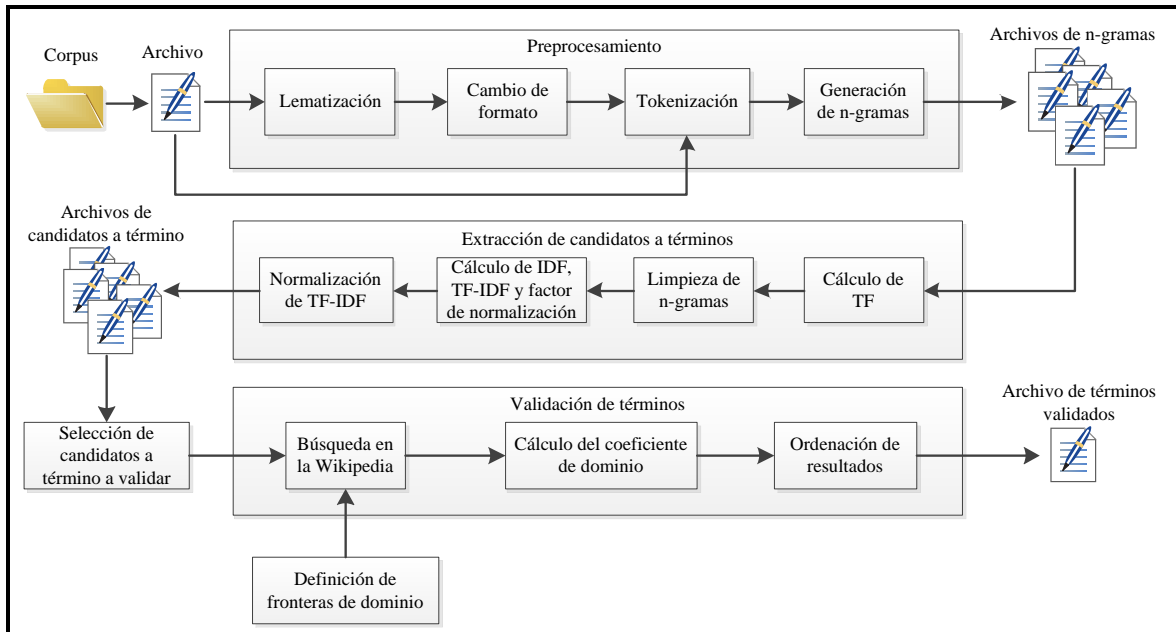


Figura 20. Arquitectura del sistema desarrollado en la tesis

La primera parte es el preprocesamiento de cada uno de los archivos del corpus, en ella se lleva a cabo la lematización, el cambio de formato, la tokenización y la generación de n-gramas (sección 3.2). Los archivos pueden ser directamente tokenizados si ya se llevó a cabo previamente una lematización y un cambio de formato, esto para disminuir los tiempos de preprocesamiento. El preprocesamiento se lleva a cabo para cada uno de los documentos seleccionados del corpus y da como resultado tres archivos por documento; estos tres archivos contienen los tres tipos de n-gramas usados en esta tesis, unigramas, bigramas y trigramas.

La segunda parte es la extracción de candidatos a término (sección 3.3), la cual se lleva a cabo, uno por uno, a los archivos generados por el preprocesamiento. Está conformado por 4 módulos; el primero de ellos se encarga de calcular el TF de los n-gramas que se encuentren en los archivos generados por el preprocesamiento. El segundo módulo quita los n-gramas de los archivos de salida del primer módulo que tengan en sus extremos palabras funcionales o números romanos, y que por tanto tienen alta probabilidad de no ser términos. El tercer módulo, primeramente, une todos los archivos creados en el segundo módulo por tipo de n-gramas (unigramas, bigramas, trigramas), y posteriormente para cada archivo de tipo de n-gramas realiza tres tareas, las cuales son, el cálculo del IDF, del TF-IDF y del factor de normalización. El último módulo de esta parte crea archivos en formato de

hoja de cálculo de las listas generadas por el tercer módulo y aplica el factor de normalización a los pesos de TF-IDF calculados.

Después de la extracción de candidatos a término, se procede a seleccionar los candidatos que serán validados por Wikipedia en función de los pesos recibidos por el análisis TF-IDF.

Finalmente, la cuarta parte consiste en la validación de los candidatos a término seleccionados en la parte anterior. Para llevar a cabo el proceso de esta parte se seleccionan las fronteras de dominio, las cuales deben coincidir con categorías existentes en Wikipedia y abarcar las de las listas de candidatos a término a analizar. Teniendo lo anterior determinado, se comienza la búsqueda de cada uno de los candidatos a término en la base de datos de Wikipedia y se realiza el cálculo del coeficiente de dominio. Finalmente, se ordena la lista de los términos validados, es decir, los resultados de mayor a menor coeficiente de dominio y se almacenan en un archivo de texto plano.

4 RESULTADOS Y EVALUACIÓN

Habiendo explicado la metodología empleada para la obtención automática de términos y su validación en el capítulo anterior, es necesario pasar a los resultados obtenidos y su evaluación.

Este capítulo estará conformado por las siguientes partes, la primera de ellas se refiere a la extracción de candidatos a término del COCIEM; la segunda es la selección de candidatos a término a validar. La tercera parte está conformada por la validación de los candidatos a término empleando Wikipedia. Y finalmente, la cuarta parte es la evaluación de los resultados obtenidos y algunas observaciones de ésta.

4.1 Extracción de candidatos a término del COCIEM

Para llevar a cabo la extracción de candidatos a términos, primeramente, se establecieron dos parámetros. El primero de ellos era la división del COCIEM en tres subcorpus, cada uno determinado por el nivel escolar, es decir, primaria, secundaria y bachillerato. La razón de realizar esta división fue para que los candidatos a término se quedaran en cada uno de los niveles educativos y no se perdieran o disminuyera su importancia por el avance del conocimiento a través de los años. El segundo parámetro era considerar cada libro de los subcorpus como un documento, es decir, los libros no se concatenaron por materia o grado escolar. Lo que se buscó con esto último es que los términos no se vieran sobre o subcalificados al hacer el análisis de TF-IDF y además se tuvieran listas de candidatos a término por libro.

Posteriormente, ya con los parámetros establecidos se procedió a preprocesar cada uno de los subcorpus y a extraer los candidatos a término siguiendo las dos primeras partes de la arquitectura mostrada en la sección 3.5. De estas dos tareas se obtuvieron tres listas de candidatos a término por subcorpus que pertenecen a unigramas, bigramas y trigramas, respectivamente; asimismo estas se encuentran divididas por libro siguiendo el formato de salida mostrado en la Tabla 13 de la sección 3.3.3. En la Tabla 16 se muestra el número de

candidatos a término con repetición que se obtuvieron de la extracción de candidatos a término.

Nivel educativo	Materia	Número de libros de texto	Unigramas	Bigramas	Trigramas
Primaria	Ciencias naturales	6	6,606	6,658	14,536
	Matemáticas	9	4,891	3,458	7,618
Total		15	11,497	10,116	22,154
Secundaria	Biología	8	13,968	17,636	32,036
	Matemáticas	24	10,728	17,945	35,127
	Física	9	7,837	19,823	38,296
	Química	8	10,812	16,583	31,745
	Educación ambiental	1	1,625	3,738	7,370
Total		50	44,970	75,725	144,574
Bachillerato	Biología	3	7,202	7,057	11,484
	Matemáticas	11	8,450	9,924	18,405
	Física	3	3,296	8,286	15,817
	Química	5	7,785	12,700	23,393
	Educación para la salud	3	3,958	8,570	11,689
	Ecología	2	4,817	5,999	10,690
Total		27	35,508	52,536	91,478
Gran total		92	91,975	138,377	258,206

Tabla 16. Número de candidatos a término con repetición por nivel educativo, materia y tipo de n-grama

Como se puede observar en la Tabla 16, en los casos de matemáticas de los tres niveles educativos el número de candidatos a término es menor a la proporción de libros que conforman a las materias. La razón de ello es que la materia de matemáticas se encuentra

constituida por varios libros de ejercicios y actividades y no tanto de teoría como en otras materias.

Los pesos de TF-IDF para la mayoría de los candidatos a término son medianos y bajos, pocos son los que obtuvieron un peso alto. Posiblemente porque al ser libros con un enfoque educativo se incluye una gran cantidad de información o ejemplos que pueden estar altamente relacionados con la vida cotidiana, como lo son las compras en el mercado o los medios de transporte; asimismo se incluye información que está frecuentemente relacionada con otras materias para comprender de mejor forma los temas explicados.

4.2 Selección de los candidatos a término a validar

Una vez finalizada la extracción de candidatos a término se procedió a continuar con la tercera parte de arquitectura del sistema de extracción terminológica mostrada en la sección 3.5. Esta tercera parte es la selección de los candidatos a término a validar, el cual es una tarea que se lleva a cabo a partir de las listas de candidatos a término obtenidas en los dos pasos anteriores dados a conocer en la sección 4.1.

A pesar de haber obtenido listas de candidatos a término para cada uno de los subcorpus del COCIEM y sus libros, solamente se eligieron tres materias para proseguir con la selección de los candidatos a término a validar. Las razones de realizar esto es que solamente se pudieron obtener 3 listas de términos creados por expertos para realizar una evaluación, de las cuales se hablarán más adelante en la sección 4.4; también se eligieron éstas porque son materias que tienen una categoría bien definida dentro de Wikipedia. Las materias seleccionadas son las siguientes:

- Matemáticas de bachillerato
- Ecología de bachillerato
- Matemáticas de primaria

Para llevar a cabo la selección de los candidatos a término a validar de estas tres materias se establecieron dos parámetros. El primero de ellos consiste en la generación de distintas listas de candidatos a término a validar usando distintos umbrales de pesos de TF-

IDF, esto para observar los distintos resultados que se obtienen al modificar el límite de lo que se considera un candidato a término a validar o no. Con base en que los pesos de TF-IDF que se obtuvieron no fueron altos sino bajos y medianos, se instituyeron los siguientes tres umbrales:

- $TF-IDF \geq 0.03$
- $TF-IDF \geq 0.01$
- $TF-IDF > 0$

El segundo parámetro que se estableció fue el de unir las listas de candidatos a término a validar por materia, en otras palabras, se unieron las listas de candidatos a término de unigramas, bigramas y trigramas de cada libro según la materia a la que perteneciera. Esto para disminuir el número de candidatos a término comunes entre libros a validar y tener al final una lista de términos validados con todas las posibles construcciones analizadas.

En la Tabla 17 se da a conocer un extracto de los candidatos a término para la materia de matemáticas de bachillerato; en la Tabla 18 se muestra un extracto para ecología de bachillerato y finalmente en la Tabla 19 se expone un extracto de las listas obtenidas por el extractor para matemáticas de primaria. Cabe aclarar que en estas tablas se mezclaron unigramas, bigramas y trigramas de los libros que conforman cada una de las materias analizadas.

Matemáticas de Bachillerato			
cuadrado	<i>0.652319739</i>	codominio	<i>0.073741693</i>
método de simpson	<i>0.450862706</i>	isomorfismo	<i>0.057477441</i>
cónica	<i>0.226610224</i>	ley de senos	<i>0.036520337</i>
triángulo oblicuángulo	<i>0.185921002</i>	escondite	<i>0.030788647</i>
binomios	<i>0.153943209</i>	reducir el contaminación	<i>0.01194413</i>
función primitivo	<i>0.150593012</i>	mauricio gómez	<i>0.011891407</i>
función exponencial natural	<i>0.125413295</i>	incluir el fórmula	<i>0.008840447</i>
ecuación trigonométrico	<i>0.092960501</i>	velocidad uniforme	<i>0.007931875</i>
construcción alternativa	<i>0.092859351</i>	voluntario	<i>0.006376907</i>
lado y ángulo	<i>0.09130083</i>	proyector	<i>0.002365373</i>

Tabla 17. Extracto de las listas de candidatos de los libros de matemáticas de bachillerato

Ecología de bachillerato			
ecosistema	<i>0.363877584</i>	fuentes de contaminación	<i>0.037250487</i>
selva tropical	<i>0.182599668</i>	diclorodifeniltricloroetano	<i>0.03127365</i>
bosque	<i>0.146871039</i>	bosque de conífero	<i>0.029398861</i>
ciclo de vida	<i>0.118631532</i>	cambiar radicalmente	<i>0.02102087</i>
ppm	<i>0.107039546</i>	región tropical	<i>0.012179076</i>
índice de natalidad	<i>0.10192139</i>	inagotable	<i>0.009825895</i>
biodiversidad	<i>0.084841923</i>	quemar el hidrocarburo	<i>0.009706804</i>
bosque de coníferas	<i>0.058240796</i>	gota de agua	<i>0.006204729</i>
organismo descomponedor	<i>0.042041732</i>	biogeografía	<i>0.004912947</i>
lechuza	<i>0.040531815</i>	cálculo	<i>0.000179262</i>

Tabla 18. Extracto de las listas de candidatos de los libros de ecología de bachillerato

Matemáticas de primaria			
números romano	<i>0.395719045</i>	línea recto	<i>0.009294068</i>
itzel	<i>0.388469382</i>	diagrama de árbol	<i>0.077376202</i>
decámetro	<i>0.257031928</i>	expropiación petrolero	<i>0.073713457</i>
multiplicación	<i>0.182761216</i>	porcentaje	<i>0.065171074</i>
mónica e itzel	<i>0.178761075</i>	encerrar el número	<i>0.050622666</i>
probabilidad	<i>0.141134807</i>	tabla de frecuencia	<i>0.038387955</i>
juanita	<i>0.118100483</i>	rombo	<i>0.035094508</i>
fracción impropio	<i>0.113304833</i>	derecha exclusivo	<i>0.028265639</i>
mínimo común múltiplo	<i>0.104221928</i>	desgraciadamente	<i>0.017478482</i>
eje de simetría	<i>0.102802493</i>	aproximación	<i>0.006075622</i>

Tabla 19. Extracto de las listas de candidatos de los libros de matemáticas de primaria

Como se puede observar en las tablas anteriores existen algunos errores que se deben a la lematización por parte de FreeLing o por problemas del texto original. Por ejemplo, en la Tabla 17 existe el caso de “construcción alternativa”, el cual no concuerda con su lema, el cual debería ser “construcción alternativo”; la razón de ello es que FreeLing consideró la palabra “alternativa” como un sustantivo porque el bigrama era un renglón del texto original y no tenía más contexto para determinar que era un adjetivo. En cambio, en la Tabla 18 se muestra el caso de “bosque de conífero” y “bosque de conifera”; el primero de ellos sí está bien lematizado, el segundo no porque le faltó un acento a la palabra “conifera” y por tanto FreeLing no pudo encontrar el lema correcto.

Además de los errores de lematización, en las tablas mostradas anteriormente existen casos en los que los candidatos a término no son en realidad posibles términos, como “desgraciadamente”, el cual es un adverbio, o “lado y ángulo” que son dos sustantivos unidos por una conjunción. Este tipo de error se debe a que se usó conocimiento estadístico para extraer los candidatos a término y no uno lingüístico o híbrido que permitiera filtrar las construcciones que posiblemente no eran candidatos a término.

Se puede observar de igual forma en las tablas que algunos de los candidatos a término no pertenecen a las materias donde se encontraron, como “juanita”, “expropiación petrolera” en matemáticas de primaria, o “reducir el contaminación” en matemáticas de bachillerato. Esto se debe por que los libros empleados en el COCIEM son de carácter educativo y frecuentemente se emplean ejemplos en el cual se emplean nombres comunes de personas o se tienen cuadros del estilo de ¿sabías qué? que incluyen información adicional o relacionada con otras materias.

En la Tabla 20 se muestra el número de candidatos a término sin repetición para cada umbral determinado para los pesos de TF-IDF. A partir de la tabla se puede verificar que la gran mayoría de los candidatos término tiene un peso de TF-IDF menor a 0.03, es decir, un peso bajo en una escala del 0 al 1. Además se puede observar que una gran parte de los candidatos a término a validar sin repetición de matemáticas de primaria tiene un peso de TF-IDF mayor o igual a 0.01 a diferencia de ecología y matemáticas de bachillerato donde poco más de la mitad de candidatos a término tiene un peso de TF-IDF menor a 0.01.

Candidatos a términos a validar sin repetición			
Materia	TF-IDF > 0	TF-IDF ≥ 0.01	TF-IDF ≥ 0.03
Matemáticas de bachillerato	<i>33,944</i>	<i>19,582</i>	<i>4,561</i>
Ecología de bachillerato	<i>19,304</i>	<i>9,428</i>	<i>583</i>
Matemáticas de primaria	<i>13,851</i>	<i>13,062</i>	<i>4,081</i>

Tabla 20. Número de candidatos a término a validar sin repetición por materia para cada uno de los umbrales de TF-IDF determinados

4.3 Obtención de términos validados por Wikipedia

Para llevar a cabo la validación de los candidatos a término por parte de Wikipedia, fue necesario en una primera instancia la selección de las fronteras de dominio para cada una de las materias que se analizaron, es decir, para matemáticas de primaria y ecología y matemáticas de bachillerato.

El procedimiento de selección de las fronteras de dominio se llevó a cabo de manera manual, buscando en la base de datos de Wikipedia las posibles categorías que serían equivalentes o que contendrían las materias analizadas. En la Tabla 21 se muestran las fronteras de dominio empleadas para cada una de las materias que se validaron.

Materia	Fronteras de dominio
Matemáticas de primaria	Matemáticas, Álgebra, Aritmética y Geometría
Matemáticas de bachillerato	Matemáticas, Álgebra, Cálculo y Geometría
Ecología de bachillerato	Ecología, Medio ambiente y Climatología

Tabla 21. Fronteras de dominio para cada una de las materias analizadas

Aunque la mayoría de las fronteras de dominio pertenecen a la primera frontera de dominio que aparece en cada renglón de la Tabla 21, se decidió indicar de manera más específica las fronteras de dominio, porque conforme se realizó el procedimiento de validación se observó que en determinados casos era necesario aumentar el límite de caminos a analizar³⁵ para encontrar la relación entre el candidato a término y la frontera de dominio; este aumento en el límite causaba que en ocasiones no se pudiera terminar la validación por el exceso de información que se generaba. Por consiguiente, para no aumentar el número máximo de caminos se decidió aumentar el número de categorías, que aunque son redundante no afectan el desempeño del programa.

Con las fronteras de dominio establecidas se procedió a validar cada una de las listas de candidatos a término dadas a conocer en la sección 4.2. El proceso de validación consiste básicamente en el cálculo del coeficiente de dominio, el cual como se vio en la sección 3.4.2, puede ser calculado de tres formas que son las siguientes:

- Número de caminos (CDnc)
- Longitud de caminos (CDlc)
- Longitud media de caminos (CDlmc)

³⁵ El límite de caminos a analizar previene que la validación de un término caiga en un ciclo infinito. En el caso de esta tesis el límite es de 500 caminos antes de detener la validación y proseguir con el siguiente candidato a término.

La información para el cálculo de estos coeficientes de dominio, la cual se explicó de igual forma en la sección 3.4.2, consiste a grandes rasgos en buscar primero la existencia del candidato a término en Wikipedia, y posteriormente en determinar si éste está ligado a una página de desambiguación o no. Si es este último caso se procede a buscar las categorías con las que está ligado el candidato a término y con cuáles están sus categorías relacionadas a la vez; con la información obtenida de esta búsqueda se calculan los tres coeficientes de dominio. De encontrarse ligado a una página de desambiguación se analizan cada una de las entradas polisémicas de forma normal, es decir, como si fuera otro candidato a término, y el que obtenga el mayor valor del coeficiente de dominio, sin importar la forma de ser calculado, se considera como la acepción buscada.

Además cabe recordar que el coeficiente de dominio (CD) tiene cuatro rangos, los cuales son los siguientes:

- **$CD = 1.00$** : El candidato a término es un término de la frontera de dominio.
- **$0.00 < CD < 1.00$** : El candidato a término está medianamente relacionado con la frontera de dominio pero se le considera como un término de la frontera de dominio.
- **$-1.00 < CD \leq 0.00$** : El candidato a término no está relacionado con la frontera de dominio y por tanto no es un término.
- **$CD = -1.00$** : El candidato a término no se encontró en Wikipedia y no se sabe si es un término o no de la frontera de dominio.

En la Tabla 22 se muestra el número de términos validados para cada una de las materias analizadas según los distintos umbrales establecidos para los pesos de TF-IDF. Se puede observar que por lo general los coeficientes CD_{nc} y CD_{lc} obtienen un número similar de términos validados, mientras que CD_{mc} obtiene aproximadamente un tercio menos de términos validados. Además, el número de términos validados aumenta conforme se disminuye el umbral de TF-IDF, esto por el aumento de candidatos a término a validar; cabe observar que estos aumentos no son tan grandes como los mostrados en la Tabla 20, la cual muestra los candidatos a término a validar sin repetición.

Términos validados									
Peso	TF-IDF > 0			TF-IDF ≥ 0.01			TF-IDF ≥ 0.03		
Métrica	CDnc	CDlc	CDlmc	CDnc	CDlc	CDlmc	CDnc	CDlc	CDlmc
Materia									
Matemáticas de bachillerato	2,909	2,830	1,806	1,575	1,524	1,099	591	577	454
Ecología de bachillerato	644	707	638	325	337	309	96	93	88
Matemáticas de primaria	1,748	1,698	901	1,368	1,334	736	535	517	310

Tabla 22. Número de términos validados por materia para cada uno de los umbrales de TF-IDF determinados

En la Tabla 23 se muestra el número de candidatos a término que tenían su respectivo artículo en Wikipedia pero no se les pudo calcular el coeficiente de dominio porque las páginas de los artículos no estaban ligados a ninguna categoría. A estos casos se les asigna un coeficiente de dominio igual a cero, el cual indica que fueron encontrados en Wikipedia pero no están relacionados con la frontera de dominio, en este caso porque no se pudo saber si pertenecían o no a ella.

Candidatos a término encontrados en Wikipedia pero sin categoría asignada									
Peso	TF-IDF > 0			TF-IDF ≥ 0.01			TF-IDF ≥ 0.03		
Tipo	Unigrama	Bigrama	Trigrama	Unigrama	Bigrama	Trigrama	Unigrama	Bigrama	Trigrama
Materia									
Matemáticas de bachillerato	289	80	28	128	60	19	46	28	5
Ecología de bachillerato	227	32	27	98	24	14	12	6	3
Matemáticas de primaria	172	30	11	132	30	11	37	19	6

Tabla 23. Número de candidatos a término que fueron encontrados como artículos en Wikipedia pero que no estaban asignados a ninguna categoría

Como se puede observar en la Tabla 23, la mayor cantidad de páginas de Wikipedia no asignadas a ninguna categoría están relacionadas con unigramas. Algunos ejemplos de candidatos a términos sin una categoría asignada son: “determinante”, “módulo”, “culiacán”, “ecuación lineal”, “señal de tránsito” o “integración por sustitución” en matemáticas de bachillerato; en ecología de bachillerato “pirámides”, “transgénico”, “control poblacional”, “cuenca hidrológico”, “valor monetario”, “campo de estudio” o “grano de polen”; mientras que en matemáticas de primaria “jugo”, “área”, “imperio inca”, “fracción equivalente”, “paleta de agua” o “suma de fracción”. Entre los ejemplos dados anteriormente existen algunos candidatos a término que realmente son términos de la materia a la que están adjudicados.

A continuación en la Tabla 24 se muestran los primeros 20 términos validados de matemáticas de bachillerato para cada umbral establecido de TF-IDF, donde cabe aclarar que se obtuvieron los mismos términos para cada uno de los métodos del cálculo del coeficiente de dominio. En cambio, en la Tabla 25 se muestran los primeros 20 términos validados de ecología de bachillerato para cada umbral determinado de TF-IDF, donde se observa que para $TF-IDF \geq 0.03$ se obtuvieron distintos términos validados a partir de la posición 16 para cada uno de los métodos del cálculo del coeficiente de dominio. En la Tabla 26 se muestran los primeros 20 términos validados para matemáticas de primaria para cada umbral establecido de TF-IDF; en este caso para $TF-IDF \geq 0.01$ varía para el cálculo de CDlc la lista de términos en la posición 19 y 20.

En general los primeros lugares de las listas de términos validados muestran términos que están altamente relacionados con la materia a la que se les adjudica. Sin embargo, existen casos en los que se aceptan términos con faltas de ortografía o que tienen una cierta ambigüedad.

Matemáticas de bachillerato		
TF-IDF>0	TF-IDF ≥ 0.01	TF-IDF ≥ 0.03
adición	adición	aritmética
afinidad	álgebra	asociatividad
álgebra	aproximación lineal	binomio
ancho	aritmética	binomio conjugado
aproximación lineal	asociatividad	binomio de newton
aritmética	binomio	bisectriz
asociatividad	binomio conjugado	cardioides
astroide	binomio de newton	caso de factorización
bidimensional	bisectriz	cifra
binomio	cálculo diferencial	circunferencia
binomio conjugado	caracol de pascal	circunferencia inscribible
binomio de newton	cardioides	circunferencia unitario
bisectriz	caso de factorización	cociente
cálculo diferencial	centena	combinaciones
caracol de pascal	centro	conjunto
característica	cicloide	construcción geométrica
cardioides	ciencia matemática	corona circular
caso de factorización	cifra	cuantificador existencial
casquete esférico	circunferencia	cubo
cateto	circunferencia inscribible	cuerpo geométrico

Tabla 24. Los 20 primeros términos de matemáticas de bachillerato para cada umbral empleado de TF-IDF

Ecología de bachillerato				
TF-IDF>0	TF-IDF ≥ 0.01	TF-IDF ≥ 0.03		
abiótico	abiótico	abiótico		
acaricida	acaricida	acaricida		
amensalismo	amensalismo	amensalismo		
biodegradabilidad	biodegradabilidad	biodegradación		
biodegradación	biodegradación	biótico		
biótico	biótico	clima		
clima	clima	contaminación		
clima árido	clima árido	ddt		
clordano	clordano	descomponedor		
clímax	clímax	diclorodifeniltricloroetano		
comensal	comensal	explotación		
condición climático	condición climático	hábitat		
conservación biológico	conservación biológico	neutralismo		
contaminación	contaminación	protooperación		
contaminante	corriente oceánico	semiárido		
corriente de humboldt	crisis ecológico	comunidad	nutriente	territorial
corriente marino	cálido	continental	vertido	vertido
corriente oceánico	ddt	ecológico	migratorio	naturaleza
crisis ecológico	descomponedor	entropía	naturaleza	migratorio
cálido	desecho tóxico	selva	territorial	ambiental
		CDnc	CDlc	CDlmc

Tabla 25. Los 20 primeros términos de ecología de bachillerato para cada umbral empleado de TF-IDF. En el caso TF-IDF \geq 0.03 se muestran a partir de la posición 16 los distintos términos validados ordenados según el cálculo del coeficiente de dominio.

Matemáticas de primaria			
TF-IDF>0	TF-IDF ≥ 0.01		TF-IDF ≥ 0.03
ancho	ancho		ancho
antecesor	antecesor		antecesor
arista	arista		cardinalidad
aritmética	aritmética		centena
billón	billón		cifra
cara	cara		cúbico
característica	característica		decena
cardinalidad	cardinalidad		diámetro
centena	centena		dígito
centro	cifra		elementos
cifra	compacto		figura
combinaciones	completar el cuadrado		figura geométrico
compacto	cuerpo geométrico		fracción impropio
completar el cuadrado	cuerpos geométricos		heptágono
cubo	curva		hexágono regular
cuerpo geométrico	cúbico		lado
cuerpos geométricos	decena		longitud
curva	decágono		líneas
cúbico	denominador	dodecaedro	matemática
decena	diámetro	denominador	minuendo
	CDnc y CDlmc	CDlc	

Tabla 26. Los 20 primeros términos de matemáticas de primaria para cada umbral empleado de TF-IDF. En el caso TF-IDF \geq 0.01 se muestran a partir de la posición 19 los distintos términos validados ordenados según el cálculo del coeficiente de dominio.

Se puede observar en la Tabla 24 que uno de los primeros términos validados es “algebra” sin acento, esto puede ser porque hubo un error en la digitalización del libro o en el libro mismo, pero Wikipedia tiene una redirección a “álgebra”; cabe aclarar que “álgebra” también aparece, solamente que algunos lugares más abajo pero también con un coeficiente de dominio igual a 1. Con respecto a la Tabla 25 no muestra errores como los anteriores, pero tiene la peculiaridad de que cambia el orden de algunos términos validados para el umbral $TF-IDF \geq 0.03$ dependiendo de la forma en que se calculó el coeficiente de dominio. En la Tabla 26 se muestra que existe un error, posiblemente por la digitalización del libro o su escritura, el cual es “denominador” el cual fue encontrado en Wikipedia porque tiene un redireccionamiento hacia “denominador”.

De igual forma, se pueden observar que existen casos en el cual la lematización no fue la correcta, como en “aritmética” o “cuerpos geométricos”, aunque sí fueron localizados en Wikipedia gracias a su estructura.

4.4 Evaluación de resultados

La evaluación de los sistemas extractores de términos es uno de los procesos más importantes que se deben llevar siempre a cabo. En esta tesis, la evaluación del sistema extractor desarrollado no es una excepción; para ello se emplean las métricas más empleadas en los proyectos relacionados con el PLN, que son precisión y cobertura.

Una de las razones por las cuales sólo se eligieron tres materias para llevar a cabo la obtención de los términos fue la falta de listas validadas por expertos para la evaluación de los resultados. Para matemáticas de primaria se obtuvo una lista de términos obtenida por el método de lista de referencia (sección 2.3.1), es decir, se analizó de manera manual el corpus, más específicamente los libros de matemáticas de primaria, para la extracción de los términos que se encontraban en esta sección del COCIEM.

En el caso de matemáticas de bachillerato, se empleó una lista de términos que fue obtenida a través del método de validación (sección 2.3.2); esta lista es la misma que fue empleada en el proyecto descrito en el artículo de Cabrera-Diego et al. (2011), donde estudiantes de ingeniería y lingüística analizaron los candidatos a término que se obtuvieron

empleando el extractor terminológico del programa WordSmith Tools 5.0 (Scott, 1996) que se basa en el logaritmo de la verosimilitud³⁶.

Finalmente, en el caso de ecología de bachillerato, se empleó igualmente el método de validación para la obtención de una lista de términos; a diferencia de la lista de matemáticas de bachillerato, en este caso se empleó la salida generada por el extractor terminológico desarrollado en esta tesis, donde un biólogo validó cada uno de los n-gramas existentes en los libros. Cabe aclarar que para ello no se le dio ninguna información sobre los pesos obtenidos en el proceso de extracción, esto para no sesgar los posibles resultados.

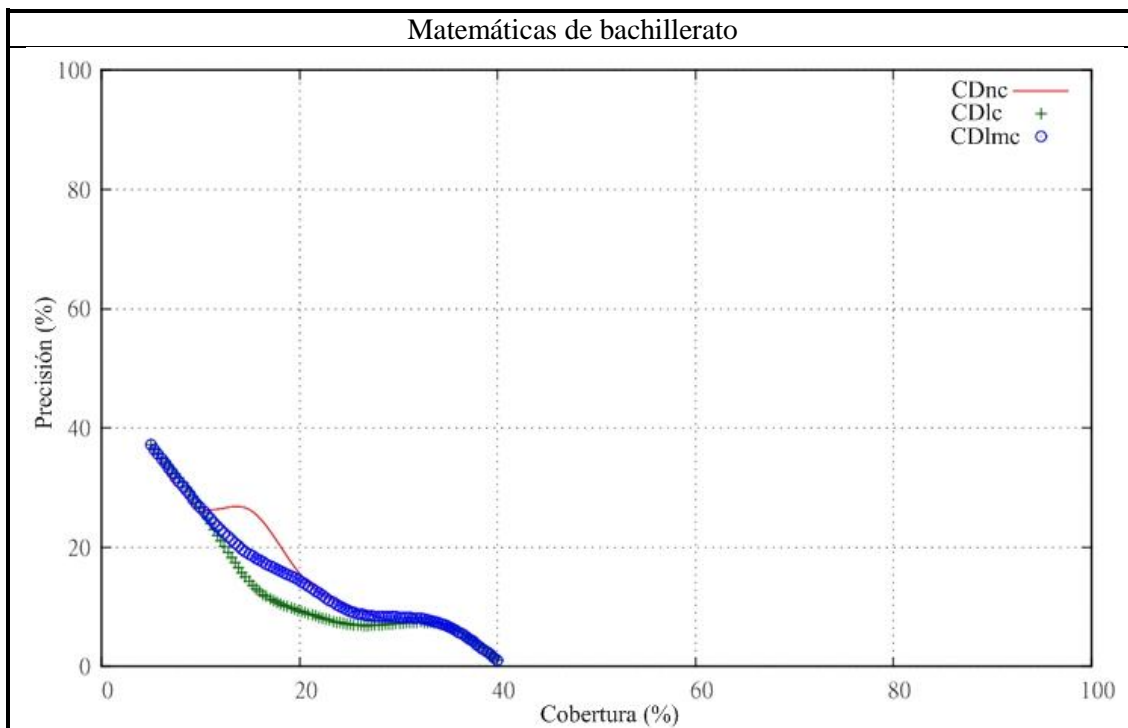
La evaluación de los resultados obtenidos por la validación de Wikipedia se llevó a cabo para unigramas, bigramas, trigramas y los tres n-gramas unidos. Para llevar a cabo esto, primero se lematizaron las listas de términos creadas por los expertos de matemáticas de bachillerato y primaria siguiendo un método similar al empleado para lematizar la base de datos de Wikipedia (sección 3.4.1.2), la única diferencia es que en este caso no existe un identificador; en el caso de la lista de términos de ecología de bachillerato esta no fue lematizada, ya que ya se encontraba en ese estado por provenir de la salida del análisis de TF-IDF. Posteriormente, las listas de términos de los expertos se dividieron en tres partes, en unigramas, bigramas y trigramas, en esta última parte se agregaron también los n-gramas que tuvieran un tamaño mayor que 3 en el caso de matemáticas de bachillerato y de primaria. Finalmente, se comparan las listas de términos validados y las listas de términos creados por los expertos; cabe aclarar que los términos validados con coeficientes de dominio mayores a cero se consideran términos y en teoría deberían ser encontrados en las listas de términos de los expertos. Con base en la información obtenida se calculan los valores de precisión y cobertura con base en las Ecuaciones (17) y (18), las cuales fueron mostradas en la sección 1.2.3.

Esta evaluación se llevó a cabo para verificar la precisión y cobertura de las listas de términos validados por Wikipedia para n-gramas de manera conjunta, es decir, los tres tipos

³⁶ El logaritmo de la verosimilitud consiste en comparar dos corpus o documentos, el primero de ellos es de donde se extraerán los términos y el otro de referencia. Para obtenerlo se emplean las frecuencias relativas de cada palabra en ambos corpus y el número total de palabras en ellos (Cabrera-Diego et al., 2011).

de n-gramas unidos, y para cada tipo de ellos (unigramas, bigramas y trigramas). Asimismo, la evaluación se llevó a cabo para conocer cómo cambiaban los resultados con base en los diferentes umbrales de pesos de TF-IDF designados (sección 4.2).

En la Figura 21 se muestran las gráficas de cobertura contra precisión de las materias analizadas para la unión de los tres n-gramas con pesos de TF-IDF mayores a cero; en ellas se muestran tres líneas que representan las diferentes formas de calcular el coeficientes de dominio (sección 3.4.2). En los Anexos B, C y D se pueden encontrar todas las gráficas de cobertura versus precisión que se generaron para la evaluación. En estas gráficas, el eje de las abscisas representa la cobertura, es decir, el porcentaje de términos que se encuentran tanto en la lista de términos validados como en la lista de términos de los expertos; en cambio, el eje de las ordenadas representa la precisión, en otras palabras, el porcentaje de términos validados que son realmente términos según la lista de términos de los expertos.



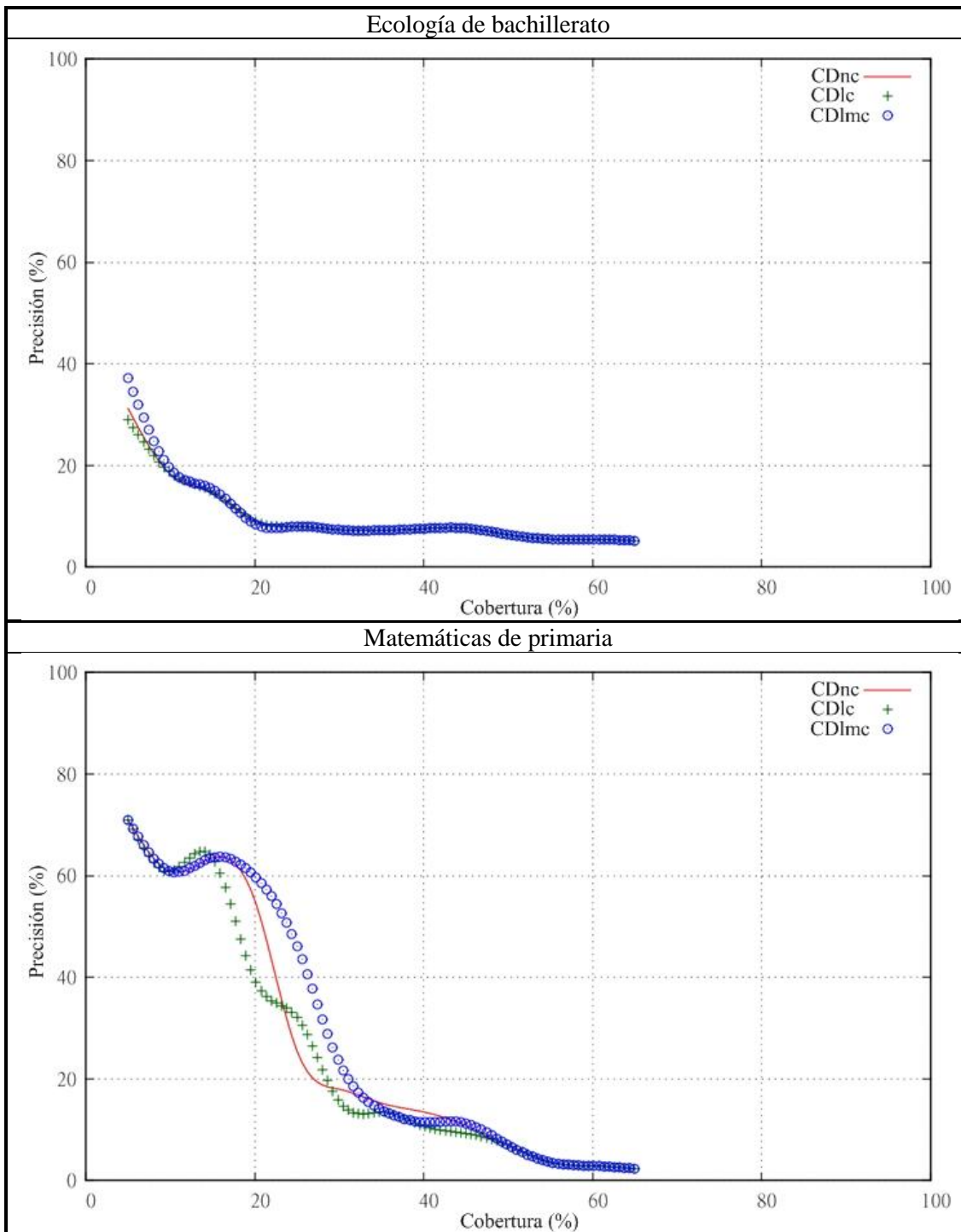


Figura 21. Gráficas de precisión y cobertura para todos los n-gramas generados para pesos de TF-IDF mayores a cero

Como se puede observar en la Figura 21, hay tres gráficas de cobertura versus precisión. La primera de ellas corresponde a matemáticas de bachillerato, donde se puede ver que los coeficientes de dominio CDnc y CDlmc empleados para el cálculo del coeficiente de dominio tuvieron mejores resultados que el coeficiente de dominio CDlc entre un 10% y 30%

de cobertura; para los tres coeficientes de dominio el máximo de precisión fue 37.14% para una cobertura del 5%, en cambio la cobertura máxima fue de 40% con una precisión de 1.05%. La segunda gráfica corresponde a ecología de bachillerato, en donde los tres métricas para el cálculo del coeficiente de dominio tuvieron un comportamiento similar; CDlc tuvo una precisión máxima de 29.02%, CDlmc de 37.19% y CDnc de 31.22% para una cobertura del 5%, mientras que los tres coeficientes tuvieron una cobertura máxima del 65% con una precisión de 5.21%. La tercera gráfica es de la evaluación de matemáticas de primaria, en la cual se puede observar que los tres cálculos del coeficiente de dominio tuvieron un comportamiento distinto entre un 10% y 30% de cobertura, siendo la métrica CDlmc el de mejor desempeño; de manera más específica las tres medidas obtuvieron una precisión máxima del 70.97% para un 5% de cobertura, mientras que se obtuvo una máximo de cobertura en 65% con una precisión del 2.33%. En todos los casos la métrica con peor rendimiento en general es la basada en la longitud de caminos (CDlc).

Finalmente, observando la Figura 21 se puede decir que la validación con Wikipedia con peor rendimiento fue la de matemáticas de bachillerato y la mejor la de matemáticas de primaria, al menos para la unión de unigramas, bigramas y trigramas y pesos de TF-IDF mayores a cero. Además, todas las gráficas muestran en un mayor o menor grado una tendencia similar a la gráfica típica de precisión y cobertura, que fue mostrada en la sección 1.2.3.

A pesar de lo anterior, es necesario analizar las razones de los resultados obtenidos y por ello en la siguiente sección se presentará una serie de observaciones sobre la evaluación realizada en esta tesis.

4.4.1 Observaciones de la evaluación de resultados

A partir de los resultados mostrados en la sección anterior, se establecieron ciertos comentarios, los cuales serán expuestos a continuación. No sólo se abordarán las gráficas de cobertura versus precisión, sino también las listas de validación y las listas de términos que se emplearon en la evaluación.

Como se pudo observar en la Figura 21, el mejor resultado de precisión se obtuvo en matemáticas de educación primaria (70.97%). En cambio, en cobertura los mejores

resultados fueron para ecología de bachillerato y matemáticas de primaria, ambos con una cobertura del 65%.

Sin embargo, debido a los resultados no tan buenos en el caso de matemáticas de bachillerato, especialmente, se decidió analizar la lista validada de términos que se empleó en la evaluación de los resultados. En este análisis se pudo observar que faltan términos en la lista validada por expertos, por ejemplo: “campana de Gauss”, “ecuación punto-pendiente”, “matriz triangular superior”, “antidiferencial”, entre otros. Esto ciertamente afecta a la precisión, debido a que la aparición de términos en la lista validada por Wikipedia pero no en la lista de evaluación (creada por los expertos) es como si se hubieran aceptado muchos no términos que lo eran en realidad. Asimismo se observó que la lista de los expertos tenía plurales que al ser lematizados en ciertas ocasiones no se volvían singulares por problemas de FreeLing, creando términos que no se encontraban de ninguna manera en las listas de candidatos de término o en las validadas por Wikipedia, provocando una falta de cobertura. Ejemplo de este caso es “ángulos interno”, el cual aparece en la lista de expertos pero no en la de candidatos a término; el más cercano, y correcto, es el de “ángulo interno”. También en la lista validada por expertos no se tenían en todos los casos los sinónimos de un término, aun cuando estos estaban en los documentos de matemáticas de bachillerato; por ejemplo, en las listas de los expertos se encontraba ecuación punto-pendiente y lado recto, pero no sus sinónimos fórmula punto-pendiente y ancho focal.

Por los problemas vistos en la lista de términos de matemáticas de expertos para bachillerato, se decidió hacer una validación de las listas obtenidas por el extractor terminológico descrito en esta tesis; para llevar esto a cabo no se utilizó de ninguna forma la información provista por la lista de términos validados por Wikipedia o de la lista de pesos de TF-IDF, esto para no sesgar la lista de validación. En la Figura 22 se muestra la gráfica de cobertura versus precisión empleando la segunda lista validada de matemáticas de bachillerato y la lista de candidatos con pesos de TF-IDF mayores a cero para la unión de los tres n-gramas empleados en esta tesis. En el Anexo E se pueden encontrar las demás gráficas generadas para pesos de TF-IDF mayores a cero.

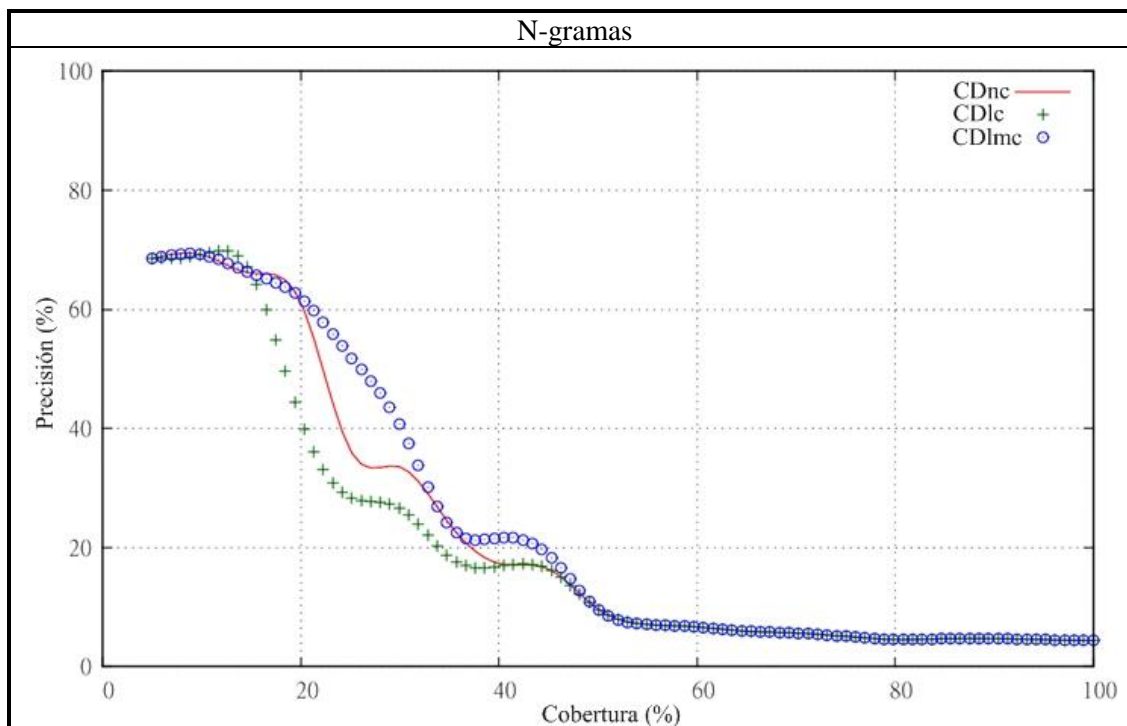


Figura 22. Gráfica de cobertura versus precisión de matemáticas de bachillerato usando la lista validada de candidatos a término obtenida del extractor terminológico desarrollado en esta tesis para n-gramas con pesos de TF-IDF mayores a cero.

Aunque, los resultados mostrados en la Figura 22 no se consideran como resultados, debido a que no se puede ser juez y parte, ciertamente la falta de términos y los errores en la lista validada por expertos afectó seriamente los resultados obtenidos en matemáticas de bachillerato. Asimismo se puede observar que al igual que en la gráfica de matemáticas de bachillerato de la Figura 21, la métrica con peor rendimiento es la que se basa en la longitud de caminos (CDlc).

Con respecto a la lista de términos de Ecología validada por un experto, existen algunos errores que son intrínsecos al trabajo manual que se realizó y algunas variaciones entre lo que se considera término ciertas veces. Sin embargo, no se observaron casos que pudieran afectar el cálculo de la precisión y cobertura en gran medida.

Sobre la lista de referencia de términos de matemáticas de primaria, la cual fue realizada por un terminólogo con conocimientos de matemáticas, sufrió de igual manera algunos problemas al llevarse a cabo la lematización; por ejemplo, la palabra “muestra” pasó a “mostrar”. Sin embargo, la lista de términos otorgada por el experto tiene algunos casos que desde mi punto de vista no son términos, por ejemplo, “al azar”, “altura de un triángulo”

o “descomposición de un número en productos de dos o más factores”. De igual forma, contiene términos que, por su estructura, no pueden ser detectados por el extractor terminológico o son eliminados en el proceso anterior al cálculo del IDF (sección 3.3.2), como los signos “<” y “%”, o como “mayor que” y “a escala”.

Además de analizar las listas que se emplearon para evaluar el sistema de extracción terminológica y su validación usando Wikipedia, se procedió a revisar las listas validadas por la enciclopedia. En esta revisión se encontraron diversos problemas que son propios de la estructura de Wikipedia; por ejemplo, se observó que había términos que se indicaban que eran de matemáticas cuando claramente no lo eran, como “mandarina” o “fiesta de disfraz” para matemáticas de primaria. Por medio de la generación de sus grafos, como el mostrado en la sección 3.4.2, se observó que la unión de ambos términos a la categoría de “matemáticas” era la categoría de “dimensión” que estaba conectada a su vez con la categoría “tiempo”. En la Figura 23 se muestra una sección de los grafos generados para “fiesta de disfraz” y “mandarina”, donde se puede apreciar la conexión de estos a la categoría tiempo. También en algunos casos las páginas de desambiguación de Wikipedia ayudaban, como en “razón” y “dividendo” para matemáticas o “residuo” para ecología, en algunos otros no lo hacía. Por ejemplo, en el caso de matemáticas la palabra “nulo” se desambiguaba como “voto nulo”, “rango” como “Rango (película)”; mientras que en el caso de ecología “reproducción” se desambiguaba como “reproducción (economía)”.

De igual manera, dentro de las listas de términos validados por Wikipedia, existen términos validados que, a pesar de formar parte de la materia y tener un coeficiente de dominio mayor a cero, en realidad fueron encontrados como otros términos, pero que de algún modo están relacionados con la materia. Por ejemplo, en matemáticas “ft” y “arreglo” son marcados como términos³⁷, los cuales sí los son en teoría, pero en realidad estos fueron encontrados como “arreglo musical” y “featuring” que de alguna manera, posiblemente por casos similares a los mostrados en la Figura 23, están relacionados con matemáticas.

³⁷ Ambos términos tienen un coeficiente de dominio igual a 0.75 y son de la lista de matemáticas de primaria para $TF-IDF > 0$ usando la métrica CDlc.

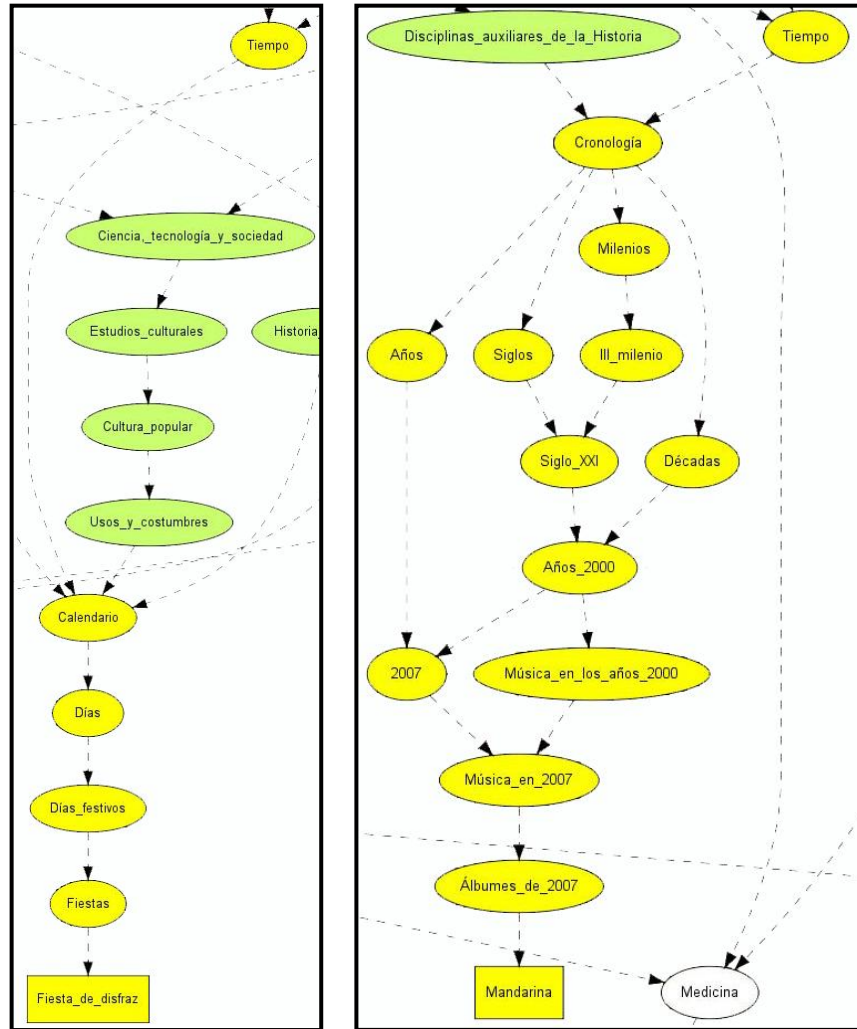


Figura 23. Secciones de los grafos de “Fiesta de cumpleaños” (izquierda) y “Mandarina” (derecha)

También como se dio a conocer en la sección 4.3 existen casos en el cual las páginas de artículos de Wikipedia no están asignadas a ninguna categoría, ya sean por errores u omisiones al momento de crear el artículo. Esto afecta la cobertura, pues hay términos que pertenecen realmente a la frontera de dominio pero que al no estar asignados a ninguna categoría no pueden tomarse como términos al momento de realizar la validación con Wikipedia, pues se les asigna un coeficiente de dominio igual a cero.

5 CONCLUSIÓN Y TRABAJO A FUTURO

En esta tesis se presentó un extractor terminológico basado en conocimiento estadístico, junto con un sistema que emplea la enciclopedia Wikipedia como recurso léxico para validar los candidatos a término. Con ello se cumplió el objetivo principal y el objetivo secundario:

- Se desarrolló un sistema extractor terminológico basado en conocimiento estadístico, más específicamente, en Term Frequency – Inverse Frequency Document (TF-IDF). El cual trae la mayor cantidad de términos posibles para obtener mejores resultados de cobertura.
- Se empleó Wikipedia como un recurso léxico para llevar a cabo la validación de cada uno de los candidatos a término que se obtuvieron del extractor terminológico.
- Se obtuvieron tres listas de términos validadas para las materias de matemáticas y ecología de bachillerato, así como la de matemáticas de educación primaria.

La realización de este trabajo de tesis permitió llegar a varias conclusiones que a continuación se describirán siguiendo la estructura de la tesis, es decir, primeramente se comenzará con la del corpus COCIEM, luego la del extractor terminológico, posteriormente con la de la validación empleando Wikipedia y finalmente la de los resultados obtenidos.

El Corpus de textos científicos en español de México (COCIAM) ha sido un gran corpus de prueba para llevar a cabo la extracción terminológica y su validación empleando Wikipedia. La razón de lo anterior es que abarca una gran cantidad de materias científicas, que van de nivel primaria a nivel bachillerato, ricas en términos que pueden ser validados empleando recursos léxicos no tan especializados o de alto nivel. Sin embargo, considero que tiene sus deficiencias como palabras mal acentuadas, la inclusión de símbolos extraños o errores en la estructura del libro debido al escaneado (partes del texto a doble columna, discontinuidad por cuadros del estilo ¿sabías qué?, etcétera), los cuales pudieron ser prevenidos en el momento del escaneado.

Con respecto al extractor terminológico ciertamente existen deficiencias en él, debido a que no se pueden obtener ciertos términos, como lo son los que empiezan con artículos, por

ejemplo “El Niño”, el cual es un fenómeno climatológico, lo anterior debido a la limpieza que se llevó a cabo en las listas de candidatos a término antes del cálculo del IDF que consistió en la eliminación de n-gramas con palabras vacías en sus extremos. Asimismo, la clasificación de los candidatos a término por los pesos de TF-IDF no fue tan buena como se hubiera esperado, pues la mayoría de los candidatos quedaban en pesos medios y bajos, y no altos, posiblemente por el alto número de palabras externas a las materias del COCIEM por su carácter educativo, como lo son el empleo de ejemplos que usan la vida cotidiana, la narración de diálogos o historias con nombres de personas para dar a conocer un tema o la inclusión de cuadros del estilo de ¿sabías qué? que describen información adicional o relacionada con otras materias para mejorar la comprensión del tema. Sin embargo, pienso que el extractor cumplió su cometido trayendo la mayor cantidad de candidatos a término posibles para tratar de tener la mayor cobertura; además creo que el uso de los pesos de TF-IDF puede ser de gran ayuda si se lleva a cabo un análisis más profundo de los pesos que otorga para calcular umbrales mucho más específicos, y por tanto separar de mejor manera los posibles candidatos a términos de los que no tienen posibilidad de serlo, ya sea por su poca relevancia en el documento o por su poca repetición en el corpus de análisis.

La validación de los candidatos a término fue un método que no sólo permitió validar los resultados obtenidos del extractor terminológico, sino que complementó en una gran medida el uso de un método estadístico para la extracción de términos, debido a que descartó un gran número de candidatos a término que ciertamente no eran términos por la estructura morfosintáctica que tenían, algo que sólo se puede realizar empleando conocimiento lingüístico en la etapa de extracción.

Los resultados, en el caso de ecología de bachillerato y matemáticas de primaria son buenos, en el caso de matemáticas de bachillerato no son tan buenos como los anteriores, pero por la prueba realizada con otra lista de evaluación se muestra que sus resultados no tan buenos fueron principalmente por la lista de evaluación original más que por la validación realizada por Wikipedia.

Ciertamente, los resultados que se obtuvieron pudieron ser mejores, pero algunos de los errores que ocurrieron en la validación empleando Wikipedia son intrínsecos a la estructura de la enciclopedia y no tanto al sistema extractor de términos. En este caso existen

ejemplos en donde la estructura enciclopédica de Wikipedia afecta los resultados obtenidos, por tener categorías como “Cultura por país”, “Biografías por actividad”, entre otras, que unen ciertos temas con categorías que no están totalmente relacionadas, como los mostrados en la sección 4.4.1. De igual forma, al ser Wikipedia una enciclopedia creada por las personas, existen errores u omisiones en la asignación de las categorías, los cuales afectan el proceso para el cálculo del coeficiente de dominio. Es necesario, de igual forma, incluir en esta parte lo de la desambiguación de términos, pues aunque en algunos casos obtenía el término correcto, en algunos otros no lo hacía, debido al método elegido (término con máximos coeficientes de dominio) y no tanto a Wikipedia en sí. A pesar de todo lo anterior, considero que Wikipedia es un excelente recurso que puede ser empleado en la validación de candidatos a término porque abarca una gran cantidad de materias (o categorías) y de temas que no se encuentran frecuentemente en un solo recurso; además de que Wikipedia está en constante crecimiento y actualización. Solamente, es necesario que se estudie más a fondo la estructura de Wikipedia para omitir ciertas características propias de ella que a veces afectan los resultados y mejorar la manera en que se desambiguan los términos.

Existe una gran cantidad de puntos a trabajar a futuro para mejorar tanto la extracción terminológica como la validación empleando Wikipedia. La primera tarea sería la de mejorar el extractor terminológico, ya sea empleando otra métrica o mezclando varias de ellas para tener distintas clasificaciones y obtener mejores candidatos a término. La segunda tarea sería desarrollar nuevas fórmulas para el cálculo del coeficiente de dominio, para ver si se pueden obtener mejores resultados. Asimismo, como tercera tarea sería encontrar una mejor manera de encontrar los términos correctos que son ambiguos, pues el método que se empleó en esta tesis no es del todo confiable. La cuarta tarea a realizar a futuro sería llevar a cabo un análisis más profundo de la estructura de Wikipedia para encontrar, y omitir posteriormente en el análisis, las categorías que pueden causar casos erróneos en la validación de los candidatos a término. La quinta tarea sería emplear la misma Wikipedia como un recurso que convierta los resultados lematizados a su forma correcta de mostrarse, es decir, que haya concordancia en género y número y no se muestren casos como “fracción impropio” sino como “fracción impropia”. Finalmente, la sexta tarea sería aplicar el método de extracción de términos y su validación empleando Wikipedia a todas las materias que pertenecen al COCIEM.

BIBLIOGRAFÍA

- Ananiadou, S. y McNaught, J., 2006. *Text Mining for Biology and Biomedicine*. Norwood, E.U.A: Artech House.
- Aronson, A.R. y Lang, F.-M., 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the american medical informatics Association*, 17, pp.229-36.
- Barrón-Cedeño, A., Sierra, G., Drouin, P. y Ananiadou, S., 2009. An improved automatic term recognition method for Spanish. En *CICLing*. Ciudad de México, México, 2009.
- Bellomo, T.S., 2009. Morphological Analysis and Vocabulary Development: Critical Criteria. *The Reading Matrix*, 9(1).
- Bourigault, D., 1994. *LEXTER, un Logiciel d'EXtraction de Terminologie. Application à l'acquisition des connaissances à partir de textes*. Paris: école des Hautes Études en Sciences Sociales: PhD Tesis.
- Bourigault, D., Gonzalez-Mullier, I. y Gros, C., 1996. LEXTER, a natural language processing tool for terminology extraction. En *7th EURALEX International Congress*. Göteborg, Alemania, 1996.
- Cabré, M.T., 1992. *La terminología : la teoría, metodología, aplicaciones*. Barcelona, España: Antartida/Empuries.
- Cabré, M.T., 1995. La terminología hoy: concepciones, tendencias y aplicaciones. *Ciência da Informação*, 24(3).
- Cabré, M.T., Estopà, R. y Vivaldi, J., 2001. Automatic term detection: A review of current systems. En D. Bourigault, C. Jacquemin y M.-C. L'Homme, eds. *Recent advances in computational terminology*. John Benjamins Publishing Company.
- Cabrera-Diego, L.A., Sierra, G., Vivaldi, J. y Pozzi, M., 2011. Using Wikipedia to Validate Term Candidates for the Mexican Basic Scientific Vocabulary. En *First International*

- Conference on Terminology, Languages, and Content Resources (LaRC 2011)*. Seúl, 2011.
- Castañeda De Isla Puga, É., 2000. *Geometría analítica en el espacio*. Ciudad de México, México: Facultad de Ingeniería UNAM.
- Cavnar, W.B. y Trenkle, J.M., 1994. N-gram-based text categorization. En *3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94)*. Las Vegas, 1994.
- Cowie, J. y Wilks, Y., 2000. Information Extraction. En *Handbook of Natural Language Processing*. Nueva York, E.U.A: Marcel Dekker.
- Cruse, D.A., 1986. *Lexical semantics*. Mánchester, Reino Unido: Cambridge University Press.
- Daintith, J., 2001. *Diccionario de matemáticas*. Translated by J.M. Castaño. Santa Fé de Bogotá, Colombia: Editorial Norma.
- Euguehard, C. y Pantera, L., 1994. Automatic natural acquisition of a terminology. *Journal of quantitative linguistics*, 2(1), pp.27-32.
- Fellbaum, C., ed., 1998. *WordNet, an electronic lexical database*. Cambridge, E.U.A: The MIT Press.
- Frakes, W.B. y Baeza-Yates, R., 1992. *Information Retrieval: Data Structures &*. Englewood cliffs, E.U.A: Prentice Hall.
- Francis, W.N. y Kučera, H., 1979. *Manual of information to accompany a standard corpus of present-day edited american English, for use with digital computers*. Rhode Island, E.U.A.: Brown University.
- Frantzi, K., Ananiadou, S. y Mima, H., 2000. Automatic recognition of multi-word terms:the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), pp.115-30.
- Gabrilovich, E. y Markovitch, S., 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34, pp.443-98.

- Gelbukh, A. y Sidorov, G., 2006. *Procesamiento automático del español con enfoque en recurso léxicos grandes*. Ciudad de México, México: Instituto Politécnico Nacional.
- Goyal, V. y Singh Lehal, G., 2008. Hindi Morphological Analyzer and Generator. En *International Conference on Emerging Trends in Engineering and Technology*. Nagpur, India, 2008.
- Heid, U., Jauß, S. y Krüger, K.H.A., 1996. Term extraction with standard tools for corpus exploration. Experiencia from German. En *Terminology and Knowledge Engineering TKE'96*. Berlin, Alemania, 1996.
- ISO 1087-1:2000, 2000. *Terminology work – Vocabulary – Part 1: Theory and application*. Génova, Suiza: ISO.
- Jackson, P. y Moulinier, I., 2002. *Natural Language Processing for Online Applications. Text retrieval, extracition and categorization*. John Benjamins Publishing Company.
- Jurafsky, D. y Martin, J.H., 2008. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Segunda edición ed. Upper Saddle, E.U.A: Prentice Hall.
- Kageura, K. y Umino, B., 1998. Methods of automatic term recognition. *Terminology*, 3(2), pp.259-89.
- Kauchak, D., Smarr, J. y Elkan, C., 2002. *Sources of success for information extraction methods*. UCSD.
- Lérat, P., 1989. L'analyse morphologique des termes nouveaux. *La Banque des mots* , pp.23-34.
- Luna Trail, E., Viguera Ávila, A. y Baez Pinal, G.E., 2005. *Diccionario básico de lingüística*. Ciudad de México, México: Instituto de Investigaciones Filológicas, UNAM.
- MacMullen, W.J., 2003. Requirements definition and design criteria for test corpora in information science. *SILS Technical Report TR-2003-03*.

- Manning, C.D., Raghavan, P. y Schütze, H., 2008. *Introduction to Information Retrieval*. E.U.A.: Cambridge University Press.
- Maynard, D., 2000. *Term recognition using combined knowledge sources*. Manchester, Reino Unido: Universidad Metropolitana de Manchester: Ph.D. Tesis.
- Medina, A. y Méndez, C., 2006. Arquitectura del Corpus Histórico del Español en México (CHEM). *Avances en la ciencia de la computación*, pp.248-53.
- Padró, L., Collado, M., Reese, S., Lloberes, M. y Castellón, I., 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. La Valletta, Malta, 2010.
- Pantel, P. y Lin, D., 2001. A statistical corpus-based term extractor. En *14th Biennial conference of the canadian society for computational studies of Intellegence*. Ottawa, Canadá, 2001.
- Pazienza, M.T., Pennacchiotti, M. y Zanzotto, F.M., 2005. Terminology extraction: an analysis of linguistic and statistical approaches. *Knowledge Mining. Studies in Fuzziness and Soft Computing*, 185, pp.255-80.
- Peters, I., 2009. *Folksonomies: Indexing and retrieval in web 2.0*. Berlín, Alemania: De Gruyter Saur.
- Ponzetto, S.P. y Strube, M., 2008. WikiTaxonomy: A large scale knowledge resource. En *18th European Conference on Artificial Intelligence*. Patras, 2008.
- Porter, M.F., 1980. An algorithm for suffix stripping. *Program*, 14(3), pp.130-37.
- Pudota, N., Casoto, P., Dattolo, A., Omero, P. y Tasso, C., 2008. Towards Bridging the Gap between Personalization and Information Extraction. En *Proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL)*. Padua, Italia, 2008.
- Ramos, J., 2003. Using TF-IDF to determine word relevance in document queries. En *First instructional Conference on Machine Learning iCML-2003*. Piscataway, E.U.A, 2003.
- Robertson, S.E., 1972. Term Specificity. *Journal of Documentation*, 28(2), pp.164-65.

- Robertson, S., 2004. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5), p.503–520.
- Salton, G. y Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), pp.513-23.
- Salton, G. y McGill, M.J., 1986. *Introduction to modern information retrieval*. Nueva York, E.U.A: McGraw-Hill.
- Salton, G. y Yang, C.S., 1973. On the specification of term in automatic indexing. *Journal of Documentation*, 29(4), pp.351-72.
- Schmid, H., 1994. Probabilistic Part-of-Speech tagging using decision trees. En *Proceedings of the International Conference on New Methods in Language Processing.*, 1994.
- Scott, M., 1996. *WordSmith Tools*. Oxford: Oxford University Press.
- Sierra, G., 2008. Diseño de corpus textuales para fines lingüísticos. En *IX Encuentro Internacional de Lingüística en el Noreste*. Hermosillo, México: Editorial Unison. pp.445-64.
- Singha, A., Buckley, C. y Mitra, M., 1996. Pivoted document length normalization. En *Research and development in information retrieval SIGIR 96*. Zurich, Suiza, 1996. ACM Press.
- Spärck-Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), pp.11-21.
- Suchanek, F.M., 2008. *Automated Construction and Growth of a Large Ontology*. PhD Tesis. Saarbrücken: Faculties of Natural Sciences and Technology.
- Toral, A., Ferrández, Ó., Noguera, E., Kozareva, Z., Montoyo, A. y Muñoz, R., 2006. Geographic IR Helped by Structured Geospatial Knowledge Resources. En *Working notes of CLEF - ECDL 2006, GeoCLEF Workshop*. Alicante, 2006.
- Torruella, J. y Llisterri, J., 1999. Diseño de corpus textuales y orales. *Filología e informática. Nuevas tecnologías en los estudios filológicos*, pp.45-47.

- van Rijsbergen, C.J., 1979. *Information Retrieval*. Segunda edición ed. Londres, Reino Unido: Butterworth-Heinemann.
- Vivaldi, J., 1995. Proyectos del IULA: El corpus técnico. En *Simposi Spanish Linguistics*. Manchester, Reino Unido, 1995. Instituto Cervantes y Manchester University.
- Vivaldi, J., 2001. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Barcelona: Universidad Politécnica de Cataluña: Ph.D Tesis.
- Vivaldi, J., Marquèz, L. y Rodríguez, H., 2001. Improving term extraction by system combination using boosting. En *12th European Conference on Machine Learning*., 2001.
- Vivaldi, J. y Rodríguez, H., 2010. Using Wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural*, 45, pp.251-54.
- Zesch, T. y Gurevych, I., 2007a. Analysis of the Wikipedia Category Graph for NLP Applications. En *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*., 2007a.
- Zesch, T., Gurevych, I. y Mühlhäuser, M., 2007b. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. En *Data Structures for Linguistic Resources and Applications*. Tübingen, 2007b.
- Zesch, T., Müller, C. y Gurevych, I., 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. En *6th Conference on Language Resources and Evaluation (LREC)*., 2008.

ANEXOS

Anexo A: Lista de palabras funcionales

a	alto	b	ciertísima	cualquiera
á	am	ba	ciertísimas	cuan
abajo	ambas	bah	ciertísimo	cuán
abril	ambos	bajo	ciertísimos	cuando
abrir	and	bastante	cierto	cuándo
acaso	antano	bastantes	ciertos	cuandoquiera
aceptar	añaño	bien	cinco	cuanta
actualmente	ante	breve	claro	cuánta
acuerdo	anteayer	buen	clic	cuantas
adelante	anterior	buena	com	cuántas
ademas	anteriores	buenas	comenzar	cuanto
además	anteriormente	bueno	como	cuánto
adios	antes	buenos	cómo	cuantos
adiós	aparecer	buscar	comoquiera	cuántos
adónde	aparte	bye	completamente	cuatro
afuera	apellido	c	comprar	cuya
afueras	apenas	cabe	con	cuyas
agosto	aproximadamente	cabo	concerniente	cuyo
agregar	aquel	cada	concernientes	cuyos
ah	aquél	calcular	concluir	d
aha	aquella	capaz	conmigo	da
ahi	aquella	capítulo	conocer	dado
ahí	aquellas	capítulos	conque	dah
ahora	aquellas	casi	consequentemente	dan
aja	aquello	cerca	considerar	dar
ajá	aquellos	cercano	consigo	de
al	aquéllos	cercanos	contigo	debajo
alcanzar	aquí	cercas	contra	deber
algo	aquí	cero	correspondiente	definir
alguien	arriba	cerrar	cosa	definitivamente
algun	arribaabajo	chas	cosas	del
algún	artículo	chau	crees	delante
alguna	asi	chaz	creo	demás
algunas	así	che	cual	demasiada
alguno	asimismo	chist	cuál	demasiadas
algunos	atrás	chito	cuales	demasiado
alla	aun	chitón	cuáles	demasiados
allá	aún	ciao	cualesquier	dentro
alli	aunque	cierta	cualesquiera	deprisa
allí	ay	ciertamente	cualesquieras	desarrollar
alrededor	ayer	ciertas	cualquier	desde

designar	en	fin	inmediatos	medio
despacio	encima	final	interpretar	mejor
despues	encontrar	g	ir	mencionar
después	enero	ganar	&	menos
determinar	enfrente	general	j	menudo
detras	enseguida	generalmente	ja	mi
detrás	entonces	gracias	jaja	mí
día	entre	gran	jam	mia
día	entretanto	grandes	jamás	mía
días	esa	guau	jamás	mias
días	ésa	h	jue	mías
dic	esas	ha	jueves	mie
diciembre	ésas	haber	julio	mientras
diez	ese	hacia	junio	miercoles
diferente	ése	hallar	junto	miércoles
diferentes	eso	harta	juntos	mio
diversas	esos	hartas	k	mío
diversos	ésos	harto	l	mios
dom	esta	hartos	la	míos
domingo	ésta	hasta	las	mis
don	está	hay	le	misma
donde	están	he	leer	mismas
dónde	estas	helo	lejos	mismo
dondequiera	ésta	hi	les	mismos
dondequiera	éstas	hola	limitar	momento
doña	este	horas	llamar	mucha
dos	éste	hoy	llegar	muchas
durante	esto	htm	lo	mucho
e	ésto	html	los	muchos
é	estos	http	luego	muy
efectivamente	éstos	https	lugar	n
efectivamente	et	i	lun	nada
eh	etc	í	lunes	nadie
ejemplo	etc.	idem	m	necesitar
el	etcétera	identificar	ma	ni
él	ex	igual	mal	ningun
ella	exactamente	igualmente	martes	ningún
ellas	excepto	in	marzo	ninguna
ello	expresar	incluso	mas	ningunas
ellos	extra	indicar	más	ninguno
email	ey	inmediata	mayo	ningunos
e-mail	f	inmediatamente	mayor	no
embargo	feb	inmediatas	me	nos
emplear	febrero	inmediato	mediante	nosotras

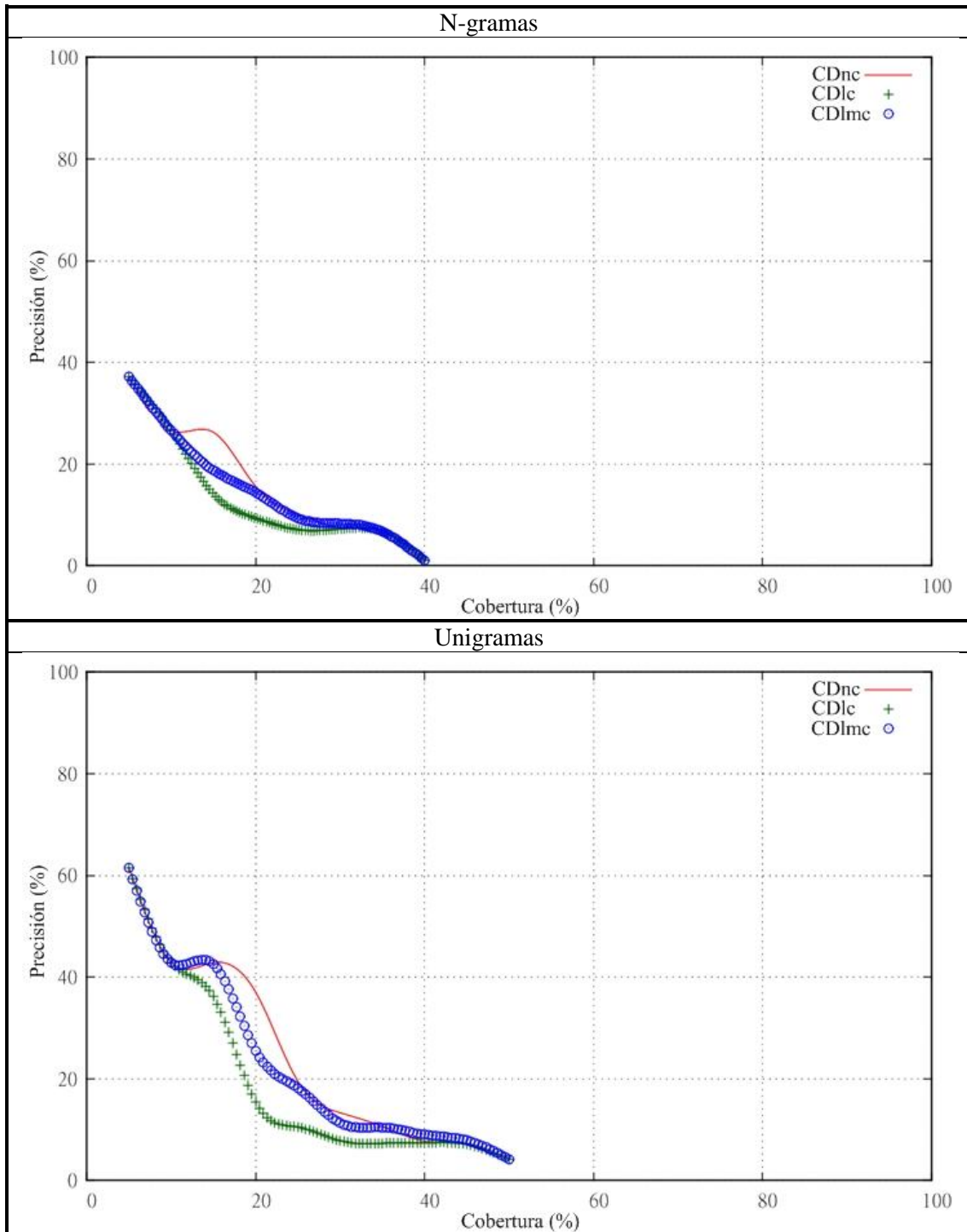
nosotros	parece	q	sabido	soyos
nov	parte	queu	se	su
noviembre	partir	quéu	sea	súbitamente
nuestra	pasada	que	seguida	subsiguiente
nuestras	pasado	qué	seguidas	suceder
nuestro	peor	queda	seguido	suficiente
nuestros	pero	quedar	seguidos	suficientes
nueva	pesar	querer	segun	supuesto
nuevas	php	quien	según	sus
nueve	plantear	quién	segunda	suya
nuevo	pm	quienes	segundo	suyas
nuevos	poca	quiénes	seis	suyo
núm	pocas	quienquiera	señor	suyos
nunca	poco	quiza	señora	t
ñ	pocos	quizá	señorita	tac
o	poder	quizas	señorito	tal
ó	pom	quizás	sep	tales
observar	por	r	sept	tambien
obstante	porque	rapida	septiembre	también
obtener	pos	rápida	si	tampoco
obviamente	posible	rapidamente	sí	tan
obvio	posteriormente	rápidamente	sido	tanta
ocho	prácticamente	rapidas	siempre	tantas
oct	presentar	rápidas	siete	tanto
octubre	primer	rapido	siguiente	tantos
of	primera	rápido	siguientemente	tarde
oh	primeras	rapidos	siguientes	te
ohh	primero	rápidos	significar	temprano
ok	primeros	raras	simplemente	tercera
ole	principalmente	realmente	simultáneamente	tercero
olé	pronta	realizar	sin	the
os	pronto	recapitulación	sino	ti
otra	propia	recordar	siquiera	tí
otras	propias	repente	so	tic
otro	propio	representar	sobre	tin
otros	propios	resolver	sobrepasar	toc
p	proxima	respecto	sola	toda
pa	próxima	resuelve	solamente	todas
pág	proximas	revisar	solas	todavía
pág.	próximas	s	solo	todavía
página	proximo	sab	sólo	todo
pais	próximo	sáb	solos	todos
país	próximos	sabado	son	tomar
para	pues	sábado	soy	total

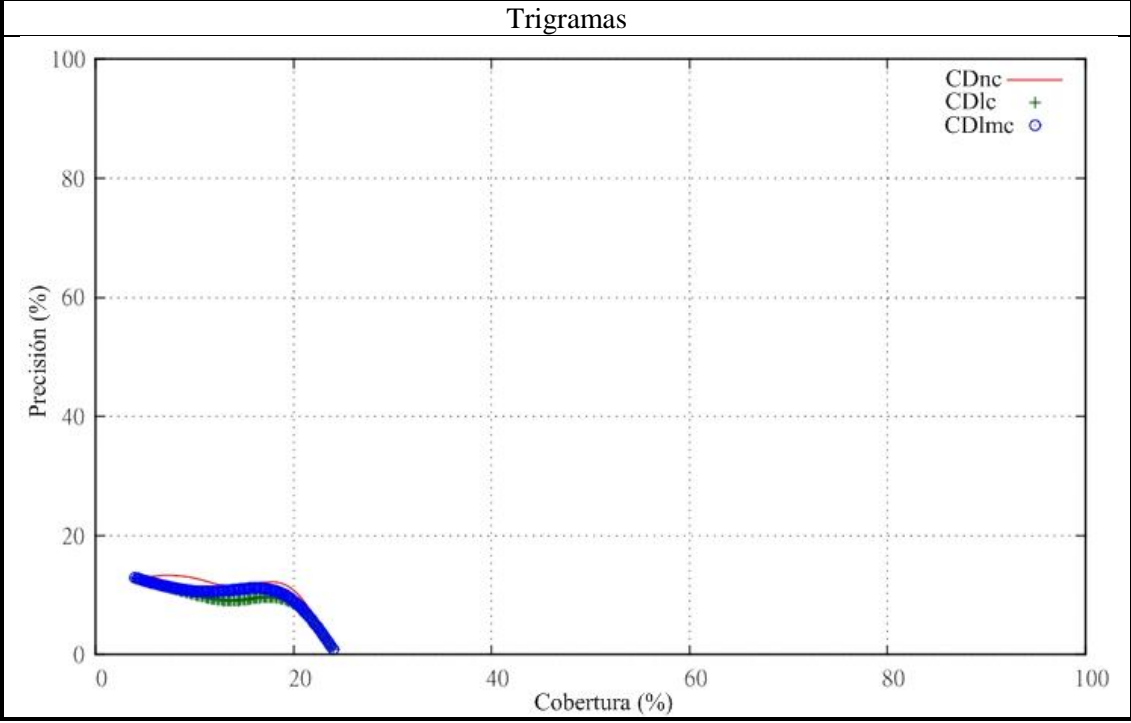
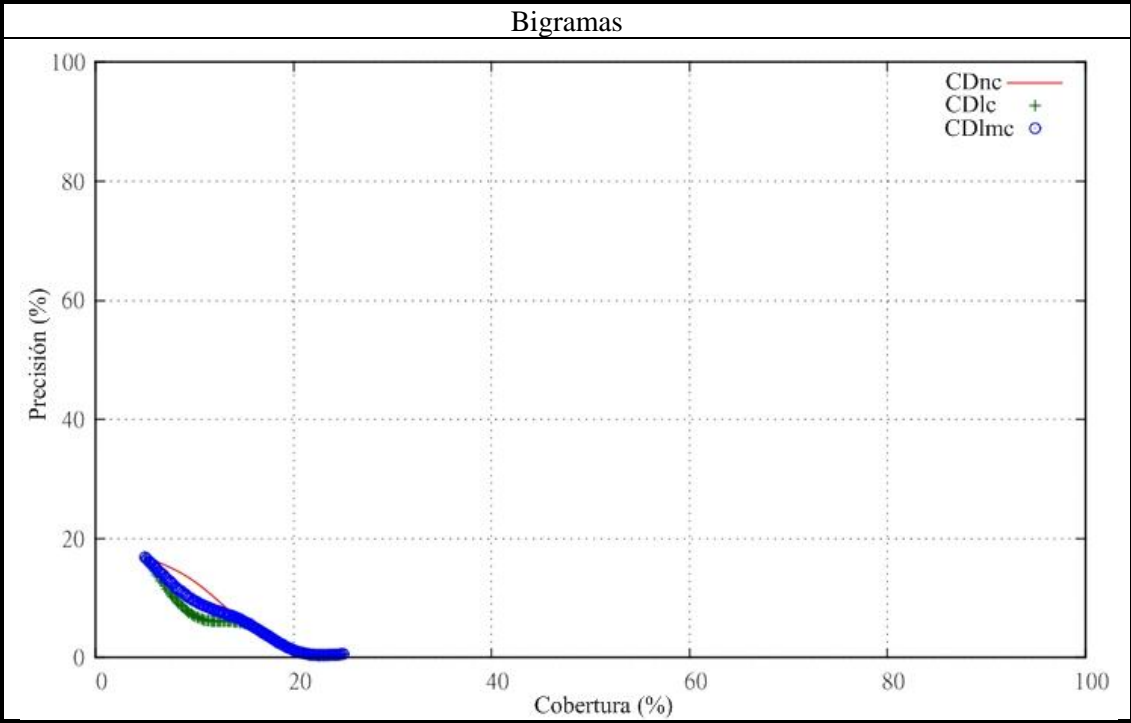
tras	usar	digan	estaréis	fuieste
trás	usted	hacer	estaremos	fuiesteis
traves	ustedes	hizo	estaría	sea
través	utilizar	hice	estaríais	seáis
trazar	uy	hagamos	estaríamos	seamos
tres	v	hicieron	estarían	sean
tu	va	harían	estarías	seas
tú	varias	hago	estás	sentid
tus	varios	haga	esté	sentida
tuya	veces	hagan	estéis	sentidas
tuyas	ver	saber	estemos	sentido
tuyo	verdaderamente	sé	estén	sentidos
tuyos	vez	sepa	estés	ser
u	vie	sepamos	estoy	será
ú	viernes	sabría	estuve	serán
uf	vos	supe	estuviera	serás
uff	vosotras	supieron	estuvierais	seré
uh	vosotros	era	estuviéramos	seréis
ultima	vuestra	erais	estuvieran	seremos
última	vuestras	éramos	estuvieras	sería
ultimamente	vuestro	eran	estuvieron	seríais
últimamente	vuestros	eras	estuviese	seríamos
ultimas	w	eres	estuvieseis	serían
últimas	web	es	estuviésemos	serías
ultimo	www	está	estuviesen	sois
último	x	estaba	estuvieses	somos
ultimos	y	estabais	estuvimos	son
últimos	ya	estábamos	estuviste	soy
um	yes	estaban	estuvisteis	ha
umm	yo	estabas	estuvo	habéis
un	y/o	estad	fue	había
una	z	estada	fuera	habíais
unas	zaz	estadas	fuerais	habíamos
unica	zap	estado	fuéramos	habían
única	decir	estados	fueran	habías
unicas	decía	estáis	fueras	habida
únicas	decían	estamos	fueron	habidas
unico	dijo	están	fuese	habido
único	dije	estando	fueseis	habidos
unicos	digamos	estar	fuésemos	habiendo
únicos	dijeron	estará	fuesen	habrá
uno	dirían	estarán	fueses	habrán
unos	digo	estarás	fui	habrás
url	diga	estaré	fuimos	habré

habréis	hubiera	tendrás	tengas	tuvierais
habremos	hubierais	tendré	tengo	tuviéramos
habría	hubiéramos	tendréis	tenía	tuvieran
habríaís	hubieran	tendremos	teníaís	tuvieras
habríamos	hubieras	tendría	teníamos	tuvieron
habrían	hubieron	tendríaís	tenían	tuviese
habríaís	hubiese	tendríamos	tenías	tuvieseís
han	hubieseís	tendrían	tenida	tuviésemos
has	hubiésemos	tendríaís	tenidas	tuviesen
haya	hubiesen	tened	tenido	tuvieses
hayáís	hubieses	tenéis	tenidos	tuvimos
hayamos	hubimos	tenemos	teniendo	tuviste
hayan	hubiste	tener	tiene	tuvisteís
hayas	hubisteís	tenga	tienen	tuvo
he	hubo	tengáís	tienes	
hemos	tendrá	tengamos	tuve	
hube	tendrán	tengan	tuviera	

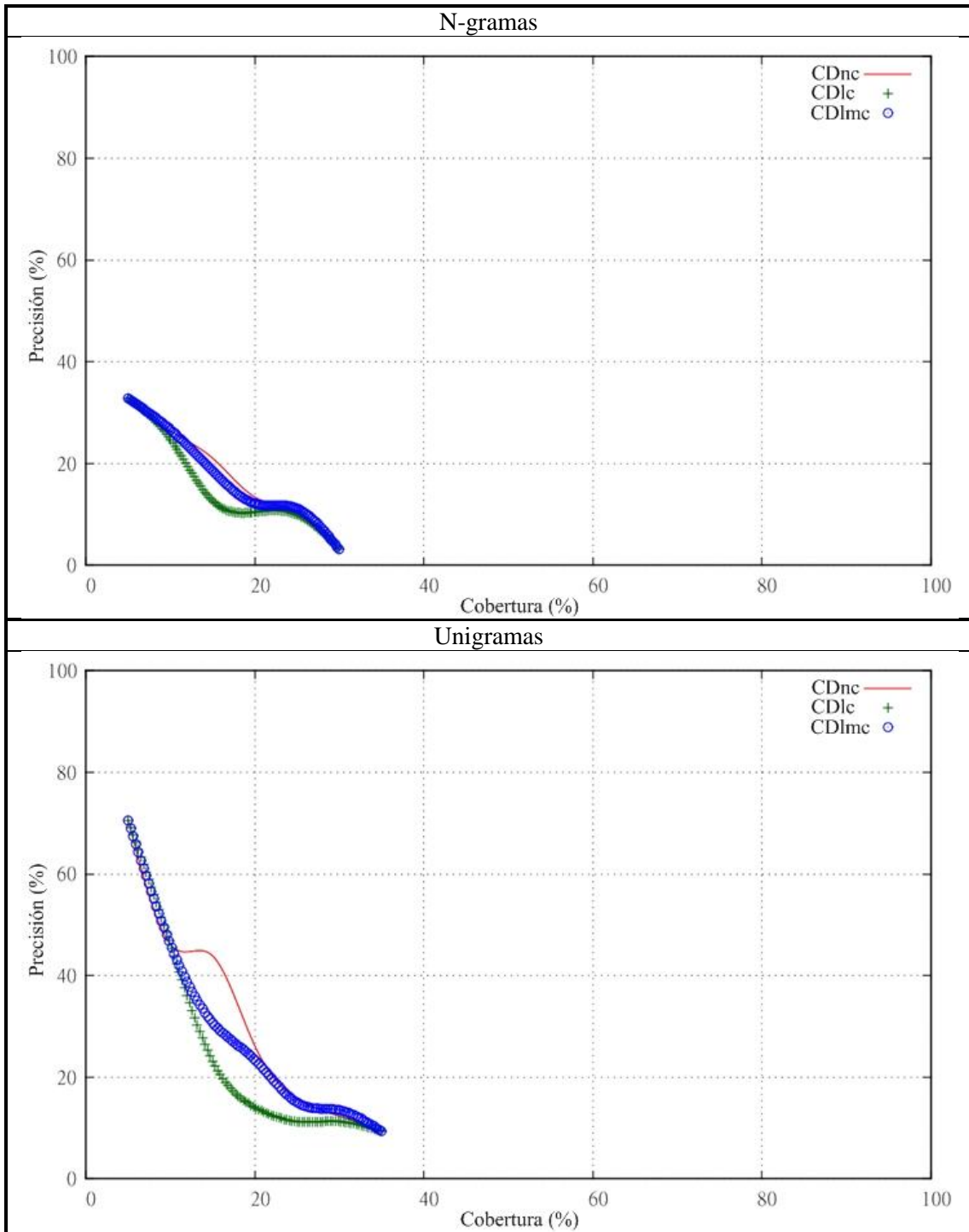
Anexo B: Gráficas de precisión contra cobertura de matemáticas de bachillerato

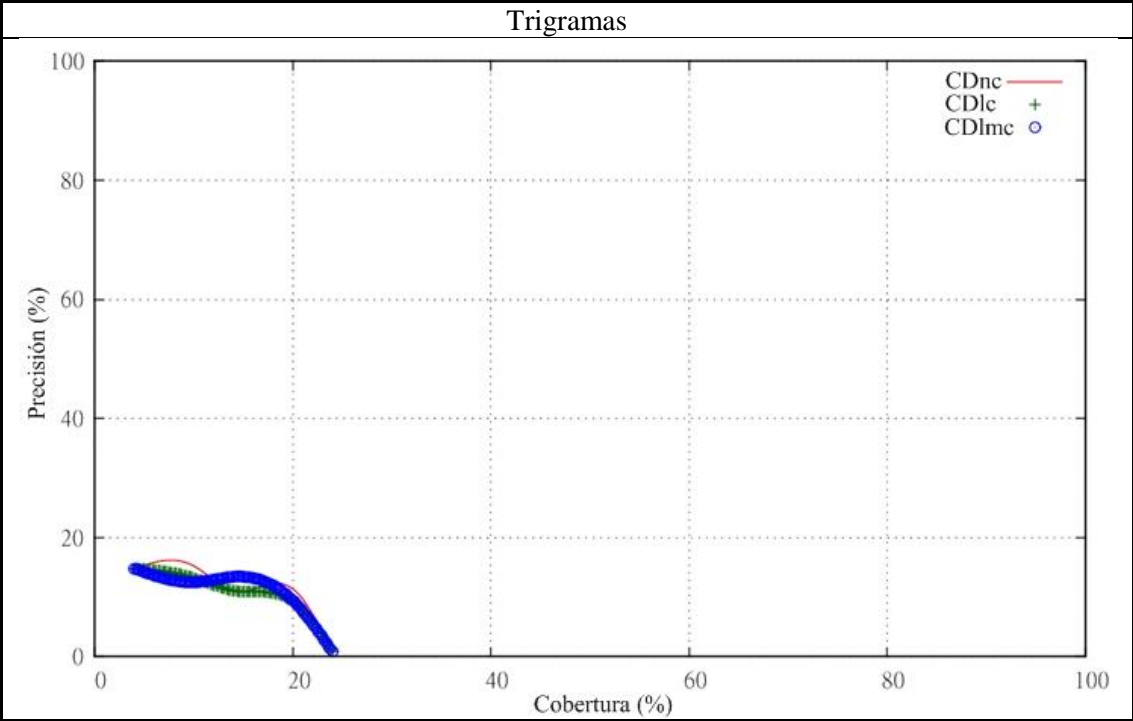
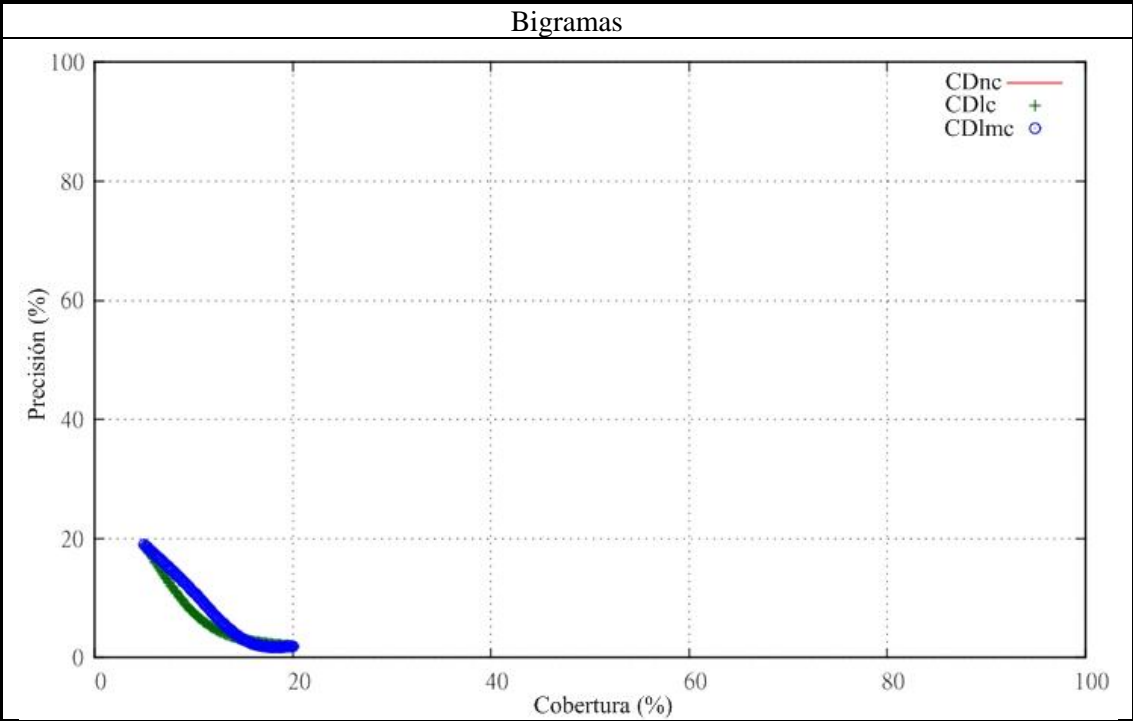
Para pesos de TF-IDF > 0



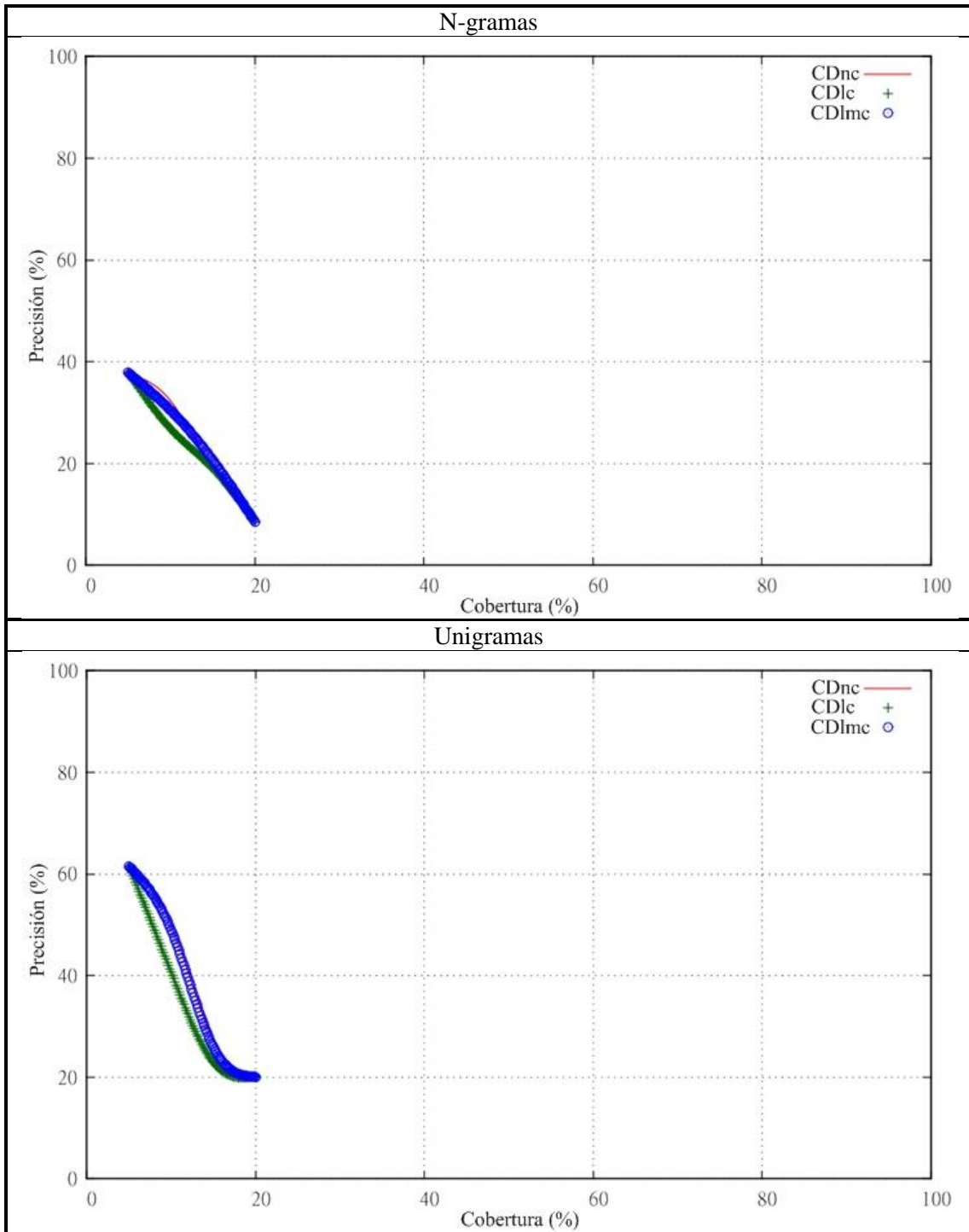


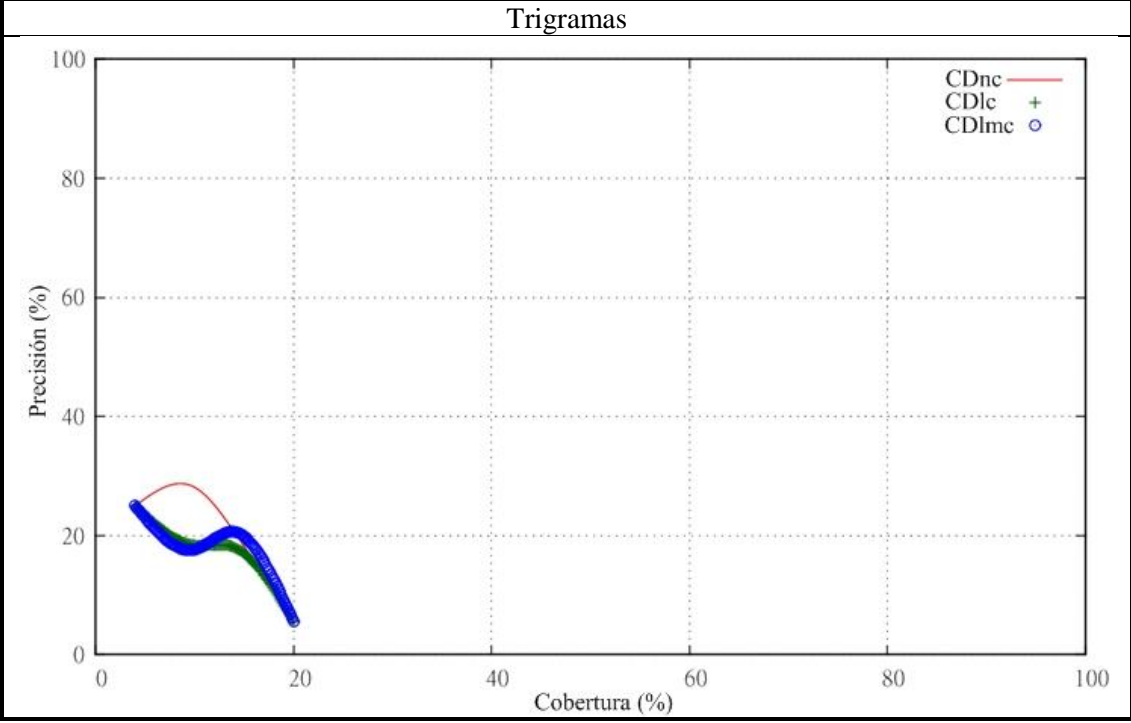
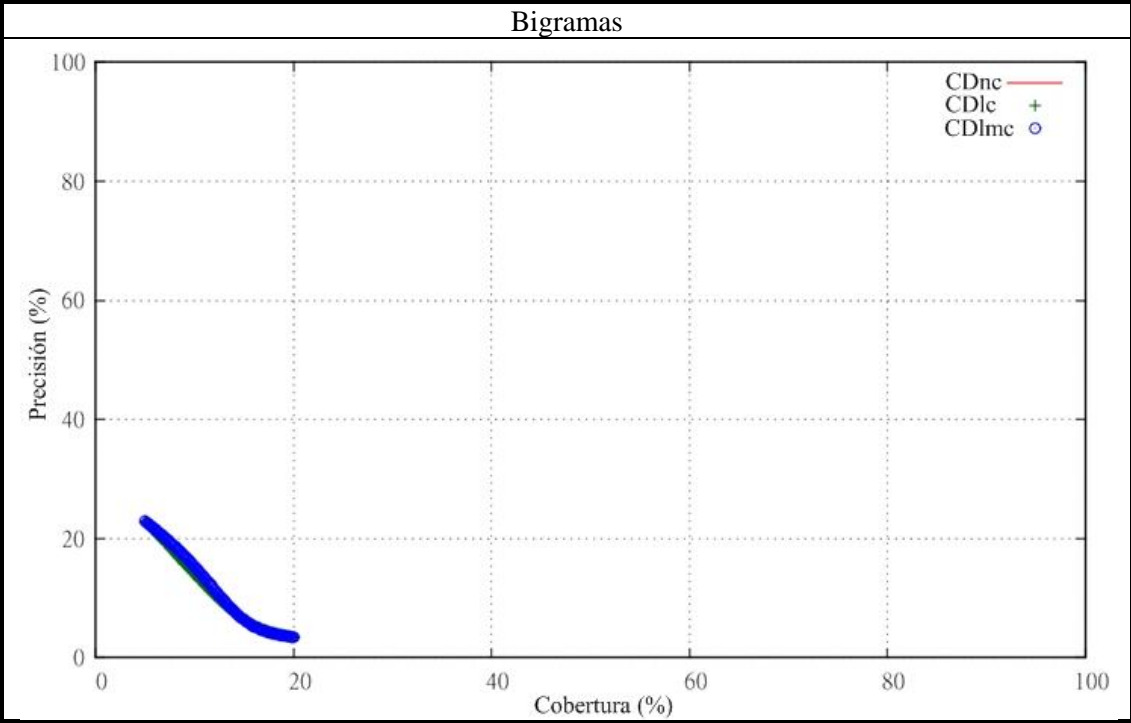
Para pesos de TF-IDF > 0.01





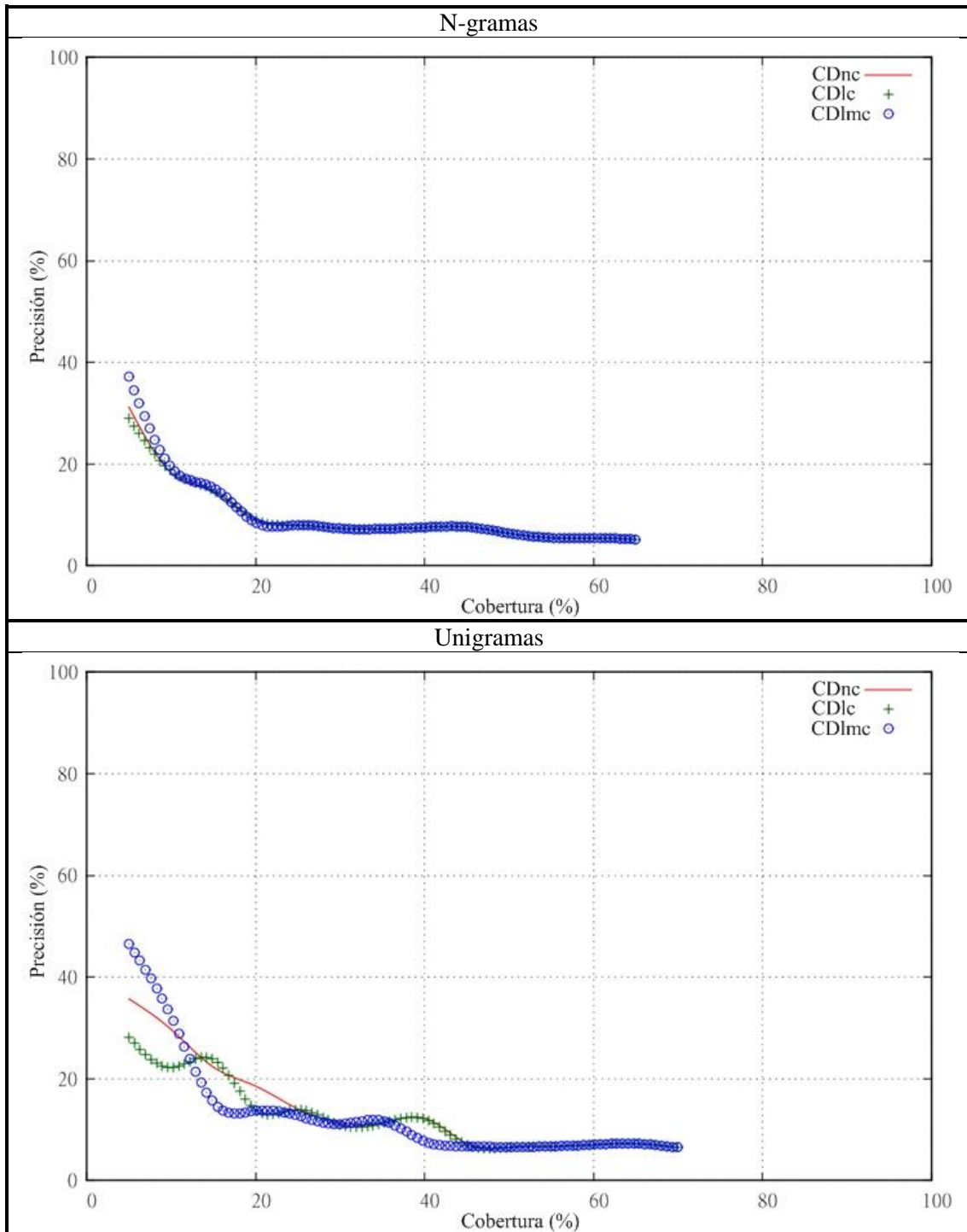
Para pesos de TF-IDF > 0.03

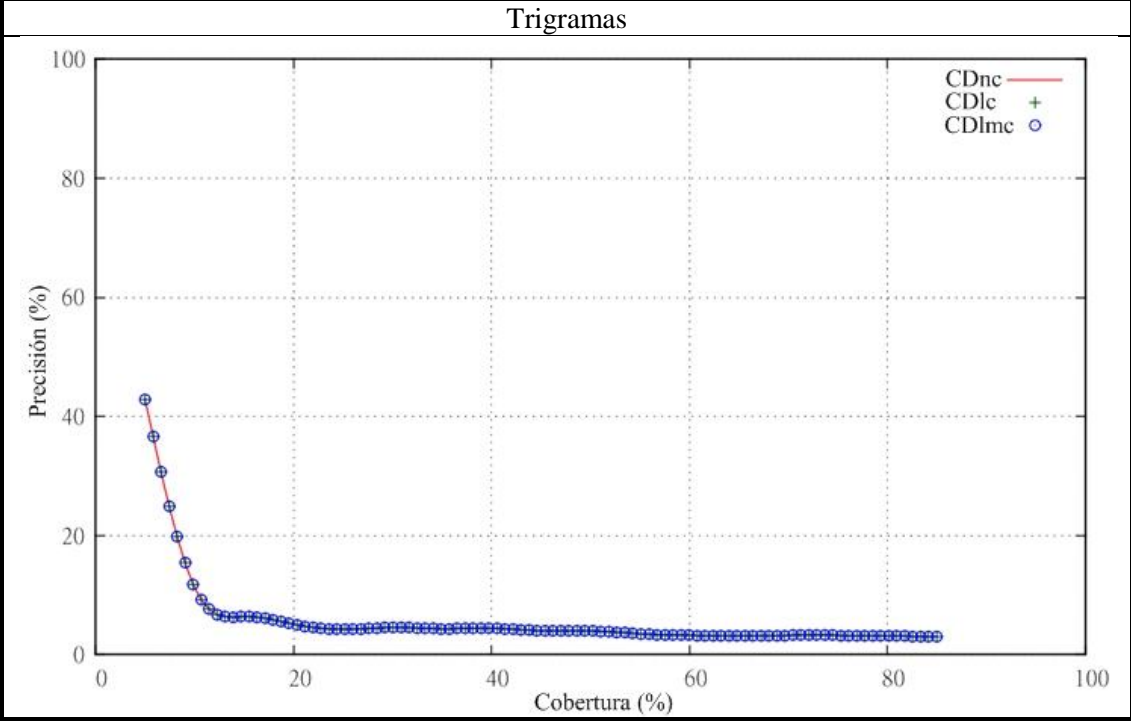
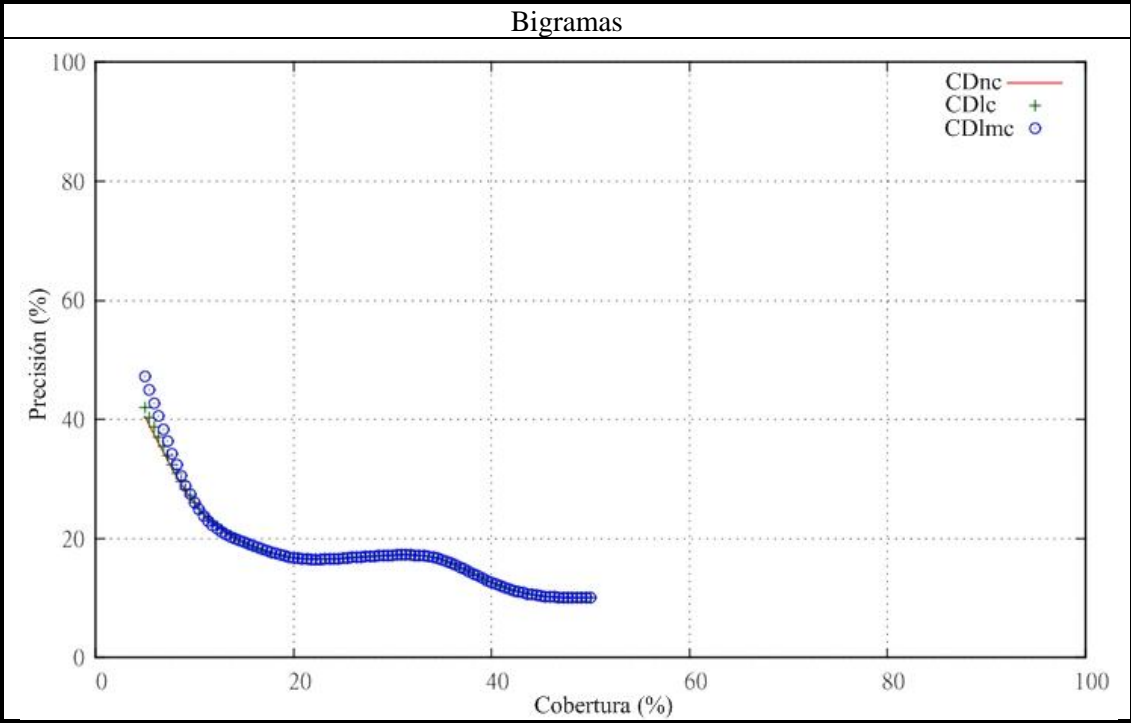




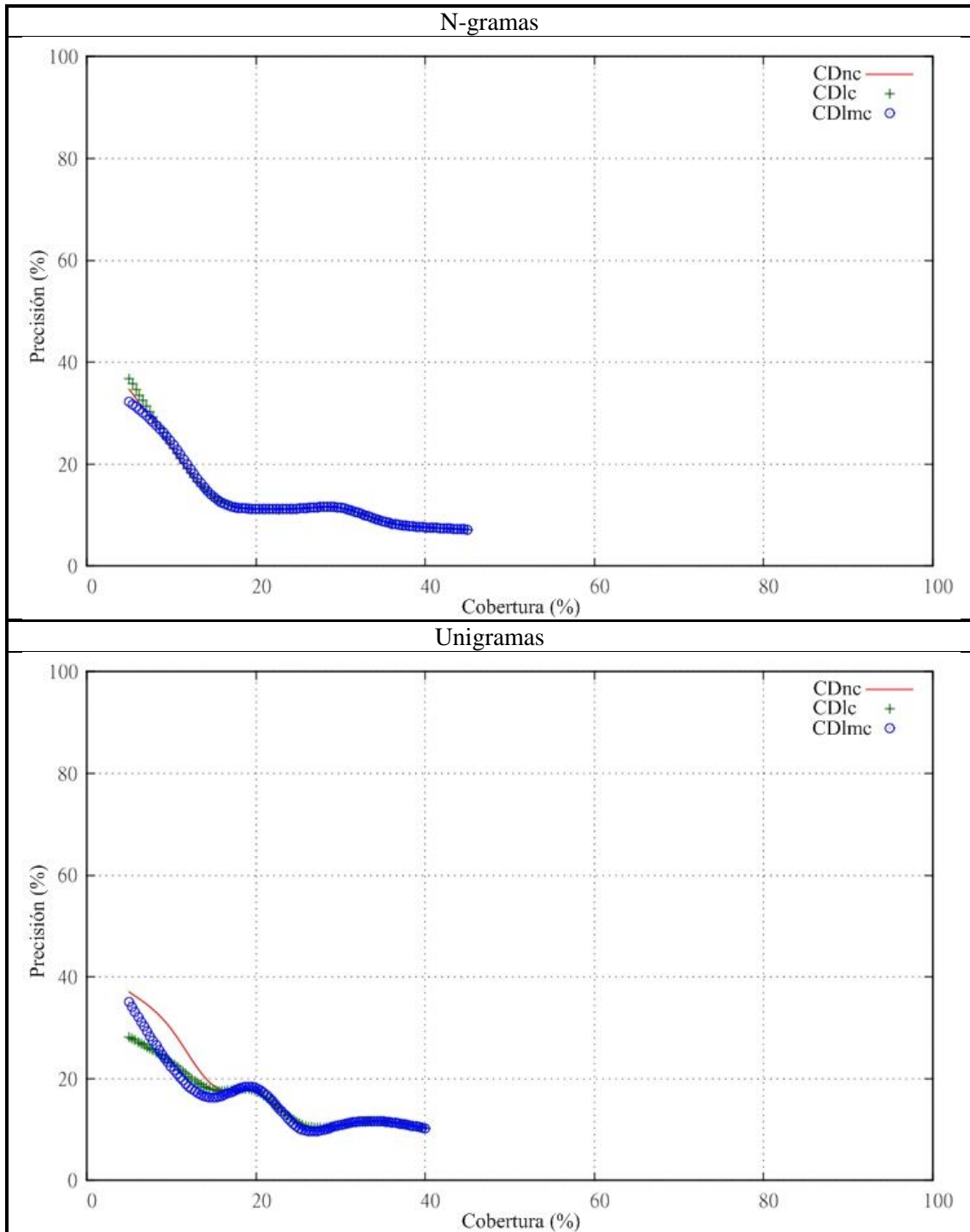
Anexo C: Gráficas de precisión contra cobertura de ecología de bachillerato

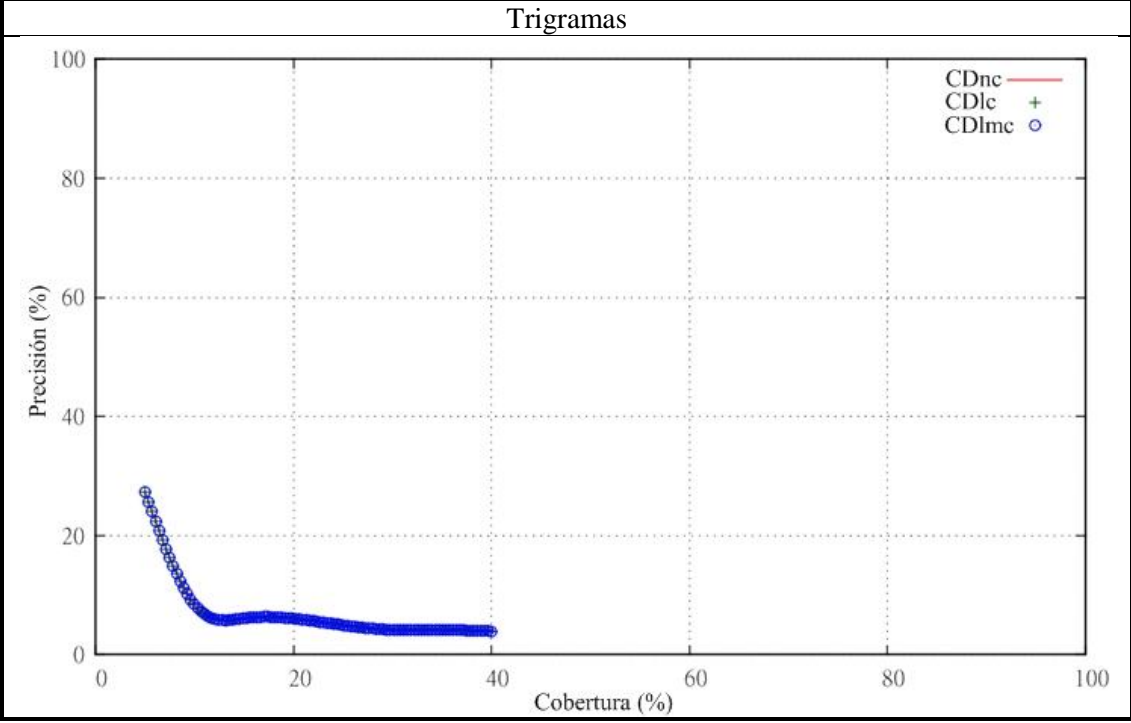
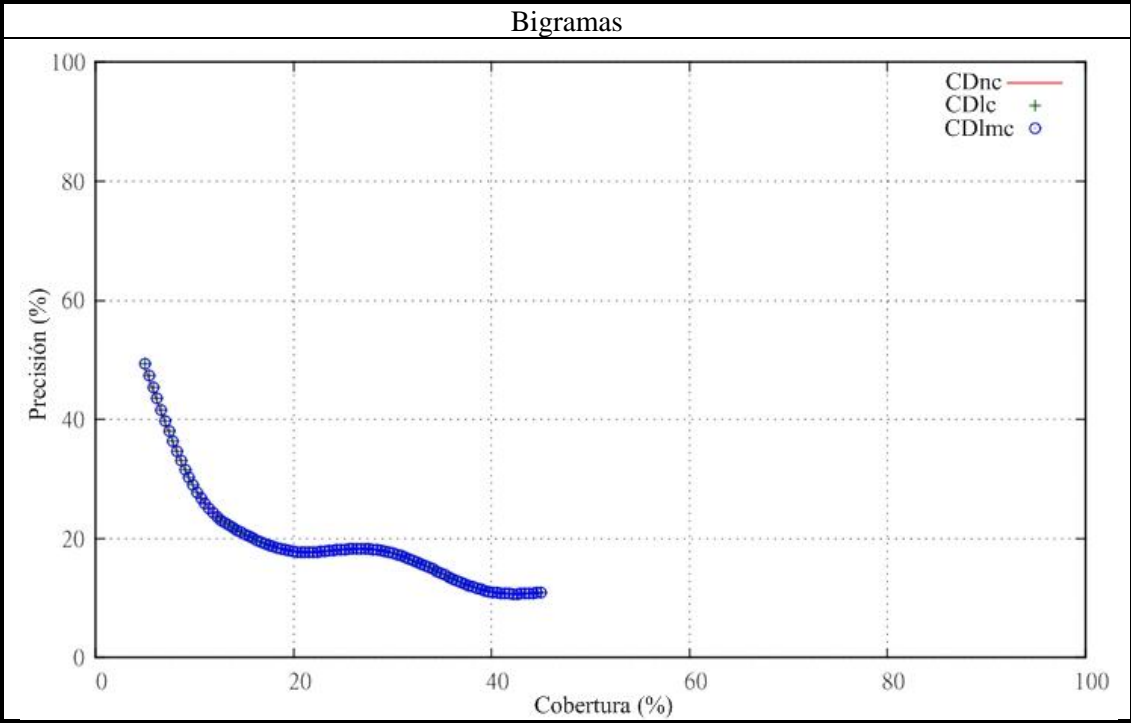
Para pesos de TF-IDF > 0



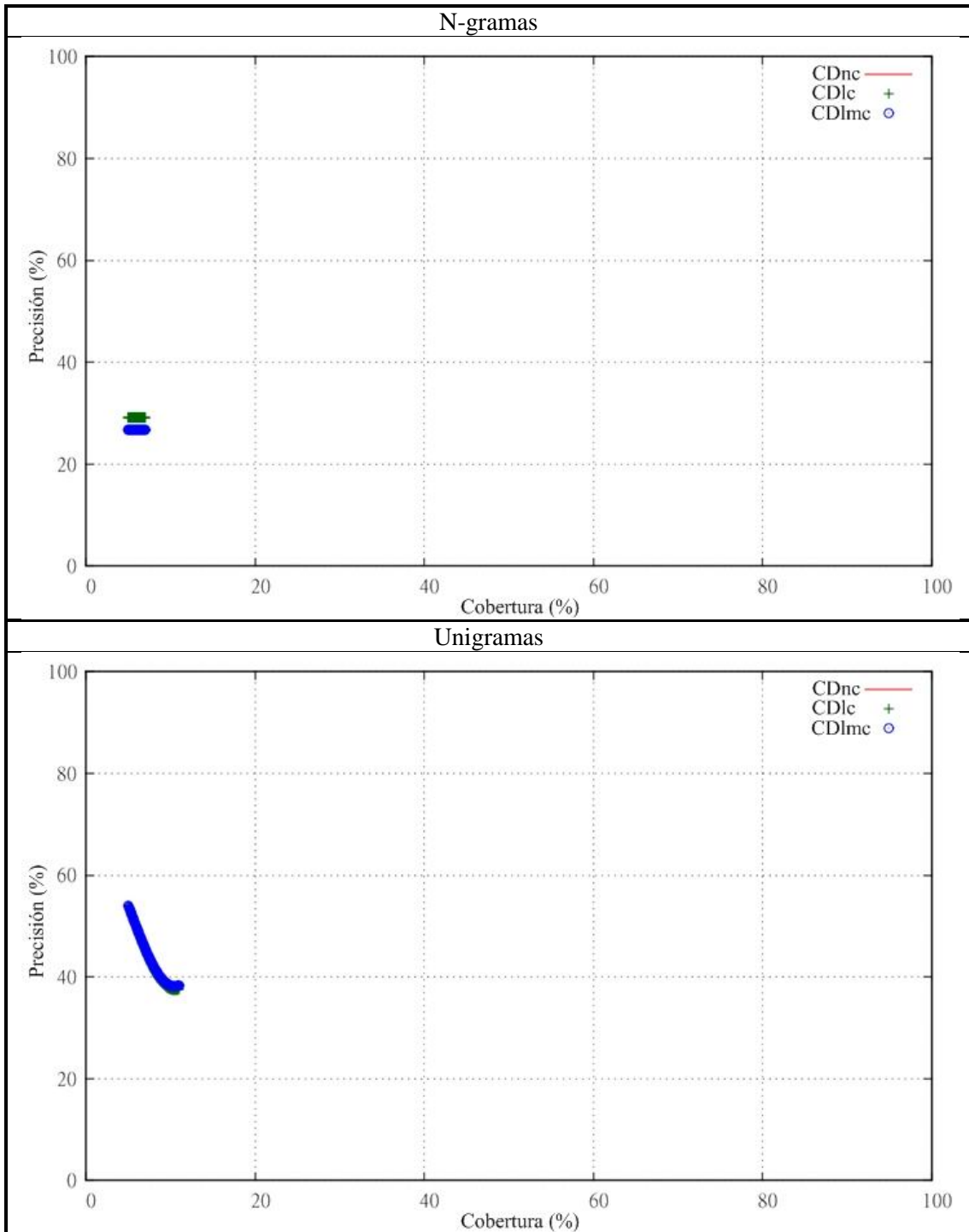


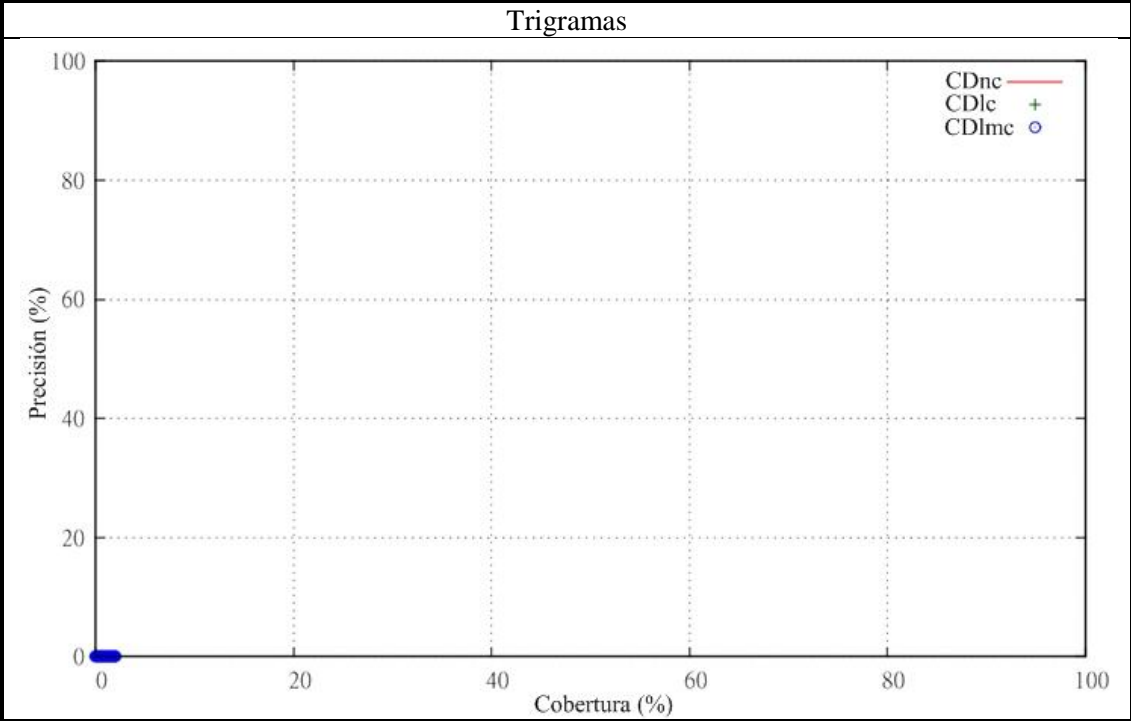
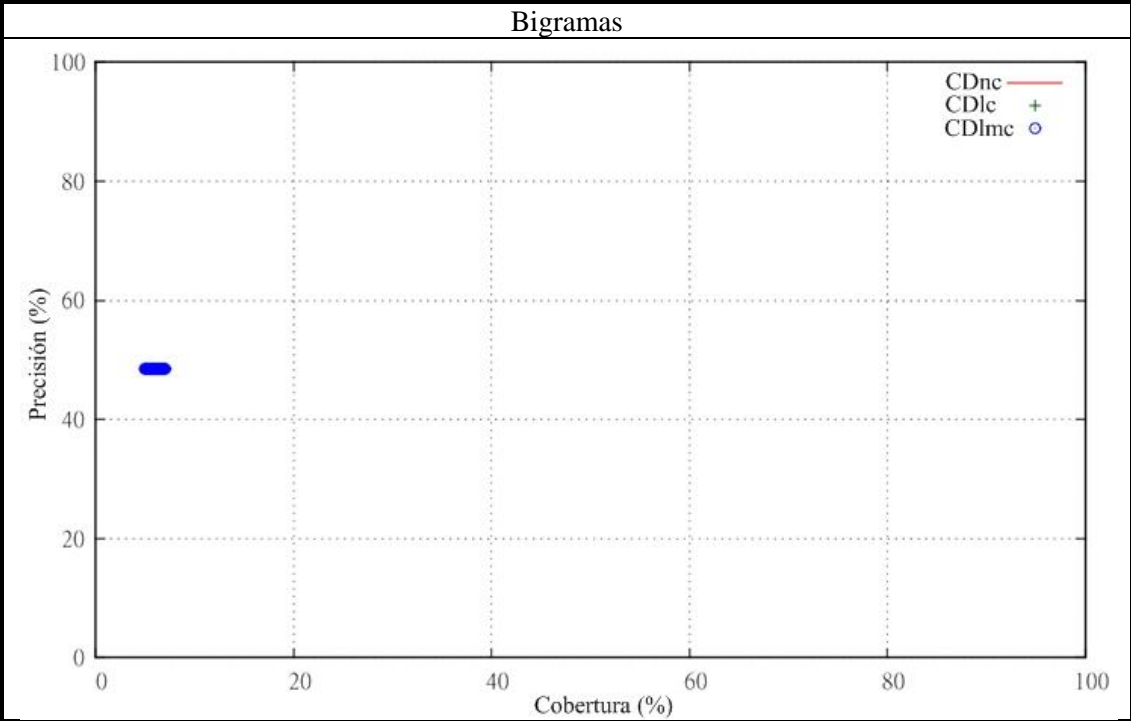
Para pesos de TF-IDF > 0.01





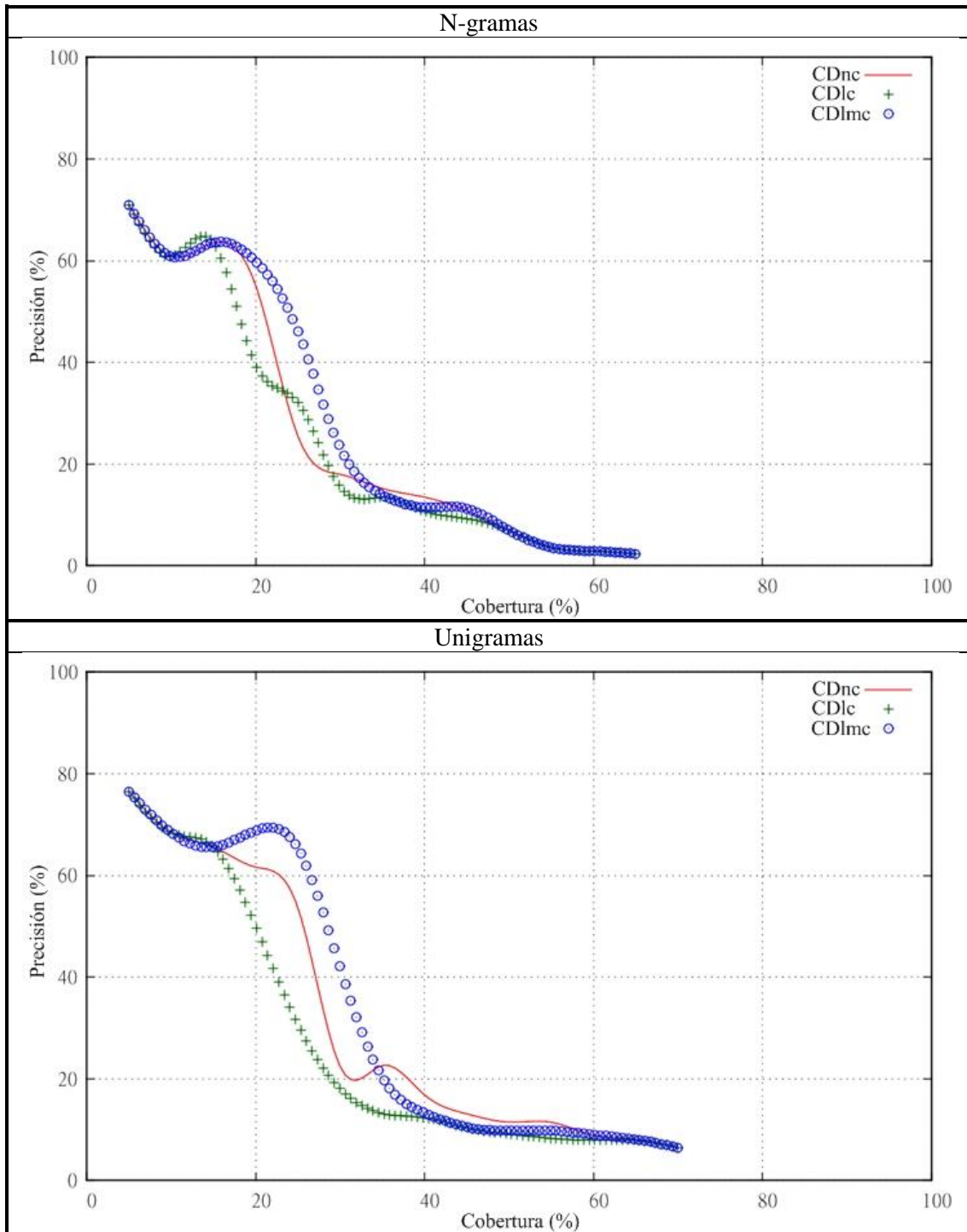
Para pesos de TF-IDF > 0.03

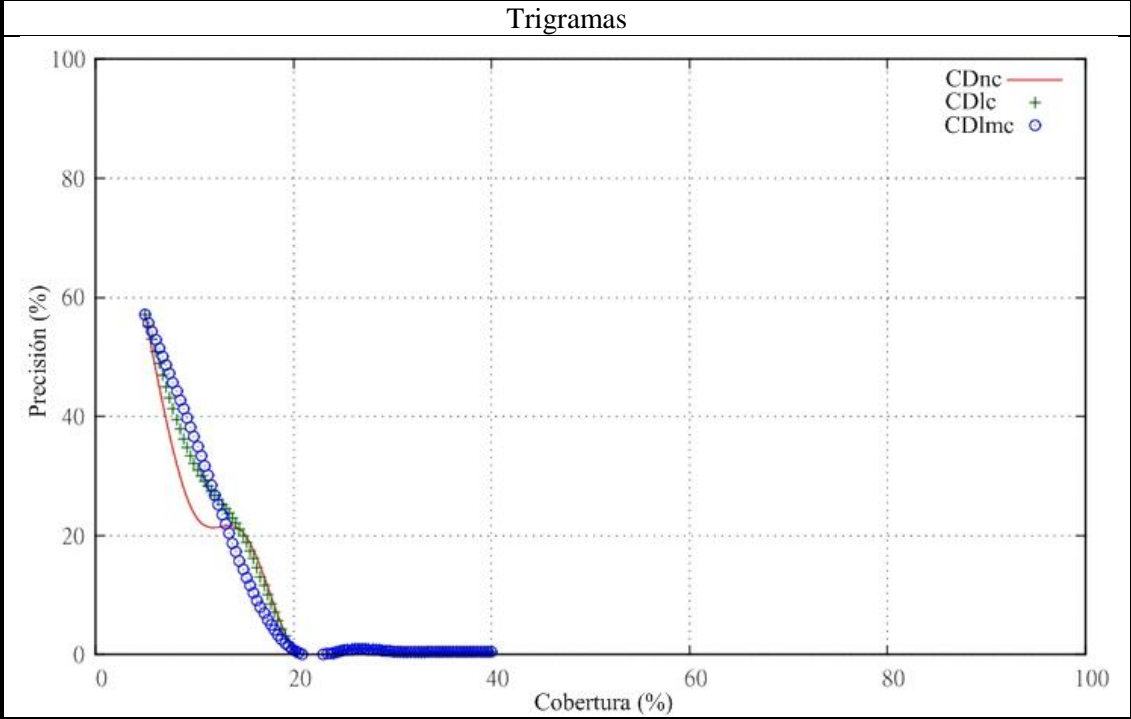
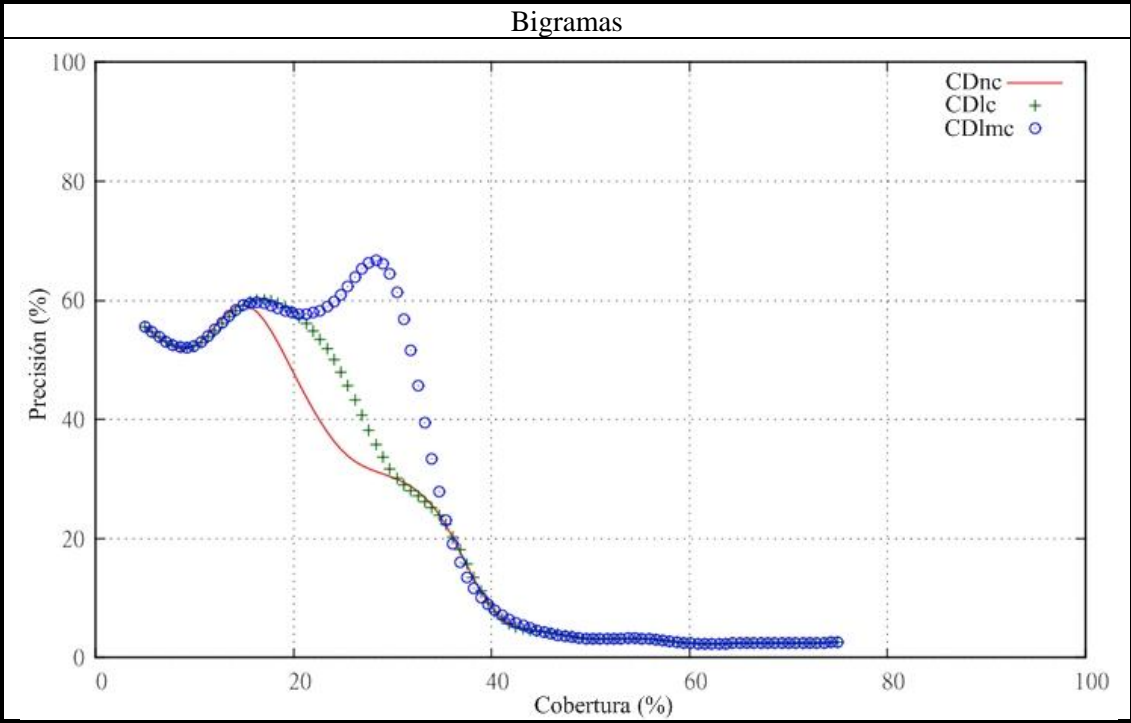




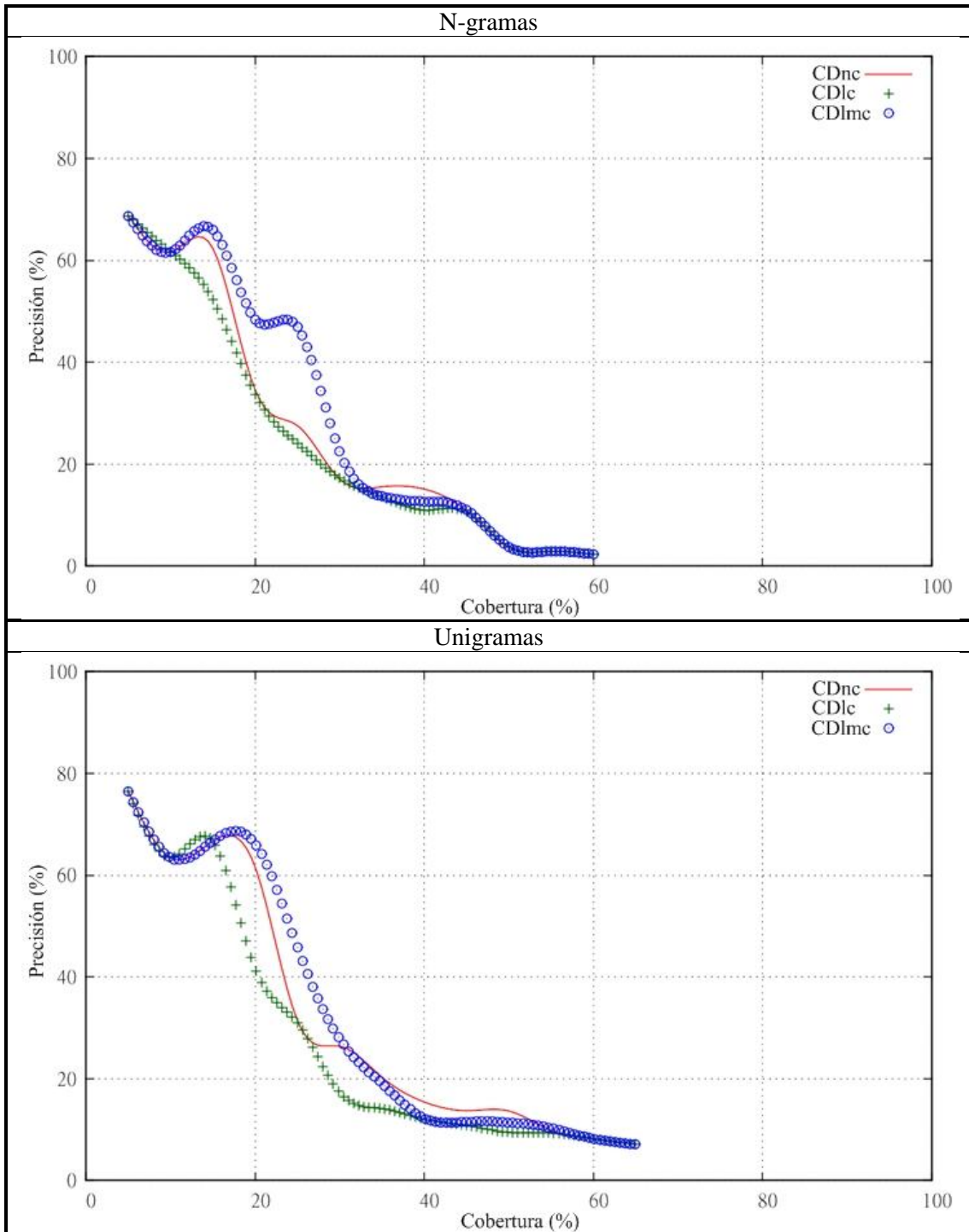
Anexo D: Gráficas de precisión contra cobertura de matemáticas de primaria

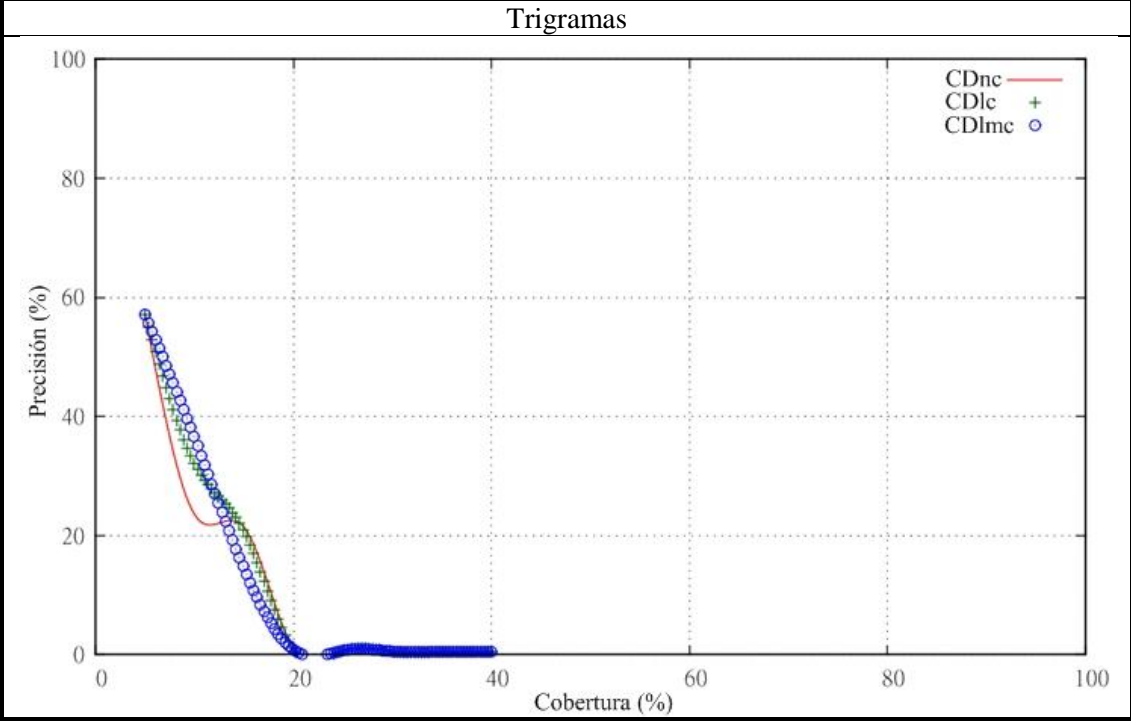
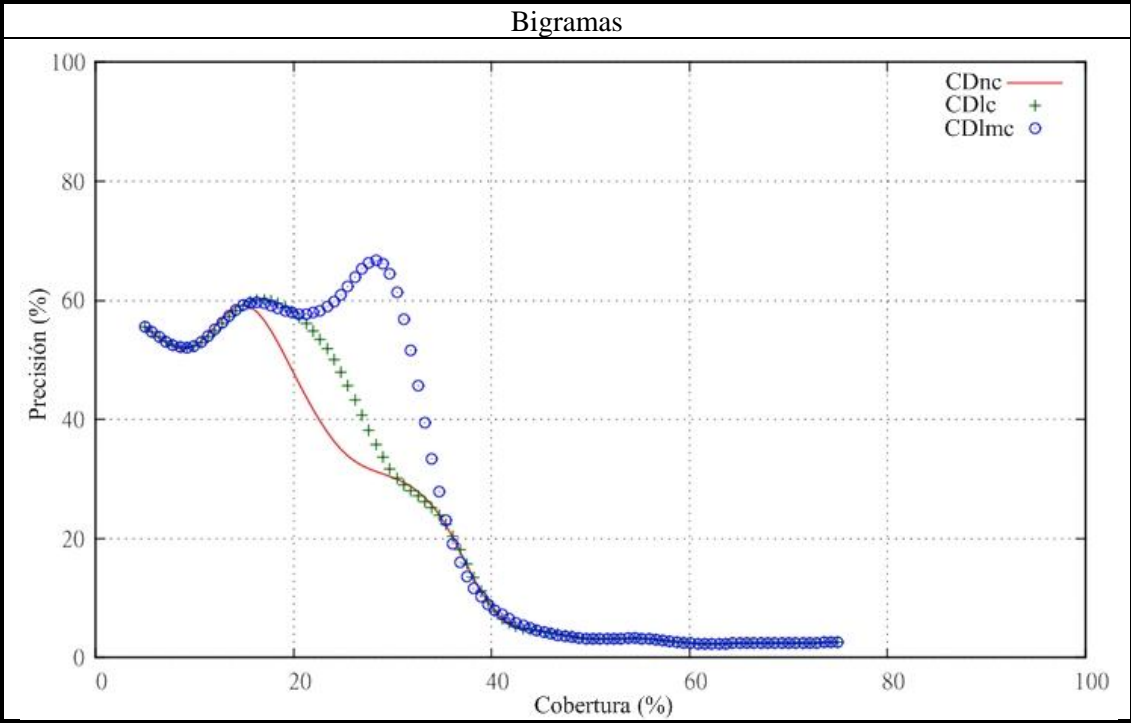
Para pesos de TF-IDF > 0



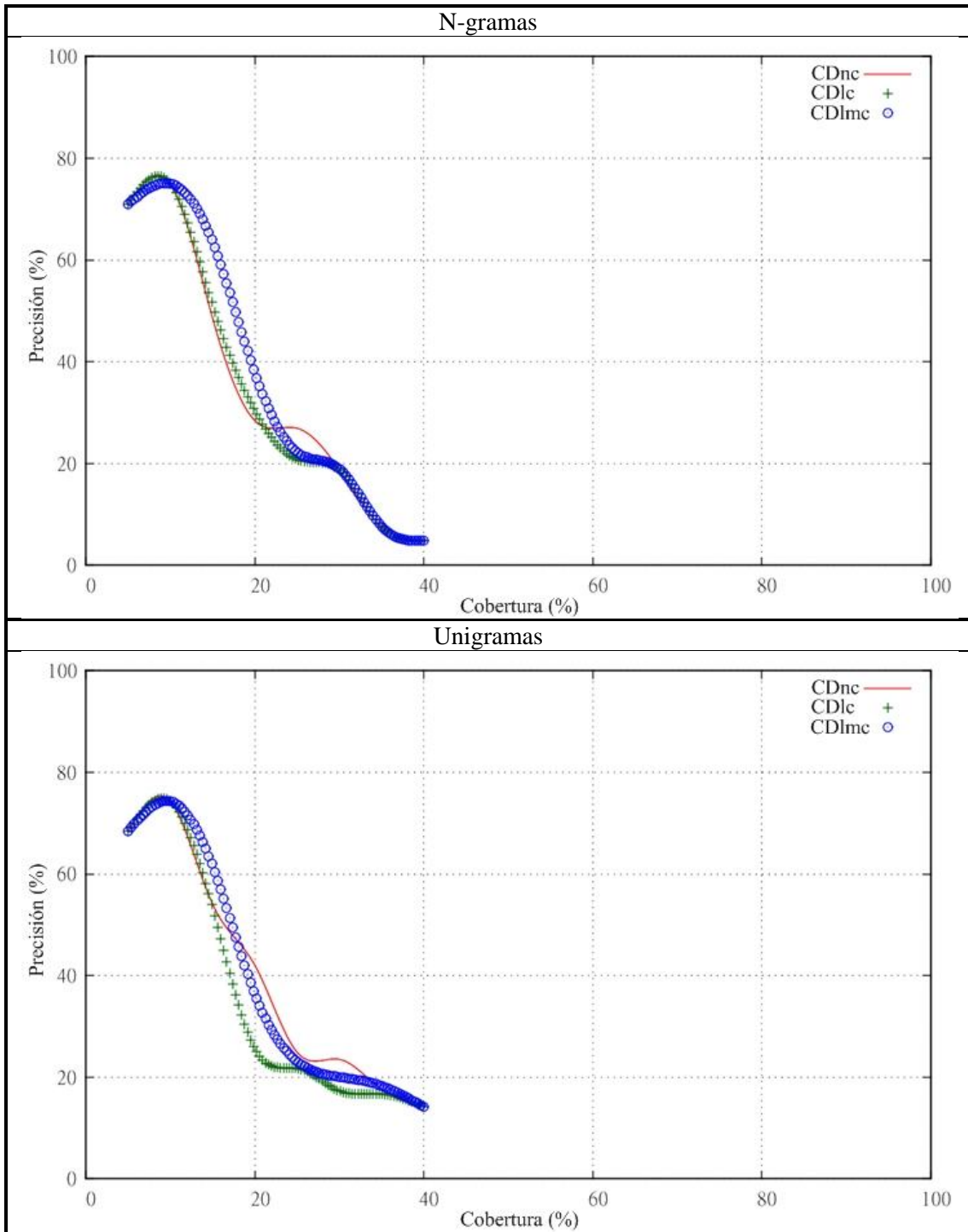


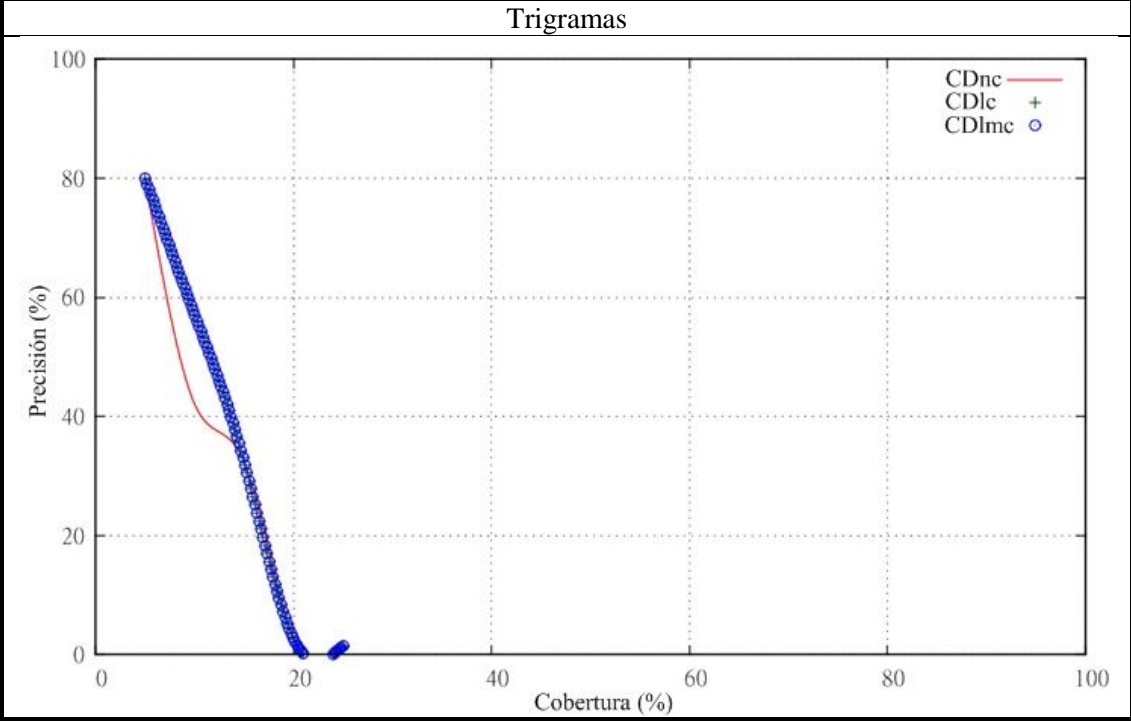
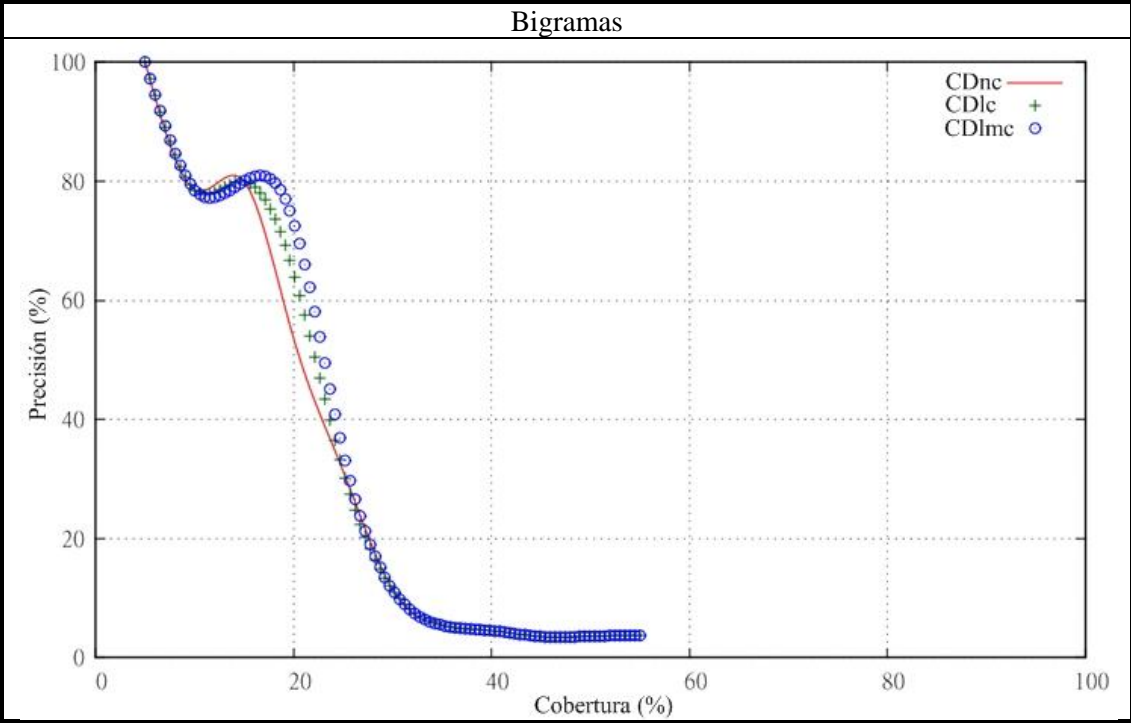
Para pesos de TF-IDF > 0.01





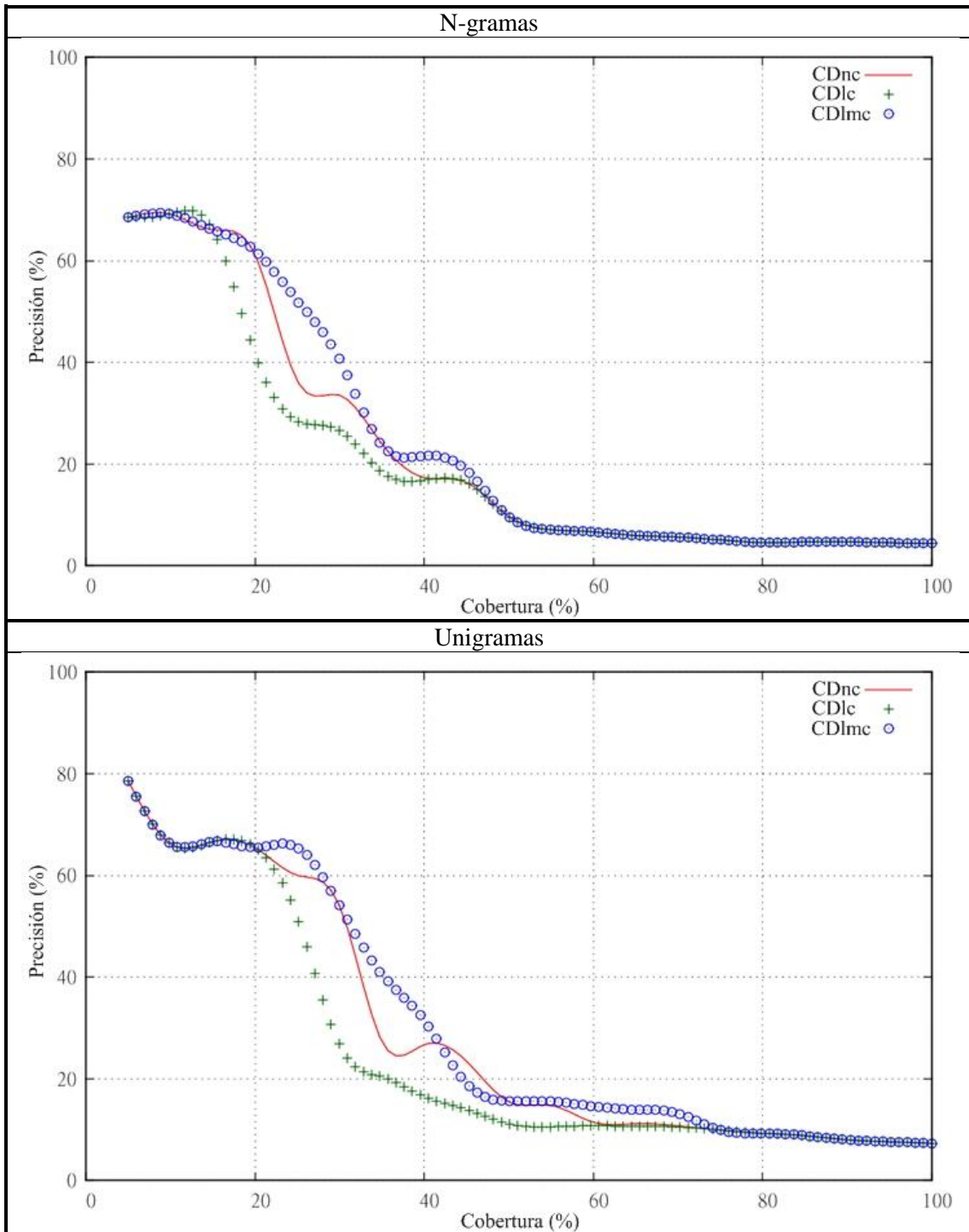
Para pesos de TF-IDF > 0.03

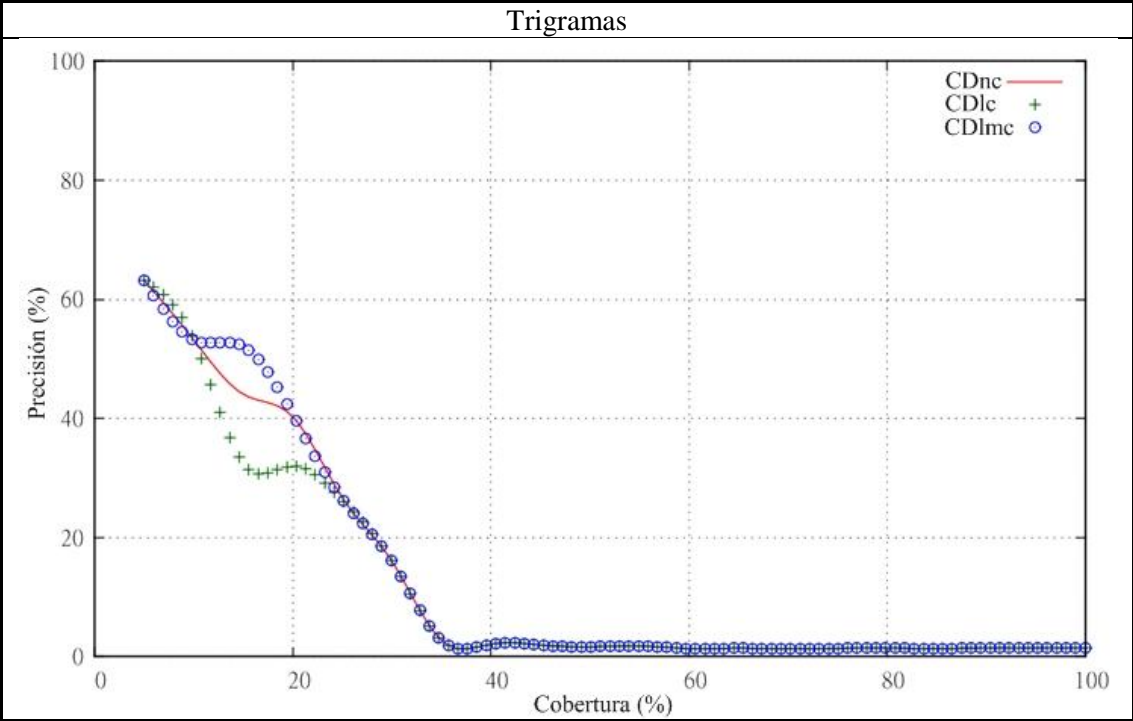
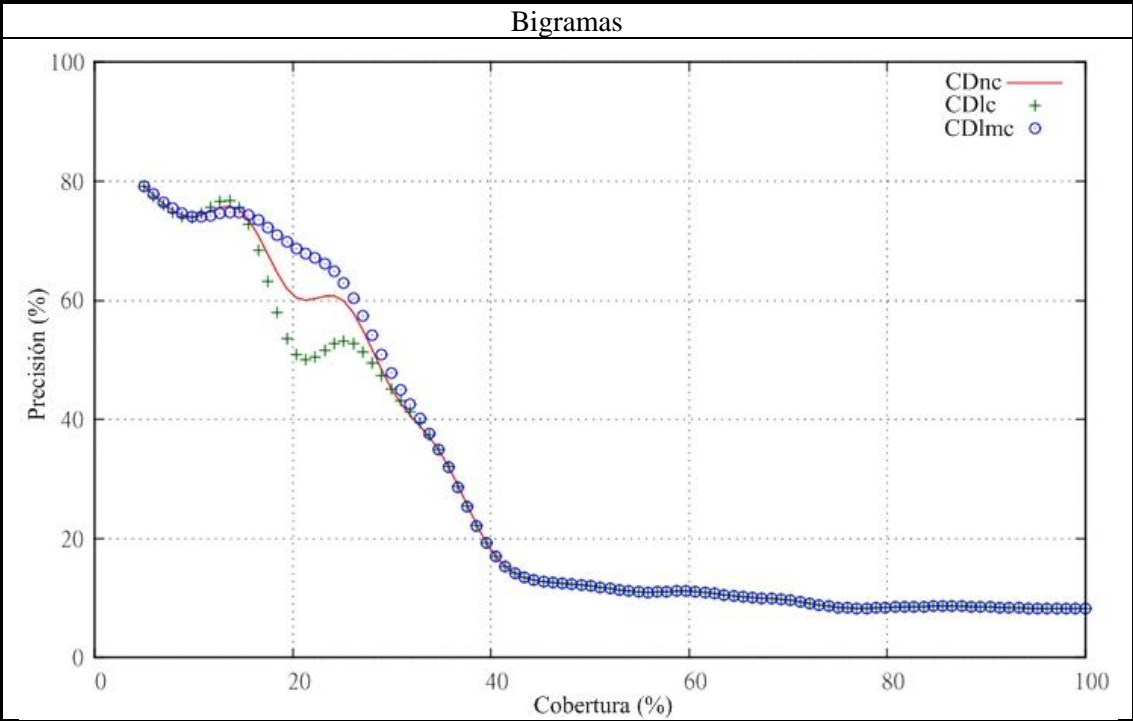




Anexo E: Gráficas de precisión contra cobertura de matemáticas de bachillerato evaluada con la segunda lista de términos

Para pesos de TF-IDF > 0





Anexo F: Lista de términos validados de matemáticas de bachillerato

La lista de términos validados presentada a continuación es la obtenida por el coeficiente de dominio con mejores resultados.

Para pesos de TF-IDF > 0 y CDnc

• adición	• conjunto finito	• distributividad	• fracción algebraico	• inverso
• afinidad	• conjunto infinito	• divisibilidad	• fracción continua	• multiplicativo
• algebra	• conjunto numérico	• divisible	• fracción continuo	• investigación de operación
• ancho	• conjunto ordenar	• división de radicales	• fracción impropio	• irreducible
• aproximación lineal	• conjunto referencial	• división euclidiano	• función circular	• isomorfismo
• aritmética	• construcción geométrico	• división sintético	• función corto	• lado
• asociatividad	• corolario	• diámetro	• función derivar	• ley de cosenos
• astroide	• corona circular	• dodecágono	• fórmula de euler	• ley de senos
• bidimensional	• cuadrado perfecto	• dígito	• generatriz	• ley de tricotomía
• binomio	• cuadrilátero	• dígito significativo	• geodésico	• longitud
• binomio conyugar	• cuantificador existencial	• ecuación canónico	• geometría	• líneas
• binomio de newton	• cuantificador universal	• elemento inverso	• geometría no euclidiano	• matemática
• bisectriz	• cubo	• elemento neutro	• geometría algebraico	• matemática aplicar
• calculo diferencial	• cuerda	• elemento oponer	• geometría clásico	• matemática heleno
• caracol de pascal	• cuerpo geométrico	• elementos	• geometría euclidiana	• matemáticas discreto
• característica	• cuerpos geométricos	• elíptico	• geometría no euclidiano	• matriz aumentar
• cardioide	• curva	• endecágono	• geometría proyectiva	• matriz cuadrar
• caso de factorización	• curva plano	• eneágono	• geometría	• matriz de cofactor
• casquete esférico	• curvar	• entero	• geómetra	• matriz diagonal
• cateto	• cálculo aritmético	• entero algebraico	• gruesa	• matriz inversa
• centena	• círculo unitario	• equidistante	• grupo cíclico	• matriz inverso
• centro	• cónica	• error relativo	• grupos cíclicos	• matriz simétrica
• cicloide	• cúbico	• esférico	• hexaedro	• matriz simétrico
• ciencia matemático	• decena	• espacio vectorial	• hexágono	• matriz triangular
• cifra	• decimal	• espiral	• hipocicloide	• matriz triangular inferior
• circunferencia	• denominador	• espiral de arquímedes	• hipotenusa	• matriz triangular superior
• circunferencia inscribir	• derivadas	• espiral hiperbólica	• homomorfismo	• matriz unitario
• circunferencia unitaria	• descomposición factorial	• existe	• icoságono	• mediatrices
• circunferencia unitario	• diagrama sagital	• existir	• idempotencia	• millón
• cociente	• diferencia de cuadrado	• exponenciación	• incentro	• minuendo
• combinaciones	• diferencia de cuadrados	• expresión matemático	• infinito	• mitad
• combinación lineal	• diferencia de cuadrados	• factor primo	• interseccion	• monomio
• completar el cuadrado	• diferenciable	• factores primos	• intersección	• monomio semejante
• congruencia	• directriz	• factorial	• intersección de conjunto	• morfismo
• conicas	• discriminante	• factorización	• invariante	• multiplicación de matrices
• conjunto	• distancia	• figura	• inversa	• multiplicación de matriz
		• figura geométrico		
		• figura plano		
		• finito		
		• forma canónico		

<ul style="list-style-type: none"> • multiplicación de radicales • máximo común divisor • mínimo común múltiplo • múltiplo • natural • nomenclatura matemático • notación decimal • numeración babilónico • numeración romana • numerador • numero real • numeros • numeros enteros • numeros racionales • numérico • número algebraico • número arábigo • número combinatorio • número componer • número de oro • número decimal • número entero • número fraccionario • número impar • número natural • número naturales • número negativo • número ordinal • número par • número perfecto • número periódico • número primo • número racional • número real • número reales • número trascendente • número triangular • números • números amigos • números compuestos • números entero • números natural 	<ul style="list-style-type: none"> • números perfectos • números primo • números primos • números racional • números racionales • números real • obtusángulo • octillón • octogonal • operación con polinomio • operación interno • operación matemático • oponer • ortocentro • ortoedro • par • par ordenar • paralelas • paralelepípedo • paralelepípedo recto • pentadecágono • pentagonal • pentágono • pentágono regular • permutaciones • perpendicular • perpendicularidad • perímetro • pirámide • plano • poliedro • poliedro regular • poligonal • polinomio • polinomios • posicional • postulado de euclides • precisión • primo relativo • primos relativos • principiar mathematica • prisma • producto directo • producto interior • producto matricial • producto notable 	<ul style="list-style-type: none"> • programación lineal • propiedad asociativo • propiedad distributivo • proporcionalidad • proporción • punto de inflexión • quintillón • radio • ratio • reciprocar • recta de euler • recta numérico • recta paralelo • rectangular • rectas paralelas • rectángulo • recíproco • regla de cramer • regla de sarrus • regla y compas • regla y compás • relación binario • relación de equivalencia • resta • restar • resto • resultante • ruleta • sección cónico • segmento • segmento circular • segmento rectilíneo • semiplano • semirrecta • sexagesimal • simplificación • simplificación de fracciones • simplificación de fracción • sistema decimal • sistema quinario • sistema vigesimal • subconjunto • subconjuntos • subtender • sumando • sumar 	<ul style="list-style-type: none"> • sustracción • sustraendo • teorema de pitágoras • teorema matemático • teoremas • teoría de conjuntos • teoría de ecuaciones • teoría de ecuación • teoría de grupo • teoría de número • tetraedro • topología combinatorio • transformación natural • trapecio • triangulo de pascal • triedro • trigonometría • trillones • trinomio • trinomio cuadrar perfecto • triángulo acutángulo • triángulo de pascal • triángulo de tartaglia • triángulo equilátero • triángulo isósceles • triángulo obtuso • triángulo obtusángulo • triángulo rectángulo • triángulo semejante • triángulos rectángulos • tronco de cono • tronco de pirámide • término algebraico • unidad • valor decimal • variable • vector columna • volumen • vértice • ví 	<ul style="list-style-type: none"> • álgebra • álgebra elemental • álgebra lineal • ángulo complementario • ángulo congruente • ángulo externo • ángulo inscribir • ángulo interno • ángulo suplementario • conjetura de taniyama-shimura • cilindrico • cilindricos • cilindro • esférica • elipsoide de revolución • conmutativa • conmutatividad • conmutativo • propiedad conmutativo • agrupación • asociativo • basilea • bolzano • búsqueda • calculo • cardinal • centímetro • ciclo • clase • columna • complejo • construcción • contenido • coordinación • correspondencia • criterio • cuártico • cálculo • cónico • diagonal • diferencia • elemento • entorno • error • espacio • estimación • estrella • existencial
---	--	---	--	---

- factor
- formula
- fundamental
- fórmula
- gráfico
- haz
- impar
- inclusión
- independencia
- inducción
- inverso
- isósceles
- kronecker
- lagrange
- landau
- largo
- lí
- línea
- matriz adjunto
- modelo
- máximos
- normal
- nulo
- orden
- paralela
- paralelo
- pascal
- pendiente
- primitivo
- puente
- puzzle
- pérdida
- raya
- recta
- recto
- reflexión
- región
- regla
- relación
- reproducción
- sector
- silla
- singular
- suma
- superficie
- tamaño
- teorema
- teorema de fermat
- traza
- triangulo
- triar
- triángulo
- tríada
- ángulo
- árabe
- \emptyset
- eje de simetría
- radio de curvatura
- esfera
- geometría esférica
- geometría esférico
- ángulo completo
- ángulo obtuso
- posición
- derivación
- agudo
- suplemento
- centroide
- mathematische annalen
- observador
- producto cruz
- vector
- william
- mínimos
- lineal
- lineales
- media aritmética
- media armónica
- media cuadrática
- media geométrica
- circular
- círculo
- problema matemático
- contemporáneo
- estéreo
- johann
- presente
- constante de gravitación
- trilateración
- belleza
- abscisa
- abscisas
- cercanía
- dada
- eje cartesiano
- error aleatorio
- ordenada
- ordenadas
- plano cartesiano
- greenwich
- kepler
- képler
- sol
- suceso
- geometría analítica
- geometría analítico
- modo
- alhazen
- aproximación
- augustus de morgan
- bernhard riemann
- bolt
- boole
- brook taylor
- cardano
- cauchy
- causa y efecto
- cigarros
- colin maclaurin
- colín maclaurin
- complicar
- continuar
- creer
- cuidado
- cálculo mental
- david hilbert
- decidir
- dedicar
- depender
- desconocido
- elie cartan
- ernst eduard kummer
- especial
- evariste galois
- exacto
- fallir
- fermat
- ferrari
- forma distinto
- frege
- gabriel cramer
- galois
- georg cantor
- george boole
- gerolamo cardano
- giro
- girolamo cardano
- giuseppe peano
- gustav lejeune dirichlet
- habilidad matemático
- haytham
- hilbert
- ida y vuelta
- inevitable
- interior
- isaac barrow
- it
- jerónimo cardano
- joseph liouville
- liouville
- lobachevsky
- mahavira
- marin mersenne
- maría gaetana agnesi
- max planck
- muhammad
- ojalá
- oro y plata
- oír
- peano
- pierre de fermat
- planck
- punto de vista
- rafael bombelli
- sebastián lerdo
- shakespeare
- sophie germain
- suma y resta
- taniyama
- tiempo límite
- vaso vacío
- viajar
- william lambton
- william oughtred
- xxy
- yutaka taniyama
- évariste galois
- ít
- conceptual
- 15
- mirar
- torre
- galardón
- premio
- bernhard
- moderno
- teatro
- dedekind
- elie
- abc
- absoluto
- ae
- agua
- alejar
- aliado
- amistad
- andrómeda
- animal
- antena
- antonio rodríguez
- apolonio
- aurelio
- azul
- año
- barra
- bata
- bay
- bloque
- bogotá
- bolívar
- bufón
- básico
- café
- camino
- canción
- carl
- carolina
- cartesiano
- casa
- castilla
- christiaan
- complejidad
- congreso
- constantino
- contacto
- coronel
- cortes
- cristóbal
- cristóbal colón
- cuarto creciente
- cuenta
- curar
- d2
- dam
- david
- diana
- dracma
- duelo
- edad
- edmund
- ejercicio
- eusebio
- fahrenheit
- fantasía
- felicidad
- francés
- franz

- friedrich
- fuego
- fé
- gerhard
- giovanni
- grúa
- guerra
- hiparco
- historia
- hora
- independiente
- inmortal
- intro
- izar
- jesús
- juan gonzález
- karl
- laura
- leonor
- ley
- llave
- london
- luis de velasco
- luna
- magia
- mamá
- man
- maxwell
- monto
- mujer
- márquez
- méndez
- méxico
- méxico
- narváez
- negro
- nicole
- niño
- nova
- novela
- obregón
- omega
- opinión
- oro
- patrick
- paular
- paz
- pedro fernández
- pepe
- ramírez
- rareza
- reforma
- revolver
- richter
- rodolfo
- rodríguez
- roque
- sergio sánchez
- sur
- sánchez
- sócrates
- tratado
- unir
- universo
- unión nacional
- valencia
- vida
- volver
- yuri
- áe
- áfrica
- ü
- juan garcía
- millar
- cuartil
- cuartiles
- desviacion media
- desviación media
- distribución de frecuencia
- distribución de frecuencias
- exactitud
- abel
- alfredo
- ali
- américa
- antonio
- antonio díaz
- antonio flores
- apasionar
- b4
- barrow
- bau
- becquerel
- bertrand
- caballero
- caballo
- cantero
- coma
- control
- corral
- cruz
- cuba
- diablo
- eduardo
- einar
- escobar
- esmeralda
- flandes
- galileo
- guevara
- hamilton
- hermano
- hernández
- imaginar
- impacto
- jj
- juana
- justo
- kummer
- leonardo
- lerdo de tejada
- lindemann
- lino
- luz
- lápiz
- marco
- marin
- matilde
- medida
- miguel
- mira
- mármol
- neumann
- océano
- pareja
- parís
- pedro hernández
- pedro suárez
- pena
- peña
- pintar
- posada
- promesa
- ricci
- rosál
- signo
- silencio
- sophie
- tierra
- troy
- valentino
- varga
- william jones
- zúñiga
- hacienda
- loco
- montevideo
- alemán
- cambio
- contrario
- córdoba
- extremo
- falta
- humanidad
- ignacio
- matías
- ortiz
- portero
- rango
- rené
- retorno
- rocha
- suárez
- teeteto
- data
- garaje
- incógnito
- m3
- mauro
- mover
- pj
- francisco
- nobel
- pavía
- persona
- sergio
- vázquez
- tom
- ac
- frecuencia
- acumular
- frecuencia relativa
- frecuencia relativo
- punto
- eugenia
- pedro
- matemático
- gauss
- jornada
- récord
- componente
- diagrama
- hipótesis
- datación
- cuadrado mágico
- sistema de numeración
- octal
- enrique garcía
- ap
- muestra
- binaria
- número binario
- sistema binario
- conejo
- multiplicación
- multiplicar
- demostración
- masa
- coordenada polar
- test
- criterio de divisibilidad
- división
- división de polinomio
- raza humano
- medalla fields
- patria
- pt
- fl2
- d1
- d3
- d4
- curva elíptico
- ál
- adán
- j2
- dimensiones
- dimensión
- habitante
- número de habitante
- blanco
- pc
- af
- dividiendo
- residuo
- áf
- problema aritmético
- aguja
- sección
- artificial
- discovery
- hombre
- función constante
- radicación
- profundidad
- profundo
- unidad de tiempo
- cartaginés

- adrien marie legendre
- legendre
- árbol
- coordenada
- coordenadas
- sistema de coordenada
- proyección de mercator
- funciones especiales
- ceros
- l2
- hb
- algoritmo de euclides
- exp
- operando
- análisis numérico
- ecuaciones
- ecuación
- forma modular
- funciones trigonométricas inversas
- función cúbico
- función lineal
- función trigonométrica
- función trigonométrica inversa
- sistema de ecuación
- sistema lineal
- sumatorio
- cantidad
- mundo virtual
- crecimiento bacteriano
- cuadrado
- pedir
- simetría
- jo
- mundial
- capacidad
- rama
- perspectiva
- paréntesis
- cartesiana
- metodo
- método
- joven
- juvenil
- viejo
- secundario
- forro
- menores
- papiro de ahmes
- papiro de rhind
- papiro rhind
- azar
- anos
- año
- leonardo da vinci
- hacendado
- axioma de peano
- axiomas de peano
- axiomatización
- axiomáticamente
- axiomático
- fallecimiento
- morir
- prisión
- abu
- agustín
- ajusco
- bienvenido
- brasilia
- capetown
- caracas
- ciudad de méxico
- congreso de berlín
- conjunto vacio
- delegación benito Juárez
- df
- docena
- erdös
- estocolmo
- iztaccalco
- londinense
- moscú
- museo británico
- norte
- obra
- otawa
- paul erdös
- royal society
- subir
- varsovia
- ciudad capital
- número de euler
- combinatoria
- dae
- daé
- ecuaciones lineales
- ecuación algebraico
- ecuación cuadrático
- ecuación cúbica
- ecuación diferencial
- exponente
- funciones continuas
- función continuo
- fórmula de stirling
- geografía humano
- inecuación
- integración numérico
- inteligencia artificial
- multiplicidad
- método de simpson
- numero complejo
- número complejo
- número irracional
- números complejo
- números complejos
- números irracional
- potenciación
- raíz cuadrar
- raíz cúbico
- sucesiones
- sucesión infinito
- sucesión numérico
- superficie de revolución
- sólido de revolución
- álgebra booleano
- radianes
- radiar
- histogramas
- ip
- calendario
- fecha
- análisis
- campeonato mundial
- inconmensurabilid ad
- logro
- β
- inclinación
- reconocimiento
- lápiz de color
- cuadro
- goma de borrar
- bp
- ctv
- daf
- p1
- afc
- carlos
- liga
- mano
- nn
- corea
- cuarto
- dormitorio
- habitación
- minuto
- an
- horizontal
- vertical
- gómez
- cheff
- isaac newton
- tacón
- est
- cénit
- alquimia
- ib
- lustro
- mes
- meses
- minutos
- septenario
- siglo
- ojo
- año solar
- annals of mathematics
- fútbol americano
- senos
- blastómero
- ayuda
- auspicio
- diagrama de árbol
- fijo
- horóscopo
- joel
- precio
- redar
- concepción
- costo
- costo fijo
- euclides
- numeración egipcio
- anual
- compuesta
- entera
- inciso
- oportunidad
- truncar
- astrólogo
- mapa mental
- almorzar
- aperitivo
- especia
- gual
- guiso
- limonada
- ordenar
- pan blanco
- pasar
- receta de cocina
- árbol frutal
- buenaventura
- binar
- carne de pollo
- frecuencia cardiaco
- presión arterial
- acariciar
- entrenador
- galleta
- protestar
- qrs
- xe
- aburrir
- actitud
- batalla de himera
- decepción
- desear
- farmacia
- locura
- medicina
- perfección
- psicología
- sufrir
- test psicológico
- simbólico
- símbolo
- símbolos
- argentino
- muerte

- | | | | | |
|--|---|---|--|--|
| <ul style="list-style-type: none"> • atención • imaginación • ensayo • escala • nao • barbilla • bicondicional • bicondicionalidad • cuantificador • cuantificadores • cuerpo humano • doble implicación • regla de inferencia • semana • oriente • equipo de fútbol • irb • nominador • dce • electrodoméstico • sistema • alameda • av • ballenato • central • claudio • cándido • código • dublín • e2 • guaraní • juan • lima • lisboa • lorenzo • margen izquierdo • mauricio • n2 • n3 • nacional • nápoles • ottawa • penetrar • prefecto • reactor • salón • satelital • sena • seña • sitio de siracusa • sumir • tehuantepec • umbral | <ul style="list-style-type: none"> • vacio • vacío • venta • Víctor • zócalo • aprendizaje significativo • batear • bolo • consolación • delantero • itziar • mónica • quíntuplo • terminar • trasero • técnica de estudio • berlín • latín • panamá • sión • sofiá • trujillo • wellington • west • basket • básquetbol • método matricial • becerril • desigualdad • inteligencia • irracional • solís • método de newton-raphson • integración • integral • marco antonio • compas • comprensión lector • compás • derivar • himalaya • reducir • sorprender • sorpresa • arrancada • excursionista • fig • notación algebraica • tenis • tt | <ul style="list-style-type: none"> • absolutamente convergente • asíntota • asíntota horizontal • asíntota oblicuo • asíntota vertical • campana de gauss • concavidad • convergencia absoluto • convexo • cóncava • ecuación cuártica • ecuación cuártico • exponencial • forma lineal • función exponencial • función implícito • función logarítmica • función logarítmico • función matemático • función primitivo • función racional • fórmula trigonométrico • gráfica de función • gráficas de funciones • identidad trigonométrico • incógnita • integración por partes • integral indefinido • integrales • ln • logaritmación • logaritmo común • número imaginario • números imaginario • números imaginarios • resolución de ecuación • tabla de verdad • valor absoluto • cp • expansión | <ul style="list-style-type: none"> • ajedrecista • ajedrez • avidez • tablero de ajedrez • ai • aí • billar • urbe • extinción • bello • comunicación • ama de casa • diagrama de flujo • revista • semáforo • pretender • cotidiano • desarrollo • cuadradas • estatura • figuras • producción • adar • antisimétrico • año bisiesto • año comercial • bisiesto • brindar • fiesta • mp • noche • sencillo • sencillos • ternario • concierto • festivo • inaugurar • milla marino • problema • aprender • aprendizaje • caminar • casado • correr • estilo de aprendizaje • fardo • perforación • signo de integral • sobrino • tutor • viaje • a0 | <ul style="list-style-type: none"> • a2 • error sistemático • pitagorismo • pitágoras • feria • aplique • armónico • coraje • diagramas • médico • sacapuntas • transportador • danza • documentos • mediante • movimiento revolucionario • tabulacion • tabulación • instrumento musical • lúe • mayoría • amabilidad • apóstrofo • código de comercio • evento deportivo • imposible • acuñación • centavo • centésimo • moneda • monedas • experiencia • moneda de oro • agua tranquilo • bola de billar • buzo • infancia • ingeniería • menor • ocio • patinador • patinaje • dis • ul • billetero • bolsillo • calculo de probabilidades • calzado • camisa |
|--|---|---|--|--|

• cálculo de probabilidad	• descripción	• combinar	• booleano	• oí
• cálculo de probabilidades	• encomio	• ficticio	• boya	• pabellón
• disfraz	• esconder	• juego	• capitalista	• hoja
• disfrazar	• escondidas	• jugar	• clases	• josé
• fórmula	• escondite	• ka	• cod	• pardo
• opción	• poema	• lápiz y papel	• concepto	• botar
• pantalón	• protagonizar	• pasajero	• construcción naval	• inicial
• terciopelo	• resumen	• tachuela	• cy	• instrumento de navegación
• tirante	• retórico	• homenaje	• descubrimiento	• rené descartes
• tocar	• ídem	• quiebra	• descubrir	• imaginario
• uniforme	• agricultura	• rotafolio	• duda	• padre
• diagrama de barra	• cambiar	• punta	• eléctrico	• instante
• gráfica	• cosechero	• rogelio garcía	• empírico	• autorización
• gráfica de barra	• embalsamamiento	• abatís	• fenómeno	• cardan
• noreste	• olvidar	• compendio	• filósofo	• efecto
• presentación	• rastrojo	• engrapadora	• filósofo	• engranar
• anónimo	• sector agrícola	• libro	• goro	• estética
• prenda de vestir	• topógrafo	• menón	• hc	• gusto
• principio de igualdad	• yac	• cara o cruz	• ignorar	• herramienta
• probabilidad	• dicho	• caracterización	• imputar	• ingeniero químico
• pulsera	• introducción	• caracterizar	• integración vertical	• interruptor eléctrico
• ropa	• metáfora	• diálogo de platón	• intuición sensible	• motor
• sortija	• modus ponens	• eneadas	• iterar	• máquina
• aplauso	• modus tollens	• isaac todhunter	• juicio de valor	• método empírico
• apodo	• suspenso	• lotería	• lancha	• operación union
• comportamiento	• cita	• rifa	• lingote	• radicales
• equipo	• congelación	• robert recordé	• mesana	• sustancia
• filibustero	• cuartillo	• suerte	• modo gráfico	• sustancia
• invención	• descomposición	• taba	• mínimo	• velocidad de transmisión
• inventar	• noticia	• águila o sol	• noción	• americano
• invento	• qi	• hijo	• novedad	• bordo
• isbn	• siembra	• individualidad	• pensar	• bra
• boda	• ejecutar	• pronósticos deportivos	• persecución	• chino
• anécdota	• emeterio	• labio	• personaje	• combate
• biográfico	• máximo	• traducción	• practicar	• cáscara
• cuentos	• equilibrio	• ama	• principio	• fondo
• miscelánea	• subnormal	• carta	• principios	• guarnición
• navegantes	• distribución de gauss	• interpolación	• práctico	• indicación
• piropo	• distribución de probabilidad	• lourdes	• raciocinio	• manzana
• tumba	• distribución normal	• p2	• realidad	• pastel
• cocinar	• distribución normal	• planeta	• recogedor	• receta
• cámara	• función de distribución	• posesión	• saqueo	• celebrar
• extensión	• función de probabilidad	• crecimiento	• simulador de vuelo	• experimento mental
• pieza	• tipo de descuento	• gallo	• software	• habilidad
• quinta	• burro	• abatimiento	• supresión	• teórico
• absurdo	• griego	• algoritmo	• tecnología	• cot
• alegórico	• piedra	• algoritmos	• trasatlántico	• e6
• antonomasia	• ao	• algorítmico	• trascendencia	• escocés
• autobiografía	• canelo	• arresto	• técnica	• fruto
• colgar		• automático	• típico	
		• barco	• oi	
		• bitácora		

• mana	• canalón	• pelota	• unidades de medida	• takahashi
• piña	• cb	• reloj	• yarda cuadrar	• babilonio
• pomo	• conocimiento	• s1	• ángstrom	• director
• tc	• ingrediente	• s6	• a1	• pantalla
• tratamiento	• intelectual	• sara	• b2	• piano
• externo	• material	• sombrero	• balón	• experto
• caja	• medición	• toba	• be	• legítimo
• carbón	• observatorio	• ultra	• conversar	• marrón
• especie	• paralaje	• íg	• dama	• monumento
• gasto	• pedestal	• canica azul	• hardy	• punto de congelación
• mente	• piso	• cen	• legado	• relojero
• pv	• plinto	• corto	• maestro	• verde
• renacimiento	• saliente	• din	• mesa	• arreglo
• bagazo	• sistema de unidad	• diseñador	• piscina	• carrusel
• bellas artes	• sistema métrico	• diseñar	• sombra	• dj
• entrenamiento	• sistema métrico decimal	• diseño	• velocidad	• escritura
• fabricación	• teoría de código	• línea base	• axial	• letra
• industria	• interés	• obrero	• bal	• patrón
• lección	• acreedor	• limosna	• completo	• presto
• ley físico	• deudor	• marino	• cup	• tema
• línea de visión	• empleado	• regalo	• eje	• amigo
• modernidad	• empleo	• rica	• gla	• ahmes
• respeto	• jornal	• voluntario	• mj	• aj
• sistema de transmisión	• salario	• convencionalismo	• oro puro	• cometer
• telecomunicaciones	• trabajador	• error tipográfico	• proyección	• componer
• telecomunicación	• varón	• referencia	• representación gráfico	• compuestas
• verificación	• dedo	• propina	• clavar	• consolidar
• actividad	• fanático	• doble	• go	• cártamo
• clavo	• m1	• fi	• brahmagupta	• diofanto
• confusión	• or	• ff	• escribir	• diofanto de alejandría
• consultar	• procesador	• triple	• experimentación	• eratóstenes
• inn	• sat	• ilustrador	• mecánico	• forma
• iu	• adulterar	• condición	• ortografía	• hecho histórico
• od	• FALSO	• reducción	• bot	• historiador
• p3	• cf	• hermano de jesús	• paso	• magüey
• pan dulce	• milán	• adarme	• stirling	• oyamel
• restricción	• afelio	• alcuota	• estructura	• papiro
• tipo	• desocupar	• bushel	• pastor	• trébol
• tópico	• estatua	• decigramo	• poseidón	• zea
• xo	• linear	• femtómetro	• tiempo	• farol
• cerradura	• perihelio	• hm3	• ys	• industrial
• daí	• semieje menor	• km3	• bbb	• operar
• escocia	• vacaciones	• litro	• célula	• antiguo egipcio
• foto	• órbita elíptico	• nm	• propiedad	• caber
• fotografía	• padre de familia	• pennyweight	• presentador	• cascarilla
• negativa	• alejandría	• pie cúbico	• visualizar	• clément
• negativo	• arte	• pulgada	• auto	• cuneiforme
• vendo	• esp	• tabla de conversión	• soldado romano	• dejar
• yi	• esquina	• unidad de longitud	• cuaderno	• epigrama
• t3	• gh	• unidad de medida	• europa	• fenicio
• billetes	• hr		• motivo	• girar
	• marta			

- hipopótamo
- jeroglífico
- mar rojo
- numerus
- nunes
- peck
- res
- rés
- watts
- capa
- carrera
- dirección
- partida
- registro
- poesía
- basura inorgánico
- basura orgánico
- hidrostática
- periodo de gestación
- auc
- acá
- estilo
- expresión
- fa
- fuente
- g5
- ameno
- artes
- bernardo
- capí
- contención
- espejar
- guardar
- iguala
- ji
- lenguaje
- madero
- naval
- renunciar
- saldar
- science
- taberna
- tío
- época
- arquímedes
- héctor
- substitución
- sustitución
- valor
- acento
- corte
- instrumento
- lectura
- revolución
- cí
- fg
- noemí
- pinta
- pían
- cuarta
- distorsión
- octava
- octavo
- onda sonoro
- radiocarbono
- animación
- catálogo
- colección
- lempira
- reserva
- visión
- escuela
- hp
- demostrar
- descenso infinito
- hz
- inferencia
- logico
- lógica
- lógica matemática
- lógica simbólica
- lógica simbólico
- lógico
- mentira
- pensamiento lógico
- razonamiento analógico
- tautología
- casilla
- pelayo
- zapato
- contingencia
- deductivo
- definición
- denotación
- escala richter
- necesario
- posibilidad
- razonar
- respuesta
- significado
- verdad
- visual
- abanico
- bingo
- continente
- nota
- st
- terrestre
- transporte
- baldor
- ciego
- espina
- arete
- chaleco
- corbata
- falda
- gabardina
- playero
- segmentar
- seguridad
- abreviación
- abreviada
- abreviadamente
- abreviar
- semántica
- diamante
- estadio
- nob
- punto cardinal
- botín
- pañuelo
- prenda
- tm
- ui
- bulbo
- nacimiento
- oliva
- doctorar
- obra de teatro
- ducto
- hectogramo
- iniciar
- proceso
- procesos
- programación
- samuel
- conversión
- creación
- diario
- memoria
- moraleja
- seguro
- síntesis
- agrario
- cuadra
- cultivo
- editorial
- einstein
- atar
- atx
- bde
- bl
- calculadora
- calculadora mecánico
- circuitos
- circuitos lógicos
- compu
- computador digital
- computador portátil
- cons
- electrónica
- electrónico
- intuición
- john von neumann
- lenguaje de computador
- lenguaje de programación
- máquina analítico
- máquina de sumar
- nx
- oscilador
- portátil
- positivo
- shift
- tarjeta perforar
- tecla shift
- tiempo real
- tso
- vme
- von neumann
- análogo
- lerdo
- miligramo
- pala
- transmisión
- abaco
- biyectiva
- biyectivo
- bnc
- cable telegráfico submarino
- cambio de variable
- charlar babbage
- codominio
- cosecante
- cotangente
- cultivo energético
- d/a
- derivada parcial
- desintegración radiactivo
- dios
- dominio de definición
- ecuación exponencial
- ecuación paramétrico
- ecuación pitagórico
- energía fósil
- experimento
- funciones inversas
- función biyectiva
- función biyectivo
- función componer
- función entero
- función escalonar
- función identidad
- función inversa
- función parcial
- función sobreyectiva
- función trascendente
- if
- interpolar
- intervalo abierto
- intervalo cerrado
- inyector
- logaritmo neperiano
- lr
- mano libre
- mantisa
- material radiactivo
- multiplicador
- progresión aritmética
- progresión aritmético
- progresión geométrico
- q8
- radiactividad
- referente
- segundero
- serie de fibonacci
- serie de maclaurin

<ul style="list-style-type: none"> • serie de potencia • serie de taylor • serie geométrico • serie y sucesión • sucesión de fibonacci • suprayectivo • taladrar • terna pitagórico • variable complejo • velocidad de corte • ábaco • if • ómnibus • bombay • capital • oz • común • albañil • cables • carro de guerra • conclusión • concreto • conjunto musical • desviación estándar • esfuerzo • fibrocemento • grupo de trabajo • hormigón • investigación • investigar • jordano • mampostería • medida de dispersión • medidas de dispersión • monotonía • observación • plano inclinar • rectificación • regla de cálculo • suavizar • telar • variancia • distribución • x2 • gabriel • antena parabólico • apisonar • bajar • buscador 	<ul style="list-style-type: none"> • bx • dó • edificio • elemental • explicación • internet • motor de búsqueda • razón de proporcionalidad • resistencia interno • seo • televisor • vínculo • áridos • convenio • hinojosa • préstamo • @ • azumbre • celemín • centro histórico • contratista • cuadrante • cámara oscuro • cántara • digitalización • digitalizar • energía eólico • fanega • finca • flexómetro • fotocopiar • granos • hectómetro • ilustrada • impresora • impresoras • kilómetro • km • lb • legua • leguas • milímetro • navegante • nonio • onza troy • palmo • qq • quintal • reloj atómico • sonido • izquierda • librería 	<ul style="list-style-type: none"> • abecedario • alfabeto • arcada • cabeza • detectar • escolástico • minuterero • ny • orden alfabético • programa de concurso • reactor nuclear • reflector • reloj de péndulo • reloj de sol • rr • vav • aberración • champaña • ron • sci • armador • boliviano • colón costarricense • cronómetro • dólar beliceño • dólar canadiense • dólar estadounidense • locación • my • peso chileno • peso colombiano • peso cubano • peso dominicano • peso mexicano • peso uruguayo • prelado • rd • sistema de escritura • telescopio • termómetro • br • hábito • autocad • cartulina • furlong • kg • kilogramo • metro cuadrar • metro lineal • metros 	<ul style="list-style-type: none"> • quark down • tonelada • tonelada largo • yarda • sim • arts • sextante • abstracto • atajo • autor • comando • editor • escobilla • fenómenos • ff • ft • galáctico • girard • i5 • idea • justificación • lab • losa • materia • navegación • objetivo • objeto • palo • pd • pensamiento • pie • pl • procedimiento • prueba • sp • suelo • topología • tránsito • vista • ámbito • í5 • óscar • serie alternar • platino • ral • avión • colorar • mer • key • mecanismo • ua • abeliano • aserrín 	<ul style="list-style-type: none"> • ban • cenefa • clérigo • costura • familia de religión • fray • jabón • joyero • lana • librado • ondular • perfume • platero • pleno edad media • previo • serrucho • terna • torneo • afi • broca • corriente • desplazamiento • interruptor • kí • mecánica • partícula • pegar • soga • antiguo grecia • augustus • biblioteca de alejandría • bruñir • califa • cigarro • circón • clasificación • cnido • digno • du • egipcio • eudoxio • eudoxo • eudoxo de cnido • evolución tecnológico • hades • hongkong • jacinto • juicio universal • ls • marfil • mesar
---	---	--	---	---

• musa	• billete de banco	• universidad de poitiers	• estadística	• jugador
• panadería	• billetes de banco	• yen	• gerente	• pago
• papel higiénico	• butaca	• adelantado	• necesidad	• pagos
• patricio	• ciu	• barra vertical	• negociación	• abismo
• piedra precioso	• escuela pitagórico	• belice	• poles	• naranja
• piedras preciosas	• fila	• beliceño	• tonel	• ra
• platón	• flotación	• firma	• transportes	• aa
• porfirio	• gremio	• morir de risa	• descuento	• acera
• puro	• isaac	• punto suspensivo	• descuentos	• aristarco
• puros	• islámico	• punto y comer	• dinero	• barrón
• quadratura	• leyes	• antiguo griego	• letra de cambio	• cerillo
• quilates	• mueblería	• excepción	• pq	• derecha
• quinario	• musulmán	• interpretación	• taller	• h3
• rm	• pitagórico	• julia	• tr	• letrero
• sufi	• principio de arquímedes	• tera	• valorar	• morelia
• varar	• rey salomón	• men	• padua	• mundillo
• vida cotidiano	• tenedor	• trabajo	• creativo	• persa
• zi	• buc	• trabajó	• repertorio	• ptolomeo
• zí	• cuantificacion	• recibo	• ganancia	• secretario
• v2	• t5	• pintura	• desplazar	• shu
• v3	• tela	• divisa	• mob	• tolomeo
• v4	• civilizar	• antigüedad	• sucedáneo	• washington
• al-mamun	• cátedra	• bonito	• vocablo	• acerar
• aum	• imperio	• cuestión	• conforme	• arquitectura
• calle	• obispo	• desorden	• acogida	• audiencia
• carretera	• abd	• lista	• be2	• benito
• católico	• estaca	• sup	• halar	• caballería
• despertar	• inglés	• anualidad	• morado	• dominar
• durmiente	• pre	• caja de madera	• pagaré	• fernández
• himera	• roldan	• capital social	• xq	• gallinero
• jesuatos	• rv	• corta	• acre	• guitarra
• jesuita	• atto	• descuento comercial	• onza	• ibérico
• judío	• brook	• espacio de probabilidad	• taquilla	• jarro
• mb	• canadiense	• especificación	• banquero	• lata
• medir	• canasta	• estuche	• cuenta de ahorro	• maravilla
• nicomaco	• centauri	• evento estadístico	• interés componer	• mará
• nicomaco de gerasa	• cesta	• fenómeno estadístico	• interés simple	• serpentina
• precursor	• esclavo	• jefe	• negociar	• sombrerera
• radiofaro	• femto	• logística	• plazo fijo	• ternero
• religión protestante	• fiore	• matasellos	• tomador	• escúpulo
• silogismo	• ghana	• posicionamiento	• crucero	• fabricar
• sj	• inquisición	• superior	• reino	• fábrica
• teano	• jarra	• tabla de frecuencia	• asegurado	• decima
• tradicional	• kilos	• táctica	• intereses	• acero
• tumo	• lin	• vida útil	• publicidad	• egipto
• via de ferrocarril	• lín	• consola	• póliza de seguro	• quilate
• áureo	• melbourne	• administrativo	• rédito	• álamo
• r1	• pueblo maya	• barril	• seguro de vida	• totalidad
• utilitario	• roc	• emprender	• seguros	• altiplano
• áy	• shanks		• usura	• cienfuegos
• adam	• sidney		• tche	• escalonar
			• pintada	• mar
			• actual	• rojo

• uc	• valfa	• loar	• viaducto	• uruguay
• voto	• rolando	• caja de ahorro	• acertijo	• madera
• ochavo	• ars	• ace	• alma	• misi3n imposible
• pedro herrera	• comandante	• arroba	• cielo	• dal
• variar	• ejercitar	• balanza	• flamenco	• avicultor
• aver	• ingeniero militar	• ciudad	• mosa	• avicultura
• famoso	• mercenario	• civilizaci3n	• noe	• avicola
• hundredweight	• mere	• ilustraci3n	• texcoco	• escala celsius
• tif	• mer3	• vara	• osamenta	• richard taylor
• tve	• proyectil	• trueque	• bolonia	• tribu
• can	• reconocimiento	• h3	• rafael	• portugu3s
• chileno	• especial	• pe	• rollo	• naturaleza
• colonia	• sargento	• ro	• ventanilla	• suborden
• efe	• teniente	• tyr	• andrew	• sub3rdenes
• fourier	• tropa	• ve	• cba	• s3nscrito
• guadalajara	• comodoro	• yr	• havre	• dag
• guara	• biblioteca	• yu	• ingenio	• frey
• malo	• bucle	• z3	• mega	• indu
• montreal	• video	• ceres	• r7	• cancela
• pri	• at	• eos	• roberval	• competer
• ricardo gonz3lez	• cab	• equis	• sima	• garraf
• sandoval	• dec	• orf	• toronto	• lañar
• santillana	• f5	• paraguay	• b3	• losar
• sec	• f8	• amperes	• ejecuci3n	• mendiz3bal
• seãan	• garrafa	• gale3n	• newman	• ura
• soa	• instrucci3n	• disciplina	• roxana	• alano
• uf	• nas	• fundici3n	• italiano	• premio nobel
• dya	• tarjeta	• hermandad	• joaqu3n	• escala fahrenheit
• eniac	• 3al	• pac3fico	• barrica	• punto de ebullici3n
• pte	• aditivo	• bandera	• financiera	• japon3s
• flaco	• afrodita	• cart3n	• s5	• carpa
• luisa	• avance	• colono	• tipo de inter3s	• pvq
• marcar	• dep3sito	• coreano	• organismo	• terreno
• mart3nez	• margen	• io	• gar3	• margen derecho
• ruiz	• palm	• lechero	• barro	• regiomontano
• simple	• tanque	• mileto	• palabra	• pavo
• bocina	• adoqu3n	• orar	• presidente	• mar3timo
• toño	• escritorio	• pf	• pul	• nomenclatura
• monje franc3s	• resoluci3n	• telecom	• publicaci3n	• arsenal
• ball	• tiro	• xp	• publicar	• anc
• casar	• trazo	• zhang	• assurbanipal	• tecla
• compartir	• v6	• converso	• sardan3palo	• nip3n
• deserci3n	• v8	• culto	• ses	• uzbekist3n
• analog3a	• volts	• disc3pulo	• vp	• isla filipinas
• contradicci3n	• a9	• faja	• copiar	
• presa	• alojamiento	• k2	• cortar	
• razonamiento	• d3	• movimiento	• diab3lico	
• ab	• glorieta	• religioso	• aõejo	
• b1	• hospedaje	• mueble	• escolar	
• c6	• kiosco	• religioso	• capitular	
• solidaridad	• leyden	• ulterior	• c3	

Anexo G: Lista de términos validados de ecología de bachillerato

La lista de términos validados presentada a continuación es la obtenida por el coeficiente de dominio con mejores resultados.

Para pesos de TF-IDF > 0 y CDImc

• abiótico	• neutralismo	• ambiente natural	• ganado caprino	• letargo
• acaricida	• organismo	• emerger	• ecosistema	• ave marino
• amensalismo	• consumidor	• chapopote	• recurso natural	• ave residente
• biodegradabilidad	• presión	• alóctono	• tierra de cultivo	• spp
• biodegradación	• atmosférico	• sedimentación	• benceno	• comportamiento animal
• biótico	• cooperación	• sedimentar	• hidrocarburo	• descendencia
• clima	• recurso no renovable	• sedimentarios	• lago oligotrófico	• descendiente
• clima árido	• red alimentario	• sedimento	• asno	• etología
• clordano	• relación biológico	• erosión hídrico	• servicio ambiental	• sociobiología
• clímax	• restauración	• desecación	• zona vulnerable	• trucha arcoiris
• comensal	• saprofito	• áreas proteger	• radiación terrestre	• visión
• condición climático	• semiárido	• especie exótico	• efecto invernadero	• viento alisio
• conservación biológico	• serie	• especie introducir	• ecología político	• reforestación
• contaminación	• simbiosis	• clorofluorocarbono	• ancianidad	• depredación
• contaminante	• simbiótico	• población biológico	• anciano	• depredador
• corriente de humboldt	• sinecología	• bióxido de carbono	• joven	• sésil
• corriente marino	• sumidero de carbono	• co2	• juvenil	• aldea
• corriente oceánico	• sustrato	• óxido de carbono	• viejo	• biodiversidad
• crisis ecológico	• tipo de clima	• gei	• biodiversidad	• bioprospección
• cálido	• tropismo	• naturaleza	• bioprospección	• diversidad biológica
• ddt	• comunal	• migratorio	• diversidad biológica	• diversidad biológico
• descomponedor	• crecimiento logístico	• pirámide de edad	• número de especie	• número de especie
• desecho tóxico	• curva logístico	• habitante	• variación genético	• variación genético
• destrucción de hábitat	• relicto	• número de habitante	• thomas malthus	• thomas malthus
• diclorodifeniltricloroetano	• territorial	• población mundial	• fértil	• fértil
• ecología de comunidad	• territorio	• crecimiento poblacional	• tasa de natalidad	• tasa de natalidad
• ecología de población	• selección natural	• óxido de nitrógeno	• índice de natalidad	• índice de natalidad
• emisión	• dinámica de población	• óxidos de nitrógeno	• regadío	• regadío
• explotación	• mar de aral	• sulfuro de hidrógeno	• riego	• riego
• extinción masivo	• hidroponia	• ciudadano	• posindustrial	• posindustrial
• geotropismo	• hidroponía	• ambiental	• globalización	• globalización
• heptacloro	• circulación atmosférico	• ambiente	• globalizador	• globalizador
• hábitat	• vertido	• medioambiental	• tienda de mascota	• tienda de mascota
• ley ambiental	• basura	• nutriente	• salix cinerea	• salix cinerea
• masa forestal	• desechar	• nutrientes	• raza humano	• raza humano
• microclima	• desecho	• vendaval	• cripsis	• cripsis
	• condición medioambiental	• viento	• matar	• matar
			• morir	• morir
			• muerto	• muerto
				• gregarismo

• acidificación	• nido	• fosa oceánico	• ine	• frecuencia
• capa de ozono	• chotacabras	• golfo	• agregado	• cebra
• celulosa	• cormorán	• orilla	• capacidad	• greenpeace
• distribución geográfico	• gaviota	• península	• índice	• evolución
• domesticación	• lechuza	• talud continental	• impacto	• apio
• domesticar	• superdepredador	• cárcava	• aral	• entorno
• afloramiento	• alcalinidad	• residuo	• migración	• life
• fondo marino	• carpa	• seguridad	• respuesta	• satélite
• surgencia	• canibalismo	• problema	• anchoveta	• primitivo
• pastizales	• mariposa monarca	• plancton	• cachorro	• barda
• gameto femenino	• pulmonar	• cerebral	• variar	• nitrógeno
• mitigación	• autotrofo	• cerebro	• huella	• escala
• combustible fósil	• autótrofo	• aparato digestivo	• control	• régimen
• calentamiento global	• autótrofos	• ecosistema acuático	• necton	• camino
• endemismo	• reproducción	• mariposa	• temperatura corporal	• distorsión
• especie endémico	• marea	• homeostasis	• polinizador	• radiación solar
• bioma	• tolvana	• carbón	• temperatura ambiente	• ecotipo
• biomas	• bosque húmedo	• gas	• terrestre	• insolación
• biogeoquímico	• selva bajo caducifolio	• bosque de chapultepec	• indico	• enredadera
• adelgazamiento	• selva húmedo	• obra	• océano pacífico	• ayuda
• adelgazar	• escama	• dulzura	• benedicto	• modelo
• banco de pez	• esqueleto	• magnesio	• mariano	• sistema
• agua marino	• proboscis	• protección	• árboles	• brócoli
• cabeza	• martín pescador	• estomacal	• recurso	• epa
• garra	• árdea	• estómago	• castor	• mutación
• cotorra	• digestivo	• hidrógeno	• pargo	• método
• ganso	• digestión	• dipnoos	• citar	• goma
• rapaz diurno	• australopithecus afarensis	• planta acuático	• epífita	• parlantes
• adrenal	• arcoiris	• vegetación acuático	• halófita	• melón
• regenerar	• nopal	• branquia	• planta trepador	• paso
• deriva genético	• emisión de co2	• singer	• rastrero	• manifiesto
• gestación	• intestino	• externo	• orden	• depredador
• bosque boreal	• degradación	• hueso	• fósil	• presa
• fauna	• renal	• co	• registro fósil	• mejillón
• neotropical	• compensar	• asma	• variación	• movilidad
• neártica	• ciudad de méxico	• neuston	• arbusto	• analogía
• abisal	• límite	• gusto	• programa	• desarrollo
• batial	• superficie	• depositar	• mayoría	• híbrido
• hadal	• constante	• depósito	• demanda	• accidente
• léntico	• oxiuro	• capacidad de carga	• demandante	• sexo
• vector	• encía	• ecologismo	• temperatura	• sexual
• ballenero	• hepático	• ajusco	• índice de mortalidad	• intocable
• cebo	• hidrosfera	• subir	• tierra	• gestión ambiental
• jauría	• dormancia	• pez ángel	• tasa	• cascabel
• extinto	• arca	• zoológico	• thomas	• pendiente
• asociación vegetal	• vómito	• zooplancton	• blanco	• pareja
• ictiosaurio	• zona intertropical	• mercurio	• wombat	• explorador
• órgano sexual	• desierto	• cría selectivo	• concentración	• ley
• transición	• relámpago	• planta suculento	• localización	• pionero
• olor	• estrecho	• suculento	• muerte	• expansión
• verdura	• estuario			• sahara
				• homo sapiens neanderthalensis

• neanderthal	• man	• pulpa	• eje	• amplitud
• neanderthalensis	• metrópolis	• alimentación humano	• radio	• umbral
• california	• miramar	• nacional	• atmósfera	• canino
• oxigenar	• mono	• o2	• contemporáneo	• endémico
• oxígeno	• movida	• regla	• fibra	• gallina
• local	• méxico	• niño	• visión	• gallinas
• dieta	• penetrar	• traza	• vista	• homo erectus
• contenido energético	• pirámide	• evapotranspiración	• antonio	• homo habilis
• fruto	• reactor	• ostra	• loyola	• mosquito
• deriva	• rosario	• delgado	• endosulfán	• ola
• caja	• snake	• tratado	• tiburón	• transporte
• energía	• venta	• rama	• calabaza	• eco
• uso sostenible	• tala	• control biológico	• agrícola	• reclutamiento
• colonia	• espejo	• flora	• bosque templar	• grave
• barracuda	• piel	• insostenible	• acuático	• altura
• hospedero	• pelo	• sostenible	• distribución	• intervalo
• comunicación	• mover	• sustentable	• línea	• agudo
• Perú	• ecuador	• erwin	• célula	• absorción
• aparato	• metamorfosis	• alvéolo	• estanque	• caverna
• órgano	• cacería	• cálculo	• síntesis	• cueva
• oso	• cazador	• muela	• fuerza	• omega
• vivo	• excepción	• muela	• bodega	• abrigo
• salvaje	• cáncer	• tiroides	• lobo	• puente
• central	• carpintero	• gavián	• machete	• valle
• humedad	• carpintero	• guacamayo	• rayo	• ozono
• reutilizar	• garza	• halcón	• arrecife	• bajío
• descomposición	• hormiguero	• corteza	• política	• plataforma
• líquen	• pico	• lechuguilla	• aire	• fundamental
• líquenes	• pingüino	• garganta	• anguila	• costa
• colectivo	• labio	• sigmoide	• matrimonio	• isla
• francés	• rural	• pájaro	• beta	• acuario
• banco	• zona rural	• n2	• volumen	• delfín
• mosca	• área rural	• papa	• aguja	• océano
• bulbo	• fósforo	• silvestre	• caballo	• talud
• asahi	• alternativa	• flor	• cabra	• micorriza
• borneo	• proyecto	• gente	• chile	
• caracol	• deforestación	• diente	• julia	
• centro	• eutroficación	• luna	• lenguado	
• concepción	• cambio climático	• cinturón	• murciélago	
• ernst	• cambio climático global	• chaparral	• picudo	
	• choque	• suelo alcalino	• quintana	

Anexo H: Lista de términos validados de matemáticas de primaria

La lista de términos validados presentada a continuación es la obtenida por el coeficiente de dominio con mejores resultados.

Para pesos de TF-IDF > 0 y CDImc

• ancho	• longitud	• proporción	• multiplicar	• oso
• antecesor	• líneas	• punto decimal	• división	• positivo
• arista	• matemática	• quebrado	• sistema de numeracion	• árbol
• aritmética	• mediana	• recta paralelo	• euros	• sureste
• billón	• minuendo	• rectangular	• dimensión	• aguja
• cara	• mitad	• rectángulo	• cilindro	• metodología
• característica	• máximo común divisor	• rectángulos	• unidad de tiempo	• nación
• cardinalidad	• mínimo común múltiplo	• redondear	• compás	• david douglas
• centena	• numeración romano	• regla y compás	• cuerda	• gamble
• centro	• numeros	• relación de orden	• simetría	• festivo
• cifra	• numérico	• resta	• cantidad	• entrada
• combinaciones	• número arábigo	• restar	• cuadrado	• pensar
• compacto	• número decimal	• resto	• quincena	• lateral
• completar el cuadrado	• número entero	• rombo	• habitante	• verano
• cubo	• número impar	• simplificación	• número de habitante	• necesario
• cuerpo geométrico	• número ordinal	• simplificación de fracción	• cálculo mental	• demostrar
• cuerpos geométricos	• número par	• sistema vigesimal	• habilidad matemático	• cumpleaños
• curva	• número primo	• submúltiplo	• ciudadana	• fiesta de cumpleaños
• cúbico	• números	• sucesor	• paréntesis	• conocimiento
• decena	• números ordinal	• sumando	• año	• especies
• decágono	• números ordinales	• sumar	• norte	• apariencia
• denominador	• números romanos	• sustraendo	• aleatorio	• fenómeno
• distancia	• números romanos	• tabla de multiplicar	• puntería	• diseñar
• diámetro	• octágono	• tablas de multiplicar	• identificación de figura	• diseño
• dodecaedro	• oponer	• trapecio	• emigrar	• fiesta
• dígito	• par	• triángulo equilátero	• lineal	• fiesta de disfraz
• elementos	• pentágono regular	• triángulo escaleno	• regla graduar	• juvenil
• entero	• perpendicular	• triángulo isósceles	• porcentaje	• viejo
• existir	• perímetro	• unidad	• lustro	• enumerar
• figura	• pi	• valor decimal	• mes	• calendario
• figura geométrico	• pirámide	• valor relativo	• perspectiva	• fecha
• figuras geométricas	• plano	• volumen	• razonar	• belleza
• fracción impropio	• poliedro	• vértice	• inferencia	• visual
• heptágono	• poliedros	• ángulo recto	• correlación	• ilusión
• hexagonal	• polígono regular	• punto	• posición	• dacca
• hexágono regular	• precisión	• coordenada	• frecuencia relativo	• accra
• infinito	• prisma	• matemático	• aniversario	• horizontal
• lado	• proporcionalidad	• multiplicación	• año escolar	• vertical
• lineas				• descubrir
				• experimento
				• respuesta

- descomposición
- azar
- principio
- rangel
- ciudad de mexico
- ciudad de méxico
- desfile
- derivar
- análisis
- bosque de chapultepec
- inventar
- metro cúbico
- entrenamiento
- lección
- gramo
- milla
- pie cuadrar
- tonelada largo
- tonelada métrico
- yarda
- coordenada cartesiano
- eje cartesiano
- busca
- fu
- ajusco
- delegación cuauhtémoc
- subir
- conversión de unidad
- tabla de conversión
- unidad de medida
- escuadra
- laboratorio
- par de base
- pedir
- kilogramo
- metro lineal
- acontecimiento
- acontecimientos
- ampliación
- octavo
- ejercicio
- adolescente
- centilitro
- centímetro cuadrar
- centímetro cúbico
- decagramo
- decalitra
- decímetro cuadrar
- decímetro cúbico
- hectolitro
- hectárea
- kilómetro cuadrar
- km2
- litro
- litros
- milla cuadrar
- milímetro cúbico
- pulgada cuadrar
- pulgadas
- unidad de superficie
- yarda cuadrar
- desigualdad
- probabilidad
- cuadro
- alambre
- plantillas
- adrián
- blas
- crispín
- marcelo
- medición
- triángulo rosa
- juegos olímpicos
- parlante
- iniciar
- información completo
- diagrama
- conclusión
- indagar
- investigación
- investigar
- sucesión numérico
- estima
- significado
- eje de simetría
- abecedario
- orden alfabético
- punto suspensivo
- derrota
- fluido
- honestidad
- zacate
- circuito
- foto
- fotografía
- fotográfico
- negativa
- diagrama de árbol
- letra griego
- teclear
- do
- barco
- iterar
- náutica
- engañosa
- rondana
- tornillos
- decámetro
- decímetro
- hectómetro
- material
- medalla de oro
- medalla de plata
- bucio
- crustáceo
- peletería
- caja registrador
- fuente de información
- edificio
- recreativo
- rectificación
- intercalar
- calculadora
- semilla
- lana
- imaginación
- percepción
- atención
- teléfonos celular
- candado
- guanábana
- accesorio
- engranar
- repuesto
- tuerca
- herramienta
- elevador
- escribir
- acceso
- menor de edad
- proceso
- cometer
- taller
- lancha
- cinta de medir
- flexómetro
- ladrillo
- ladrillos
- cursar
- experto
- inscripción
- nt
- oxford university
- panadería
- mecánico
- invitado
- invitar
- visita
- visitar
- cuadrante
- barco de vapor
- fabricante
- fábrica
- adecuación
- expresión oral
- tendencia
- extremo derecha
- arquitecto
- modo
- refrigerador
- clíper
- carreta
- minuto
- alquiler
- concreto
- fisura
- atar
- pata
- respaldo
- casete
- secundario
- bajar
- internet
- rafia
- cache
- vacío
- vacío
- cacahuete
- trébol
- básica
- muestra
- francisco toledo
- apogeo
- tarima
- escultura
- arcilla
- artesano
- orfebrería
- tapete
- parcela
- numeral
- colores
- chimenea
- fachada
- fachadas
- piso
- tabique
- petrolero
- encimar
- diccionario
- iniciado
- torneo
- basura
- basura orgánico
- noticiero
- mantel
- concurso
- libélula
- domo
- linear
- canica azul
- costumbre
- líquido
- minuterio
- báscula
- pesa
- simular
- creer
- sober
- delf
- fotocopidora
- fotocopiar
- hilos
- transatlántico
- bordado
- cable eléctrico
- greca
- moldura
- national
- disco compacto
- museum
- marisa
- libro
- libros
- torre latinoamericana
- linear azul
- línea azul
- cucharada
- renacuajo
- censo
- continuar
- fermat
- ida y vuelta
- pierre fermat
- salvador guerrero
- segar

- suma y resta
 - premio
 - hormiga
 - cartulina
 - sacapuntas
 - milla náutica
 - primar
 - protestar
 - ceremonia
 - gasolinera
 - cipriano
 - clasificación
 - changuito
 - umbral
 - cacerola
 - forma
 - descripción
 - fraseo
 - dinosauria
 - dinosauría
 - tanque de gas
 - camisa
 - chamarra
 - disfraz
 - gorra
 - pantalón
 - pantalón corto
 - suéter
 - tocar
 - uniforme
 - lodo
 - pisar
 - terminar
 - ropa
 - vestido
 - balneario
 - croquis
 - adentro
 - amarillo azul
 - aproximación
 - cuidado
 - decidir
 - exacto
 - hours
 - mandarina
 - obra completo
 - oír
 - raro
 - romper
 - tesoro
 - viajar
 - equivocar
 - cohecito
- libro de texto
 - libros de texto
 - pitagórico
 - calle
 - calles
 - calcetín
 - cinturón de seguridad
 - gorro
 - media
 - suela
 - terciopelo
 - azabache
 - imperio
 - pulsera
 - guante
 - colgante
 - canguro
 - rópodos
 - erar
 - verso
 - abreviar
 - abreviatura
 - cuento
 - cuento de hada
 - cuentos
 - auditorio
 - deportivas
 - conejos
 - rinoceronte
 - color naranja
 - forrar
 - oxford university press
 - dilo
 - figura rojo
 - regreso
 - tonatiuh
 - tonatúh
 - oración
 - baile
 - experiencia
 - mural
 - ardilla
 - chon
 - cochino
 - delfines
 - gato siamés
 - hipopótamo
 - oso hormiguero
 - venado
 - bermudas
 - morado
- eje
 - derecha
 - américa latina
 - latinoamericana
 - latinoamérica
 - bailar
 - carol
 - sida
 - color amarillo
 - color rojo
 - color rosa
 - color verde
 - rosado
 - verde
 - medir
 - programa de televisión
 - cima
 - dotar
 - huichol
 - huicholes
 - jacinto
 - fila
 - dromeosáuridos
 - feng shui
 - kiribati
 - antiguo egipto
 - egipto antiguo
 - aprender
 - aprendizaje
 - malvado
 - vaca
 - vacuno
 - bahamas
 - problema
 - gal
 - euro
 - garcía
 - cultura de egipto
 - egipcio
 - mesar
 - ventanal
 - gorrión
 - solear
 - variar
 - sombrilla
 - simple
 - palillo
 - jeroglífico
 - cuadrada
 - figuras
 - clip
 - aplauso
- meter
 - méto
 - completo
 - casa de muñeca
 - recortable
 - alacena
 - casillero
 - sillón
 - presentación
 - gaviota
 - bolita
 - canica
 - canicas
 - juguete
 - juguetes
 - rehilete
 - platicar
 - extremo izquierda
 - indo
 - papalote
 - cine
 - silla
 - bolsa de papel
 - canasta
 - canastilla
 - apodo
 - trinchador
 - dejar
 - muñeco
 - títere
 - anaranjado
 - castigo
 - curiosidad
 - mercancía
 - marinero
 - hilera
 - lata de sardina
 - kilo
 - actuar
 - frasco
 - numeración egipcio
 - trenza
 - vacunación
 - diagrama de barra
 - gráfica
 - gráfico de barra
 - sonreír
 - marca registrar
 - techo
 - sorprender
 - sorpresa
 - roser
- limpieza
 - obrero
 - medicamento
 - sexualidad
 - pitágoras
 - aceptación
 - análisis de sangre
 - aburrimiento
 - temor
 - participación ciudadana
 - paleta
 - cuarto
 - tío
 - glóbulo rojo
 - caber
 - corto
 - sarampión
 - berthá
 - autoestima
 - opción
 - fabla
 - tomate
 - época
 - elemental
 - medallas
 - punta
 - violencia
 - adicción
 - abuso
 - brinco
 - contreras
 - henar
 - leones
 - mejorar
 - montón
 - naval
 - science
 - socorro
 - shui
 - pelos
 - isbn
 - siamés
 - chite
 - círculos
 - centesimo
 - centésimo
 - moneda
 - monedas
 - lucir
 - miriam
 - mónica
 - paola

- | | | | |
|--|---|--|---|
| <ul style="list-style-type: none"> • rosalba • sandra • pirámide pentagonal • bárbara • guardar • quintar • tazón • yoyo • menor • símbolo • conocimiento previo • capitalista • feria • actividad deportivo • deportes • deportista • islam • islámico • andador • biberón • caminar • correr • velódromo • expresión • balón de fútbol • milpa • sector privar • torniquete • musulmán • sustitución • abono • asir • brazo • puño • alimento • bebidas • refresco • natación • raqueta • cambiar • cosecha • cosechar • helado • karate • karateca • payaso • estatura • festival • galleta • acetona | <ul style="list-style-type: none"> • almuerzo • carne • carne moler • comida • golosina • golosinas • mechar • ordenar • pastelero • quesos • gallina • pollito • plastilina • tenis • ciclismo • padre de familia • empacadora • viñedo • galería de arte • museo • ajedrez • tablero de ajedrez • chocolate • lechuga • tania • oeste • frijol • fríjol • espagueti • frutas • larga • lenteja • millo • pasar • pastelería • pistache • árbol frutal • tijera • tripa • carrera de relevo • postal • tarjeta postal • bretaña • puso • jugador • ismael • socializar • legumbre • arroz • guayaba • jitomate • miel • uva | <ul style="list-style-type: none"> • cancha • estadio olímpico • estadios • escocia • colgar • cuadrangular • triciclo • despertar • juicio • justificación • recién nacer • patinaje • remolque • maratón • papal • bota de oro • martina • equipo • dante • borrador • misael • ooo • fumador • juego • juegos • jugar • efraín • raquel • corazón rojo • gis • razonamiento • anar • convivio • árbol de navidad • u3 • cria • varita • ascendente • joel • pulido • colaboración • susto • continente americano • cincar • combinar • abuso de autoridad • dólar estadounidense • peso mexicano • comentario • plomo • entretener | <ul style="list-style-type: none"> • calceta • paso • revista • pecera • gusto • casita • abuelo • bisabuelo • sobrino • cúbito • producción • electoral • bonito • consentimiento • dragón • individual • libreta • primavera • celebrar • suerte • carrera de caballo • sardina • apuesta • juego de azar • indicador • lotería • revolución • trabajador • camp • asociación • asociación • idea • vacaciones • planta • precio • precios • tecla |
|--|---|--|---|