UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN LINGÜÍSTICA

AGRUPAMIENTO TEMÁTICO DE DOCUMENTOS BASADO EN UN MODELO DE SIMILITUD TEXTUAL

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN LINGÜÍSTICA HISPÁNICA

PRESENTA:
VÍCTOR GERMÁN MIJANGOS DE LA CRUZ

TUTOR:
DR. GERARDO EUGENIO SIERRA MARTÍNEZ
INSTITUTO DE INGENIERÍA, UNAM

MÉXICO, D. F. JUNIO 2015

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Esta tesis se llevó a cabo con el apoyo de la beca de maestría otrogada por el CONACyT en el período 2012-2014. Además, fue completada gracias al proyecto DGAPA IN400312 y al proyecto CONACyT 40832.

Agradezco a mi tutor y a mis sinodales por todo el apoyo a través del desarrollo de esta tesis. A mi familia, por estar a mi lado en esta etapa, y a mis familiares y amigos.

1

Introducción

Dada la gran cantidad de información que se produce hoy en día, se ha vuelto necesario manejarla de una forma eficiente para poder conseguir de ésta el mayor provecho posible. Sin embargo, ya que se trata de una cantidad enorme de documentos, se vuelve una tarea sumamente exhaustiva (si no es que imposible) para un ser humano. Por tanto, se hace necesario la utilización de herramientas automáticas que sean capaces de "entender" lo que se ha escrito en un texto y proporcionar a las personas la información que ellos necesitan: es precisamente en este punto donde la lingüística computacional ha venido jugando un papel de suma importancia en los últimos años.

Para manejar y obtener información de un conjunto de documentos, la categorización ayuda a estructurar la información y la hace accesible: la categorización tiene una gran importancia para la tarea humana, pues es la forma en que los seres humanos estructuran el mundo (Gentner and Ratterman, 1991; Goldstone, 1994; Tversky, 1977). De esta forma, la categorización automática se vuelve un proceso primordial dentro de la lingüística computacional. Dentro de esta área, se han venido desarrollando diferentes métodos que tratan de capturar la escencia del lenguaje, de tal forma que puedan manejar la temática de los textos y así detectar grados de similitud entre uno y otro documento. A esta última se le ha dado el nombre de *similitud textual*.

Actualmente, y todavía en muchos lugares, la tarea de categorizar documentos recae en manos humanas: los llamados *curadores*. La curación es una tarea que consiste en la revisión, clasificación y procesamiento manual de documentos: «Literature curation is a labour-intensive and time-consuming task, during which researchers extract relevant knowledge from a massive amount of

literature available in multiple repositories» (Sateli and Witte, 2012, p. 2).

La labor de los curadores no puede ajustarse al ritmo de la producción de material documental. Si bien los recursos informáticos han abordado dicha tarea, dado que la información se encuentra en lenguaje natural, es necesario no sólo manejar herramientas computacionales, sino tener un conocimiento basto del comportamiento del lenguaje en estos ámbitos.

La categorización es un problema que se puede entender desde la similitud: la igualdad o diferencia entre dos entidades determinará su pertenencia o no a un grupo determinado. Tratándose de categorización temática, además se tiene que tomar en cuenta la similitud "semántica"de los documentos, por lo que se requiere conocimiento lingüístico. Este tipo de conocimiento puede ser modelado para permitir un procesamiento computacional. Por tanto, el presente proyecto se enfoca en identificar las características temáticas que permitan determinar la similitud entre documentos y, de esta forma, poder categorizarlos.

En este trabajo se exploran varios métodos particularmente utilizados dentro de la literatura de la lingüística computacional y la similitud textual. Pero además se propone un método que explota los modelos ya explorados y que se basa en las ideas de Harris (1954) para representar las oraciones de un documento como vectores de ocurrencia en un contexto. Se parten de ideas generales sobre la similitud temática entre documentos, principalmente que:

1. Se puede cuantificar la información temática de un texto, tal que ésta pueda ser procesada por una computadora.

2. Existen palabras con mayor peso temático (llámense términos) que tendrán mayor influencia en la detección de características temáticas de un texto (Jones, 1972; Manning and Schutze, 2000; Manning et al., 2008; Wu et al., 2008).

3. La temática de un texto, en gran medida, determinará el uso terminológico que se haga dentro de éste y dispondrá un comportamiento distribucional (Harris, 1954).

Los puntos anteriores ya han tenido una amplia discusión dentro de la literatura de la lingüística computacional y han tratado de ser resueltos a partir de diferentes enfoques. El primer punto ha tenido diferentes aproximaciones tanto del ámbito de la lingüística computacional como de la computación pura. De esta forma, se ha tratado la similitud como un problema de distancia dentro de un plano o espacio vectorial, o bien, se han hecho aproximaciones que tratan de

precisar la naturaleza léxica e incluso semántica de los documentos. El segundo punto, igualmente, ha tenido una amplia exploración dentro de la lingüística computacional y otras áreas como la recuperación de información. Por último, el tercer punto enmarca una idea que se ha hecho popular dentro de la lingüística computacional y que radica en la hipótesis distribucional de Harris.

Con estas ideas en mente, a continuación se presentan los puntos generales sobre la tesis: se discute la problemática, el objetivo general y los objetivos particulares; se plantea también el ámbito de pertenencia de la tesis, es decir, el marco de la lingüística computacional en el cual se inserta el presente trabajo; y, finalmente, se presenta la distribución de la tesis.

1.1. Problemática

Se ha mencionado la gran cantidad de textos que se encuentran disponibles actualmente debido a la cantidad de medios de comunicación y difusión de información que han surgido. También se ha mencionado la cantidad de esfuerzo que para un humano sería manejar tanta información, y cómo este tipo de tarea se ha manejado por curadores. Ante tal problemática, se ha planteado como solución la automatización de dichos procesos. Sin embargo, la automatización presenta serios problemas que corresponde al ámbito de la lingüística computacional.

Se ha planteado ya que la problemática de similitud y de agrupamiento automático son parte de la lingüística computacional. El planteamiento del presente trabajo es integrar las dos problemáticas con un mismo fin, pues se pretende realizar un agrupamiento de documentos a partir de su similitud textual. De igual forma, se han planteado 3 ideas generales que implican la detección de similitud textual en documentos, que presentan en sí mismas un reto para la lingüística computacional. La detección de similitud textual entre dos documentos es un problema complicado incluso para un humano, pues se necesita el conocimiento del área que se está tratando para poder determinar con certeza si dos textos son similares o no. Por ejemplo, se pueden considerar los dos fragmentos siguientes:

> Normalmente vive en el intestino del hombre y de los animales y no suele causar ningún tipo de problema, es más, es necesaria para el funcionamiento correcto del proceso digestivo. Sin embargo, algunas cepas por intercambio de material genético, han adquirido la capacidad de causar infecciones y provocar diarreas sangrantes. Este tipo de Eschirichia coli es capaz de producir una potente toxina llamada Vero-citotossina (Vtec), sustancia que actúa como un veneno.
>
> Karmali en 1983, la asoció con casos aislados de síndrome urémico hemolítico (SUH) caracterizado por daño renal agudo, trombocitopenia y anemia hemolítica microangiopática, precedida por diarrea con sangre, con la presencia de heces de E. Coli productora de una citotoxina con actividad de células Vero, por lo que se le llama verotoxina (VT), y las cepas capaces de producirla se les denominó E. Coli verotoxigénicas (VTEC).

Se trata de dos fragmentos textuales que comparten temática, pero que sin embargo discursivamente pueden verse distintos. Para un humano inexperto en el área el mayor indicador de similitud podría ser el que ambos textos tienen cierto léxico en común, tal como *Eschirichia Coli* o *E. Coli* y *VTEC*. En un análisis un poco más profundo se puede notar que se trata de términos del documento y que estos términos precisamente son los más relevantes para ambos textos. Si, por ejemplo, se le diera la tarea a un humano de agrupar dichos textos y ponerles una etiqueta de clase, bien se podría elegir el término *Eschirichia coli* como etiqueta de la clase que agrupe los textos, pues precisamente es el tema que tratan ambos.

Si bien se sabe que los términos pueden ser un factor importante para la detección de características temáticas, queda abierto el problema de cómo detectarlos automáticamente y asignarles pesos de tal forma que los términos más relevantes sean cuantitativamente más representativos que aquellos menos importantes temáticamente; en la Sección 3.2.2 se hablará del asunto. Empero, la determinación de términos es sólo un paso dentro de la similitud temática. Se necesita determinar cuántos textos comparten un término y qué tanto valor cuantitativo representa que compartan tal término. Además estos valores deben ser procesados para obtener una *distancia* que indique qué tan cercanos están varios términos y poder así agruparlos dentro de grupos similares.

En un plano más básico, se necesitan definir con precisión el concepto de similitud y de agrupamiento. Ambos conceptos tienen un amplio panorama dentro de la lingüística computacional, por lo que será necesario estudiar los métodos que mejor se adopten a los objetivos de la tesis. De esta forma, la problemática que presenta la automatización del proceso de agrupación por similitud textual presenta un panorama complejo que se tratará de resolver a lo largo de la tesis.

1.2. Objetivos

El objetivo general de la tesis es proponer un método de agrupamiento de textos dada su temática, tomando en cuenta un modelo de similitud textual que capture las características temáticas de cada uno de los textos. Se busca centrar en el problema de similitud textual, área de la lingüística computacional que ha tenido diferentes perspectivas y aplicaciones, principalmente para el descubrimiento de relaciones léxicas. En este trabajo, empero, se busca aplicar las ideas de similitud textual para identificar características de los textos y poder, así, cuantificar una medida que sea capaz de ser interpretada por un algoritmo de agrupamiento para realizar una categorización en grupos textuales.

Para lograr este objetivo, primero se presentarán los panoramas generales de la categorización y agrupamiento, así como ideas sobre la similitud, con énfasis en la similitud entre elementos lingüísticos. Esto permitirá plantear diferentes concepciones sobre el lenguaje que den paso a representaciones formales que permitan identificar y cuantificar la similitud.

1.2.1. Objetivos particulares

A lo largo de este trabajo también se han planteado objetivos particulares, los cuales se presentan a continuación:

- Buscar comprender los elementos lingüísticos capaces de permitir la clasificación temática de los documentos.
- Abarcar diferentes modelos orientados a la similitud textual, que permitan un procesamiento computacional, y compararlos.
- Estudiar y discutir los diferentes métodos propuestos en la literatura para categorización y/o agrupamiento de documentos.

- Analizar la estructura de documentos escritos para determinar el mejor modelo de lenguaje que permita su procesamiento computacional.
- Describir la similitud entre dos textos distintos de tal forma que se pueda obtener un argumento cuantitativo de esta descripción.
- Formalizar tal descripción para que sea capaz de ser automatizada por un sistema computacional.
- Llevar a cabo un agrupamiento temático de textos en diferentes niveles (de general a especializado).

1.3. Marco de la tesis

Cabe resaltar que la presente tesis, como ya se ha venido señalando, aborda un tema específico dentro de la lingüística computacional. Por tanto, parece indispensable dar una breve introducción precisamente sobre la lingüística computacional. En primer lugar, se debe señalar que la lingüística computacional, como parte de la lingüística, integra recursos y métodos computacionales a la metodología lingüística. Para Payrató (1998) la lingüística computacional implica «las aplicaciones y aportaciones de la informática al análisis del lenguaje». Si bien no se puede negar que parte de la lingüística computacional sea traer al estudio del lenguaje las herramientas computacionales, un panorama que sólo describa así a la lingüística computacional será incompleto.

La lingüística computacional es también parte de la inteligencia artificial, y como tal busca entender y emular los procesos humanos que conlleva la producción lingüística. Desde este ámbito, se ha planteado la pregunta de si es posible interactuar con una máquina a partir de un lenguaje natural como el español. Se plantean Bolshakov and Gelbukh (2004, p.15) que «The great challenge of the problem of intelligent automatic text processing is to use unrestricted natural language to exchange information with a creature of a totally different nature: the computer».

En términos menos generales, la lingüística computacional ha cobrado un auge importante en la era de la información, pues a ésta le corresponde el manejo de las grandes cantidades de datos que se producen en lenguaje natural. Se ha visto a la lingüística computacional como una ciencia capaz de generar conocimiento y aplicaciones útiles para las tareas más exhaustivas de la vida

cotidiana, que van desde determinar el correo electrónico no deseado, la traducción automática a diferentes idiomas o el mismo agrupamiento temático de datos textuales, hasta aquellos problemas más especializados, como la detección de categorías gramaticales, detección de lindes morfológicos, determinación de estructuras sintácticas, generación de gramáticas formales o la producción automática de lengua natural. Sin duda, muchas de estas áreas requieren todavía de mucho trabajo que realizar, pero se han hecho múltiples esfuerzos por obtener recursos computacionales que puedan llevar a cabo estas tareas.

La lingüística computacional, como área de la lingüística, se encuentra con la intersección de otras subáreas de la lingüística, como se puede observar en la Figura 1.1. Principalmente resaltan el área de la lingüística aplicada, la psicolingüística y la lingüística matemática. La lingüística aplicada porque la lingüística computacional se ha visto como una forma de crear herramientas y aplicaciones que sean de utilidad tanto para la lingüística como para otras áreas. La psicolingüística también juega un papel importante, pues parte de la lingüística computacional, como ya se ha señalado, se interesa por los procesos cognitivos que permitirían emular el lenguaje humano. De igual forma, un área que es de primordial importancia y que ha dado grandes aportes a la lingüística computacional es la lingüística matemática, sobre esta se puede señalar que:

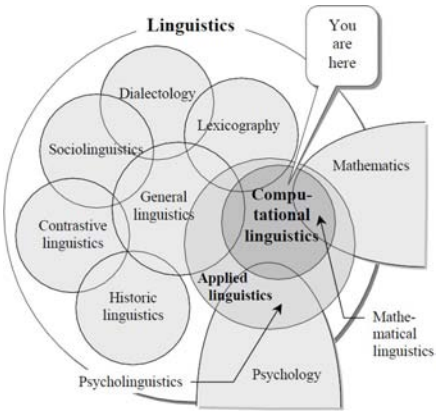

Figura 1.1: Estructura de la ciencia lingüística (Bolshakov and Gelbukh, 2004).

> There are two different views on mathematical linguistics. In the narrower view, the term *mathematical linguistics* is used for the theory of formal grammars of specific type referred to as *generative grammars*. [...] Alternatively, in the broader view, mathematical

> linguistics is the intersection between linguistics and mathematics, i.e., the part of mathematics that takes linguistic phenomena and the relationships between them as the objects of its possible applications and interpretations (Bolshakov and Gelbukh, 2004, p. 23).

Dentro de la lingüística matemática se inserta también la *lingüística estadística*, que estudia el lenguaje a partir de diferentes rasgos cuantitativos, como la frecuencia de las palabras, sus combinaciones y construcciones en los textos. También se le puede llamar a esta área de la lingüística interesada en datos cuantitativos *lingüística cuantitativa*.

De esta forma, convergen varias áreas de la lingüística (de las cuales no se han mencionado todas) dentro de la lingüística computacional. Por tanto, la lingüística computacional puede verse como la intersección de todas estas áreas para generar herramientas que sean capaces de manejar, procesar y generar lenguaje natural. Con esto en mano, cabe señalar los problemas a los que se enfrenta la lingüística computacional. Para Dougherty (1994) los problemas de la lingüística computacional pueden separarse en dos grandes grupos: i) *problemas conceptuales*, y ii) *problemas técnicos*. Los primeros se tratan de problemas más teóricos que involucran la comprensión del lenguaje para generar procesos o reglas formales que sean computables; se cuestiona cosas como la naturaleza del lenguaje, qué es el significado, cómo se adquiere el lenguaje y cuestiones que involucran la naturaleza conceptual del lenguaje para que permitan emularlo por una máquina. El segundo grupo de problemas es quizá el más explotado dentro del área y se refiere a cuestiones más prácticas que involucran la representación del lenguaje dentro de una computadora, su codificación y decodificación, su estructura y la identificación de ésta (detección automática de morfología y sintaxis), entre otras cuestiones que atañen al procesamiento automático del lenguaje.

Dentro de esta categoría de problemas técnicos se pueden insertar aquellos que han sido abordados dentro de la literatura de la lingüística computacional, y que Gelbukh and Sidorov (2006) clasifica en: ayuda a preparación de texto, como corrección ortográfica, detección de características gramaticales y de estilo; búsqueda de información, que implica la ayuda al humano de buscar información precisa en grandes cantidades de datos; búsqueda de documentos, que podría verse como una especificación de la anterior orientado a la información lingüística codificada en texto; gestión inteligente de documentos, que implica el manejo de textos de forma "inteligente"; interfaces humano-máquina;

traducción automática; generación de lenguaje; y otras aplicaciones.

En un panorama general, lo que se busca dado este tipo de problemas es generar sistemas computacionales que sean capaces de realizar tareas de lenguaje natural de la misma forma en que las realizaría un humano. Esto, empero, resulta ser una tarea complicada. Un esquema general, resumido por Gelbukh and Sidorov (2006), puede ayudar a plantear la dificultad de este tipo de hazañas. Así, la creación de un sistema de lingüística computacional implicaría:

1. La transformación del texto en una representación formal que preserve las características relevantes para una tarea dada.

2. Manipular esta representación, transformándola según la tarea y buscando en ella subestructuras necesarias al objetivo específico.

3. La transformación, si es necesario, de los cambios hechos a la representación formal para que ésta pueda ser devuelta por el sistema en lenguaje natural.

El primer punto que se toca aquí representa ya una necesidad de conocimiento lingüístico que permita ser capaz de detectar cuáles son las estructuras y/o patrones específicos que determinan un comportamiento dado del lenguaje. Por ejemplo, para los propósitos de esta tesis, se necesita saber en qué parte del lenguaje se puede recuperar información temática. Además de dicho conocimiento, se necesita ser capaz de representar la información en una estructura formal que sea computable. El segundo punto requiere ser capaz de abstraer determinados procesos lingüísticos para indicar a la computadora cómo manipular la información que ya se le ha dado con el primer paso. Por último, dado que se ha procesado ya el lenguaje y se ha obtenido un resultado a partir de las herramientas computacionales, se debe ser capaz de interpretar este resultado (sea dado cuantitativamente o de alguna otra forma) para poder traducirlo a lenguaje natural. Cada uno de estos procesos presenta su grado de dificultad y son requeridos para la creación de un sistema dentro de la lingüística computacional. A través de la tesis desarrollaremos estos puntos con el caso específico que aquí se plantea.

De esta manera, se ha visto que la lingüística computacional requiere de varios ámbitos de la lingüística y que, además, se hace acompañar por conceptos provenientes de áreas como la computación o incluso la matemática. La similitud textual ha sido un problema presente dentro de la lingüística computacional, ya que muchas aplicaciones lo requieren. Por mencionar algunas se puede traer

a escena la creación de herramientas para la detección de plagio, la creación de taxonomías, el descubrimiento de relaciones léxicas, la desambiguación de estructuras lingüísticas semánticamente ambiguas y el agrupamiento de documentos, que es el tema a tratar aquí. Éste se ha insertado dentro de lo que se ha llamado *agrupamiento de documentos* y es parte de la gestión inteligente de documentos, la búsqueda de documentos y la búsqueda y recuperación de información.

1.4. Distribución de la tesis

Para los objetivos que se plantea la tesis, se configuró la siguiente distribución, tratando de abarcar los temas más generales al principio, para así poder adentrarse en la aplicación de diferentes modelos del lenguaje para una aplicación en lingüística computacional:

Agrupamiento de documentos. En este capítulo se hará, primero, una breve introducción sobre el problema de categorización, se verán sus aplicaciones, y se adoptará una posición que contrasta este término con el de agrupamiento. A este último, se le enmarcara dentro del concepto de aprendizaje de máquina. Se expondrá cómo se entiende el concepto de agrupamiento dentro del aprendizaje de máquina. Se expondrán, asimismo, la clasificación de los algoritmos de agrupamiento y se presentarán algunas implementaciones.

Similitud textual. En esta sección se adentrará al concepto de similitud y, más específicamente, al de similitud textual dentro de la lingüística computacional. Se expondrán diferentes métodos para detectar la similitud textual que se han expuesto en la literatura (métricas, taxonomías, información y espacios semánticos) y se discutirán sus ventajas y desventajas.

Metodología. Aquí se presentarán la metodología adoptada y se expondrán las características del corpus utilizado. En este capítulo se presentarán los detalles técnicos de los experimentos. Se expondrá el pre-procesamiento de los documentos que constó de la validación de los términos y la aplicación de los algoritmos de *stemming*. Se presentará la estructura del sistema computacional desarrollado.

Análisis y evaluación. Aquí se discutará la implementación de los métodos y el algoritmo propuesto a un corpus de evaluación. Además, se expondrán

los métodos utilizados para la evaluación y se discutirán los resultados obtenidos.

Conclusiones. Finalmente, se hará un resumen de la tesis y se presentará una discusión de la metodología y los resultados obtenidos para valorar sus aportaciones y sus carencias.

Agrupamiento de documentos
En este capítulo se trata el conpcepto de agrupamiento. Se hará un contraste entre lo que aquí se toma por categorización. Se hará, asimismo, una breve introducción de las propuestas del aprendizaje de máquina, marco dentro del cual se toma el término de agrupamiento. De igual forma, se aborda el problema de la representación de datos textuales dentro de este mismo panorama. Se presentará la definición de agrupamiento como parte del aprendizaje de máquina, su hipótesis y su problemática. Asimismo, se expondrá una clasificación de los diferentes métodos de agrupamiento, y algunos algoritmos que se presentan en la literatura y se mostrarán sus ventajas y desventajas. Finalmente, se explicará el algoritmo de MajorClust que, dada sus características, fue el que se ha utilizado para los experimentos mostrados en este trabajo.

1.5. Categorización y agrupamiento

Para la labor humana, una actividad relevante para el desempeño de diversas tareas es la categorización de elementos similares dentro de grupos; para el ser humano (sobre todo en sus etapas más tempranas), este proceso de categorización implica una parte importante del proceso de aprendizaje (Kaufman and Rousseeuw, 2005). Por ejemplo, un niño aprende a distinguir entre diferentes tipos de animales, como gatos y perros, así también distingue entre hombres y mujeres, y entre muchas otras cosas; de esta forma, crea un esquema mental que categoriza al mundo dentro de grupos de elementos similares (Goldstone, 1994; Kaufman and Rousseeuw, 2005).

Es imposible pasar por alto que el proceso de categorización implica así mismo el concepto de similitud. Según señala Goldstone (1994) «categorization behaviour depends on similarity between the item to be categorized and the categories' representation» (Goldstone, 1994, p. 127). Este mismo autor apunta dos maneras en que la categorización puede hacerse dado dos grupos distintos A y B. Señala que i) un objeto es categorizado en A y no en B si es más similar

a la mejor representación de A (su prototipo[1]) que a la de B; y ii) un objeto es categorizado como A y no como B si es más similar a los elementos individuales que integran A que a aquellos que integran B.

Dentro de la ciencia, la necesidad por categorizar los elementos de estudio ha sido una preocupación constante. Para un biólogo, por ejemplo, es importante encontrar grupos que categoricen a animales o insectos, lo que permite estudiarlos con mayor precisión. La necesidad por la categorización de objetos de estudio ha hecho que desde tiempo atrás se haya practicado el agrupamiento. Señala Kaufman and Rousseeuw (2005) que «In the past, clustering was usually performed in a subjective way, by relying on the perception and judgement of the researcher» (Kaufman and Rousseeuw, 2005, p. 3); esto, sin embargo, ha venido cambiando a través del tiempo y los procesos de categorización cada vez dependen menos de la subjetividad del ser humano y se han creado métodos automáticos para llevar a cabo esta tarea.

Kaufman and Rousseeuw (2005) señala que «Cluster analysis is the art of finding groups in data» (Kaufman and Rousseeuw, 2005, p. 1); el término de *agrupamiento* (o *clustering* en inglés), dada las tendencias de automatización, se ha utilizado de manera específica por varios autores como un modelo de aprendizaje de máquina. De esta forma, señala Manning et al. (2008) que «Clustering is the most common form of *unsupervised learning*. No supervision means that there is no human expert who assigned documents to classes». Por su parte, Segaran (2007) apunta que «Clustering is an example of *unsupervised learning*» y Marmanis and Babenko (2009) que «clustering belongs in the category of machine learning known as *unsupervised learning*»[2].

Por tanto, en este trabajo se ha optado por referir como *agrupamiento* al modelo de categorización perteneciente al aprendizaje de máquina; en adelante cuando se refiera al término de *agrupamiento* se entenderá como parte del aprendizaje de máquina. En cualquier otro caso se utilizará *categorización*. Dada esta perspectiva, se expondrán los conceptos de *aprendizaje de máquina*, *aprendizaje supervisado* y *no supervisado*, y se ahondará en el concepto de agrupamiento dentro del marco del aprendizaje de máquina. Pero antes se analizarán las aplicaciones de la categorización y, por ende, del agrupamiento.

[1] Aquí se entiende prototipo como el elemento cuyos rasgos mejor representan a una clase.

[2] Otros autores, como Bijuraj (2013), utilizan el término *clustering* o *agrupamiento* independiente del aprendizaje de máquina, de forma similar a lo que aquí se toma como *categorización*.

1.5.1. Aplicaciones

Dada la necesidad de categorización del ser humano, los modelos de categorización han facilitado en gran medida tareas en diferentes áreas. Dentro de la minería de datos, por ejemplo, la categorización es uno de los primeros pasos, pues es relevante para la organización y el procesamiento de datos. Asimismo, los motores de búsqueda que trabajan con datos textuales requieren de los métodos de categorización. Bijuraj (2013) señala como aplicaciones de estos métodos las siguientes:

- Minería de datos
- Reconocimiento de patrones
- Análisis de imágenes
- Bioinformática
- Minería de voz
- Procesamiento de imágenes
- Motores de búsqueda web
- Minería de texto
- Aprendizaje de máquina

Aquí se pone especial atención en la minería de texto. Para ésta también se hace importante el proceso de agrupamiento; el texto necesita categorizarse para poder obtener de éste información relevante (Bijuraj, 2013). Un resumen breve de sus aplicaciones en la minería de texto se muestra en la Tabla 1.1.

Cabe aclarar que el mismo Bijuraj (2013) propone como aplicación de la categorización el aprendizaje de máquina. Sin embargo, el aprendizaje de máquina en esta tesis se ve más como un método de categorización que como una aplicación, puesto que por medio del aprendizaje de máquina, como se verá a continuación, se puede realizar categorización de elementos a partir de los algoritmos no-supervisados.

Tabla 1.1: Aplicaciones de los algoritmos de agrupamiento (Manning et al., 2008, p. 351).

Aplicaciones	**Qué se agrupa**	**Beneficios**
Agrupamiento de resultados de búsqueda	Resultados de una búsqueda	Presentación al usuario de información más efectiva
Scatter-Gather	Subconjunto de una colección	Interfaz alternativa de usuario: "Buscar sin teclear"
Agrupamiento-de-colección	Una colección de documentos	Presentación de información efectiva para búsqueda exploratoria
Modelado de lenguaje	Una colección de documentos	Incremento de la precisión y/o la exhaustividad
Recuperación basada en grupos	Una colección de documentos	Mayor eficiencia: Búsquedas más rápidas

1.6. Aprendizaje de máquina

En el área de la lingüística computacional, los métodos de *aprendizaje de máquina* están siendo utilizados en una amplia gama de tareas, como el etiquetado de categorías gramaticales, reconocimiento de opiniones, clasificación de emociones y otras tantas aplicaciones en las que resalta el agrupamiento de documentos. La ventaja que estos presentan ante otros modelos de procesamiento del lenguaje es la capacidad que tienen de inducir características del lenguaje que de otra forma comportarían una tarea sumamente exhaustiva. Por ejemplo, en una tarea de clasificación de emociones, sin un sistema de aprendizaje, se requerirían de diccionarios que fueran capaces de abarcar por lo menos la mayoría del léxico que determina un sentimiento en documentos textuales; un sistema de aprendizaje induce a partir de un corpus (o conjunto de datos) este léxico y puede asignarle pesos al mismo. Otro problema que se ve beneficiado por los sistemas de aprendizaje de máquina es cuando la tarea que se aborda suele cambiar con determinadas variables; un ejemplo claro es la detección de *spam*, pues éste puede ser percibido de forma distinta por una persona que por otra que puede considerar *spam* lo que la primera persona no. De esta forma, un sistema de aprendizaje de máquina «aprende» del comportamiento de cada uno de los individuos y así determina lo que para cada una es *spam*.

El concepto de *aprendizaje de máquina* se basa en la idea de emular los procesos de aprendizaje del ser humano a partir de un programa de computadora. De esta forma, se han propuesto diversos métodos que traten de cumplir esta

tarea. Como punto de partida, se debe tener en cuenta qué se entiende por *aprender*. Una definición general es dada por Michalsky et al. (1986) citando a Minsky, quien entiende al aprendizaje como un proceso que realiza cambios útiles en nuestras mentes. Esto ya deja ver que el aprendizaje es una forma en que se optimiza una tarea a partir de adaptación a los cambios o la experiencia. Una definición más específica puede ser:

> Learning denotes changes in the system that are adaptative in the sense that they enable the system to do the same task or tasks drawn from the same population more effectively the next time. (Michalsky et al., 1986, p. 10)

Está es una definición más amplia, en donde el sistema puede verse como un sistema biológico, es decir, el cerebro humano, o uno artificial, es decir, una computadora. Con precisión a este último aspecto, una definición sobre el aprendizaje de máquina que después han adoptado diversos autores es la dada por Mitchell (1977), quien menciona que:

> Un programa de computadora se dice que **aprende** a partir de experiencia E, con respecto a una tarea T y una medida de desempeño P, si su desempeño en la tarea T, medida por P, mejora con la experiencia E.

Esta primera definición de aprendizaje de máquina pone en la mesa tres elementos básicos que debe tener todo sistema de aprendizaje: experiencia (E), una tarea (T) y una medida de desempeño (P). Los tres elementos son variables y van a depender en gran medida de T, es decir, de la tarea que se requiera realizar. Cambronero and Moreno (2006) determinan, igualmente, otros tres conceptos básicos de un sistema de aprendizaje de máquina: el **conjunto de datos** (o en el caso de la lingüística computacional, un corpus), en el cual se distinguen el *conjunto de entrenamiento* y el *conjunto de prueba*; en general, se puede ver al conjunto de entrenamiento como la experiencia, E. El segundo elemento es el **modelo**, que, según lo definen los autores, es una conexión entre las variables dadas y las que se predicen por el sistema; éste se puede ver como un modelo matemático que determina las condiciones que permitirán hacer una predicción a partir del conjunto de datos. Por último, se habla del **aprendiz**, que lo definen los autores como «cualquier procedimiento utilizado para construir un modelo a partir del conjunto de datos de entrenamiento»; es un paso posterior

al modelo, pues a partir del modelo es que se obtiene el aprendiz, el cual se puede ver como una serie de reglas determinadas por el modelo que permitirán la predicción por parte del sistema.

Cabe también señalar que dentro del conjunto de datos existen dos aproximaciones generales (y se puede hablar de una tercera) en la forma de abordar el aprendizaje de máquina. Éstas son el aprendizaje supervisado y el no supervisado. A continuación se abordarán las nociones básicas de ambas perspectivas, para posteriormente enfocarse en el aprendizaje no supervisado que es el que interesa para los fines específicos de esta tesis.

1.6.1. Aprendizaje supervisado y no-supervisado

Dentro del *aprendizaje de máquina* destacan dos vertientes que varían en cuanto al tipo de información que le proporcionamos a la computadora. Estas vertientes son el *aprendizaje supervisado* y el *aprendizaje no-supervisado*. La principal característica que los diferencia es que el primero requiere de un conjunto etiquetado de entrenamiento; es decir, se necesita dar a la computadora información previa acerca de la tarea que va a realizar y que la computadora aprenda a distinguir los ejemplos a partir de clases determinadas por un experto. En general, se puede entender este tipo de métodos como «Techniques that use inputs and outputs to learn how to make predictions» (Segaran, 2007, p. 29). Así, para un sistema de *aprendizaje supervisado* se necesita de un corpus etiquetado manualmente por un humano. En tal corpus se debe indicar la clase de cada uno de los ejemplos; de esta forma, la computadora aprenderá los rasgos propios de cada clase y podrá decidir, cuando se le dé un objeto no clasificado, a qué categoría pertenece. Formalmente, el conjunto de datos S que se requieren para el aprendizaje supervisado se puede representar como:

$$S = \{(x_i, y_i) | x_i \in \mathbb{R}^d, y_i \in Y, i = 1, ..., m\}$$

Donde $\mathbb{R}^d$ es un espacio vectorial que en este caso está formado por los d ejemplos del corpus de entrenamiento (x_i), mientras que Y es el conjunto de etiquetas (y_i) asociadas a cada uno de los ejemplos. De esta forma, se nota que cada uno de los ejemplos del corpus tiene que tener una etiqueta asociada que lo describa.

Por su parte, el *aprendizaje no-supervisado* no tiene etiquetas asociadas a los ejemplos del corpus, sino que la computadora agrupa los ejemplos a partir de las características de cada uno de ellos; para Segaran (2007): «unsupervised

learning algorithms are not trained with examples of correct answers. Their purpose is to find structure within a set of data where no one piece of data is the answer» (Segaran, 2007, p. 30). Por tanto, a diferencia del aprendizaje supervisado el conjunto de datos $\mathcal{U}$ que requieren los algoritmos de este tipo no contempla un conjunto de etiquetas:

$$\mathcal{U} = \{u_i | u_i \in \mathbb{R}^d, i = 1, ..., M\}$$

En este caso, no se requiere de un experto humano que determine las etiquetas de las clases en que se deben agrupar las muestras. Esto, si bien implica la eliminación de la exhaustiva tarea del etiquetado, implica también mayor dificultad para el sistema pues no se tiene ninguna información previa. Precisamente, este tipo de métodos se utiliza cuando se desconocen las clases en las que deben agruparse las muestras. Dentro de la categoría de *no-supervisado*, por tanto, es donde se encuentra el *agrupamiento*.

Un tercer método dentro de los algoritmos de aprendizaje de máquina es el que se ha dado en llamar *aprendizaje parcialmente supervisado*, que es la conjunción de los métodos supervisado y no-supervisado, tal que se puede definir como $T = S \cup \mathcal{U}$ (Friedhelm and Trentin, 2014). En este caso, el conjunto de datos se puede formalizar de la siguiente forma:

$$S = \{(x_i, y_i) | x_i \in \mathbb{R}^d, y_i \in \Delta^L, i = q, ..., m\}$$

Donde $\Delta^L = \{y \in [0, 1] | \sum_{j=1}^{L} y_j = 1\}$ es un conjunto que contiene el equivalente a las etiquetas en el método supervisado, pero con la diferencia de que en este caso las etiquetas son «difusas» y toman valores del tipo $\frac{1}{L}$, donde L es el número total de grupos.

Aquí, se presta especial atención a los algoritmos no-supervisados, especialmente a los algoritmos de *agrupamiento*, pues son estos los que se aplican al caso del agrupamiento temático, en donde las clases de los documentos son variables y no se conocen *a priori*.

1.7. Representación de datos textuales

La representación de los datos textuales es uno de los factores claves para la aplicación y procesamiento de texto por un sistema computacional. La forma en que estos textos se representan para que sean manejables por la computadora

implica diferentes perspectivas, pero sobre todo requiere que los datos sean presentados en un espacio vectorial. En este apartado se presentan las formas en que pueden ser representados los datos textuales y se define el concepto de *espacio vectorial*, que posteriormente será ampliado en lo que se ha dado en llamar «espacio semántico».

En un panorama general, Metzler et al. (2007) propone tres formas de representación de texto:

- **Representación superficial:** No existe ninguna manipulación manual o automática de los datos textuales, sino que estos aparecen tal como son; es decir, no se aplican algoritmos de *stemming* (véase la Sección 3.2.3) que den las raíces de las palabras, ni transformaciones de otro tipo.
- **Representación lematizada:** Las palabras son llevadas a su forma de diccionario o se cortan para conservar las bases, de tal forma que el texto sea «normalizado». Con ello, se requiere de un mayor esfuerzo que en la representación superficial, pero los resultados suelen ser mejores.
- **Representación expandida:** En este tipo de representación se hace una «expansión» del texto; es decir, se incluye información contextual relevante al texto, tal como indicaciones de título, autor, información semántica o pragmática, etc. Sin embargo, dicha expansión requiere de recursos externos que aporten este tipo de información (como puede ser Wikipedia, diccionarios electrónicos, tesauros), lo cual lo hace más costoso y requiere de un mayor esfuerzo.

Este tipo de representaciones se basan en cómo se manejarán las palabras que conforman cada uno de los textos. Además, como señala Segaran (2007), la forma común de preparar los datos para un proceso de aprendizaje es determinar un conjunto común de atributos que puedan expresarse cuantitativamente. Por tanto, un paso posterior es la representación del espacio vectorial, pues en este espacio se tomará en cuenta si el documento tiene una representación superficial, lematizada o expandida. Este tipo de representación suele afectar los resultados finales del algoritmo. Debe tomarse en cuenta que una representación superficial es la más sencilla, sin embargo suele tener ciertos problemas, mientras que la representación expandida requiere de recursos que pueden ser difíciles de conseguir, además de un trabajo previo exhaustivo que no garantiza que los datos de expansión hagan el sistema eficiente. La representación lematizada es la más difundida y suele obtener buenos resultados. En la Sección 3.2.2

se discuten la representación superficial y lematizada aplicadas a los objetivos de este corpus.

El siguiente paso será la representación en un modelo de espacio vectorial, por lo que aquí se define este concepto y se discuten las nociones básicas. Una definición muy básica es dada por Howland and Park (2003), quienes dicen que el modelo de espacio vectorial «represents documents as vectors in a vector space» (Howland and Park, 2003, p. 3). En forma más explícita, lo que el modelo de espacio vectorial se propone es representar conceptualmente un documento a partir de un vector de palabras claves extraídas del documento y asociadas con pesos que representen la importancia de estas palabras clave (Lee et al., 1997).

Si bien la asociación de un peso a las palabras es un concepto importante, como muestran Lee et al. (1997) la asociación de las palabras se puede hacer simplemente con las frecuencias absolutas de aparición de un término en el documento. Sin embargo, como exponen estos mismos autores, la implementación de algoritmo de pesado más elaborado puede llevar a mejorar el desempeño del sistema. Uno de los métodos más comunes para asignar pesos es *tf-idf*, que se discute en la Sección 3.2.2.

Una vez que se han asignado los pesos, la representación en un espacio vectorial se hace por medio de una matriz $\mathbb{X}$ de $m \times n$, donde cada columna representa un documento y cada entrada $\mathbb{X}_{i,j}$ muestra el peso de un término i en un documento j (Howland and Park, 2003). Se puede decir que cada columna es asimismo un vector que representa un documento. Este vector tiene ciertas características que se explican en Raghavan and Wong (1986):

1. La dimensión del espacio vectorial es n, donde n es el número de términos distintos.

2. Los vectores se emparejan ortogonalmente [3].

3. d_{ij} es el componente del documento j sobre la dirección del vector de términos i.

Estos tres puntos resaltan las características que otorgan al texto su representación en un espacio vectorial. Sea $t_1, t_2, ..., t_n$ los términos que representan un documento; se asume que t_i son vectores de longitud de una unidad; luego se toma un documento d_j, donde $1 \leq j \leq m$, como un vector expresado en

[3] Se dice que son ortogonales cuando cumplen la propiedad $\mathbb{X} \cdot \mathbb{X}^T = I$ donde I es la matriz de identidad; es decir, una matriz de ceros con unos en la diagonal.

términos de t_i. Así $d_j = \{a_{1j}, a_{2j}, ..., a_{nj}\}$ donde $a_{ij} \in \mathbb{R}$, es decir, se trata de números reales que reflejan la importancia del término i en el documento d_j.

Ligado al espacio vectorial, podemos hablar de *espacio euclidiano*. Supóngase un espacio vectorial, V, con el *producto escalar* $x \cdot y$ de dos vectores $x, y \in V$, las magnitudes $||x||, ||y||$ y $\cos\theta$, donde $||x||$ y $||y||$ son las longitudes de los dos vectores y θ es el ángulo entre x y y. Un espacio vectorial con un producto escalar es llamado *espacio euclidiano*. Este concepto es importante, pues la similitud entre dos vectores de documentos depende en gran medida de las operaciones que se puedan hacer dentro del espacio euclidiano; así, una medida común de similitud es la distancia euclidiana, pero en esta tesis también se proponen medidas de similitud que hacen uso de los productos escalares o productos internos. Estos conceptos se expondrán más adelante.

1.8. Algoritmos de agrupamiento en aprendizaje de máquina

Se ha mencionado ya que para este trabajo *agrupamiento* es un proceso de categorización que corresponde al *aprendizaje no supervisado*; como tal, no requiere un conjunto de entrenamiento al que un humano le haya asociado clases. Se puede decir que:

> Clustering is probably the most important task in the unsupervised learning scenario. Here the training data are given as a finite data set $\mathcal{U}$ withouth any additional prior information. Then the major goal of clustering is to group the data vectors into k steps, with $1 < k < M$. Data clustering relies some kind of similarity (or distance) measure, in combination with an overall objective function (Friedhelm and Trentin, 2014, p. 8).

Un algoritmo de agrupamiento categoriza un conjunto de elementos dentro de grupos; se busca que los elementos dentro de un grupo tengan características que los identifiquen y, al mismo tiempo, los diferencien de los elementos dentro de otros grupos. En otros términos, un *algoritmo de agrupamiento* busca agrupar documentos similares. Manning et al. (2008) definen la hipótesis del grupo de la manera siguiente:

> **Cluster Hypothesis:** Documents in the same cluster behave similarly with respect to relevance to information needs.

Queda claro que para el aprendizaje de máquina un proceso de entrenamiento será una tarea en la que no se conocen clases, y en donde se busca agrupar elementos similares dentro de un mismo grupo (*similitud intraclase*), de tal forma que éstos contrasten con los elementos de diferentes grupos (*disimilitud interclase*). De nuevo, y como señala Friedhelm and Trentin (2014) se hace relevante el concepto de similitud.

Ante la búsqueda de grupos similares, uno de los puntos claves del agrupamiento es identificar una medida que determine la identidad entre un documento y otro. En Manning et al. (2008) se dice que «The key input to a clustering algorithm is the distance measure. [...] Different distance measures give to rise to different clustering. Thus, the distance measure is an important means by which we can influence the outcome of clustering» (Manning et al., 2008, p. 350). La distancia a la que se hace referencia no es otra cosa que una medida cuantitativa de la similitud entre un documento A y un documento B. Las distancias utilizadas para los algoritmos de agrupamiento se repasarán en la Sección 2.2.

1.8.1. Problemática del agrupamiento

Para entender la problemática del agrupamiento, cabe señalar los elementos con que se cuenta en un proceso de agrupamiento, estos son:

1. Un conjunto de documentos, $D = \{d_1, d_2, ..., d_n\}$.
2. Un número deseado de grupos, k, que puede estar no definido *a priori*.
3. Una función objetiva, que evalúe la calidad de un proceso de agrupamiento.
4. Una función de asignación γ, tal que $\gamma : D \to \{1, ..., k\}$.

La problemática del agrupamiento, así, radica en que dados D y k se minimice la función de asignación a partir de γ. En otras palabras, se busca que los grupos dados por γ sean coherentes en cuanto a la similitud de los elementos que los constituyen, ya que «[t]he objective function is often defined in terms of similarity or distance between documents» (Manning et al., 2008, p. 354). Desde tal perspectiva, lo que se pretende en esta tesis es encontrar una medida de similitud que sea capaz de cumplir lo mejor posible estos requerimientos.

1.9. Clasificación de algoritmos de agrupamiento

Los algoritmos de agrupamiento se pueden clasificar de diferentes formas. En cuanto al modo en que se realiza el agrupamiento se tienen diferentes aproximaciones. En Manning et al. (2008) se reconocen dos grandes grupos: *agrupamiento plano* y *agrupamiento jerárquico*; por su parte Steinbach et al. (2000) reconoce el *agrupamiento jerárquico* y el *agrupamiento particional*. Los términos de «plano» y «particional» se utilizan para describir el mismo fenómeno, por lo que aquí se tratan en un sólo apartado. Además, para la presente tesis se incluye el algoritmo de MajorClust, que no se describe por ninguno de los autores anteriores y que tiene características específicas, por lo que se considerará aparte.

1.9.1. Agrupamiento jerárquico

Este tipo de algoritmos devuelve un agrupamiento en forma de una estructura jerárquica. En esta estructura, todos los grupos dependen de un grupo mayor que se encuentra en la parte superior de la jerarquía. Este tipo de algoritmos se basan en grupos binarios, pues cada nivel inferior al superior se puede ver como un agrupamiento de dos grupos o documentos (en el nivel más bajo). Los algoritmos jerárquicos se pueden visualizar como una especie de árbol, llamado *dendograma*.

En el dendograma se representan las conexiones que se dan entre los diferentes elementos que se encuentran en cada grupo; además muestra la distancia que existe entre cada elemento y grupo de elementos. En general, los algoritmos jerárquicos se dividen, asimismo, en *algoritmos aglomerativos* y *algoritmos divisivos*.

Agrupamiento aglomerativo

Este tipo de algoritmo, en forma general, comienza con grupos individuales; en los pasos posteriores del algoritmo, se busca los grupos que tengan mayor similitud y son estos los que emergen. Este proceso se itera por pares de clúster hasta llegar al nivel superior. Sin embargo, la problemática con un algoritmo de este tipo radica en la selección de un corte para el dendograma que mejor satisfaga la condición expuesta en la problemática de agrupamiento. Manning et al. (2008) proponen diferentes criterios para determinar los grupos:

- Cortar en un nivel de similitud previamente especificado.
- Cortar el dendograma donde el hueco entre dos combinaciones sucesivas de similitud sean más amplias.
- Aplicar la ecuación $K = argmin_{k'}[RSS(K') + \lambda K']$ donde K' referencia al corte de la jerarquía que resulta en K'; RSS es la suma residual de los cuadrados y λ es la penalización por cada grupo adicional.
- Pre-especificar el número de k y cortar donde se obtenga ese número de k.

Agrupamiento divisivo

En este tipo de algoritmo jerárquico, a diferencia del aglomerativo, se comienza con un sólo grupo donde se incluyen todos los documentos. En cada paso del algoritmo, el grupo se va separando hasta que, al final de la cadena, se obtienen grupos con un sólo elemento. Se puede ver el agrupamiento divisivo como el proceso inverso del agrupamiento aglomerativo. Con este método se obtiene también un dendograma y, por tanto, se enfrenta también al problema del método aglomerativo: esto es, la selección de un corte en el dendograma que permita obtener el número de grupos más óptimo para la tarea especificada. En la Figura 1.2 se puede ver un dendograma en donde la flecha hacia abajo indica la construcción por medio de un algoritmo divisivo, mientras que la fecha ascendente muestra el proceso del algoritmo aglomerativo. El proceso consiste en generar jerarquías de grupos a partir de la medida de similitud que se ha determinado (véase 2.2.1). Sobre esta jerarquía se pueden realizar distintos cortes. Por ejemplo, en la figura, un corte en 0.5 agruparía el documento 1 con el documento 2, el documento 3 con el documento 4, y el documento 5 quedaría aislado.

1.9.2. Agrupamiento particional o agrupamiento plano

Los algoritmos de agrupamiento particional, también llamados de agrupamiento, se pueden ver como una implementación computacional del concepto de similitud geométrica de Tversky (1977) que se expondrá más adelante; se plantean representar los elementos en conjuntos dentro de un espacio y determinar la cercanía o lejanía de los elementos para asignar grupos a estos. Usualmente

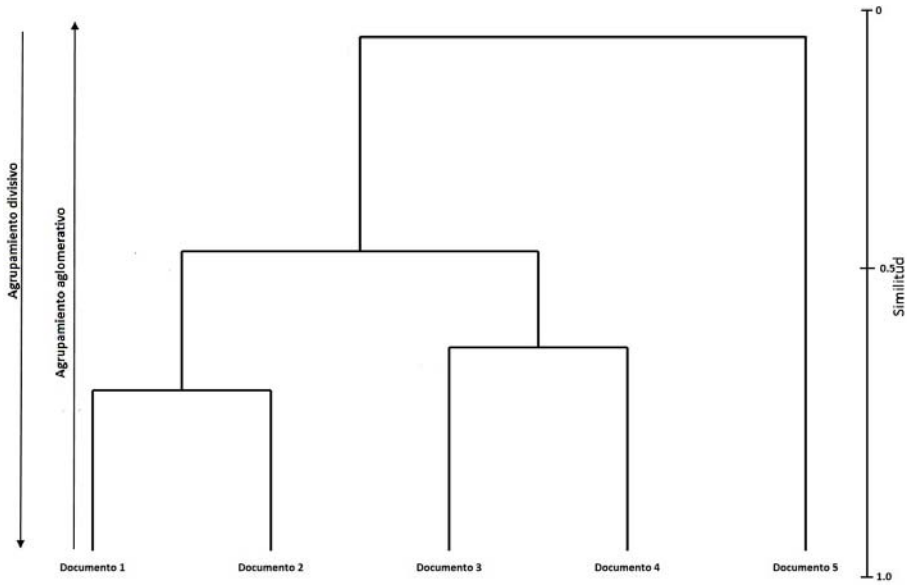

Figura 1.2: Dendograma que muestra el método aglomerativo y divisivo para agrupamiento jerárquico

comienzan con la asignación de una partición aleatoria y a partir de allí refinan esta partición en consecuentes iteraciones del algoritmo.

Para Manning et al. (2008) dicho agrupamiento crea un conjunto plano de grupos sin ninguna estructura explícita que relacione a los grupos entre sí. Esta sería la principal diferencia entre el agrupamiento jerárquico y el plano o particional, pues en el jerárquico si existe una estructura.

Entre este tipo de algoritmos se encuentra *k-medias* que se expone a continuación.

Algoritmo de k-medias

K-medias es uno de los algoritmos con más tradición dentro del área de agrupamiento. Este algoritmo es introducido por Macqueen (1967) quien se propone un método capaz de particionar una población de elementos con características n-dimensionales (es decir, n cantidad de rasgos) en k conjuntos con base en una muestra. Es decir, lo que el algoritmo de k-medias busca es agrupar elementos representados en un espacio n-dimensional dentro de k número de grupos. Este número k es determinado *a priori*, lo que quiere decir que para la implementación de este algoritmo se requiere del conocimiento del total de grupos en

que convergerá el proceso de agrupamiento. Esto ha sido uno de los mayores problemas presentados por este algoritmo, ya que muchas veces se desconoce el total de grupos de un proceso de agrupamiento.

La forma en que el algoritmo de k-medias determina los grupos es a partir de un centroide, esto es, de un elemento que funge como el elemento representativo de un grupo y que generalmente se representa por la media de los puntos o elementos que conforman un grupo.

Para este algoritmo, dado que k es determinado por el experto, primero, debe determinarse el número de k que se requiere. Posteriormente se toman aleatoriamente los primeros k puntos como medias iniciales, es decir, los centroides. Los puntos o elementos se asignan al centroide más cercano con base en una distancia (véase Sección 2.2.1) que generalmente es una distancia euclidiana. Cuando se ha hecho una primera elección de grupos, el algoritmo asigna un nuevo centroide a partir de la media de los puntos conformados por cada uno de los grupos; una vez reasignado un nuevo centroide, el algoritmo itera los procesos para asignar nuevos grupos a los elementos con base en nuevo centroide. Este proceso se realiza hasta que no se hacen nuevas reasignaciones.

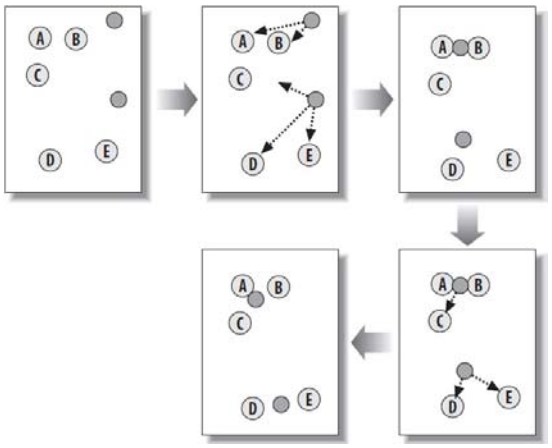

Figura 1.3: Ejemplo de aplicación del algoritmo de K-means (Marmanis and Babenko, 2009)

Un ejemplo de la aplicación de este algoritmo puede verse en la Figura 1.3; aquí se tienen cinco elementos que se distribuyen en dos grupos, según se observa. La primera iteración asigna el centroide aleatoriamente; en una segunda iteración los centroides se mueven a un nuevo espacio que es la media entre los grupos formados por los elementos cercanos. De esta forma, se reasigna un

nuevo centroide a partir de la distancia media entre los puntos de cada grupo. Este proceso se repite una vez más hasta que por último no se realizan más asignaciones; éstos son los grupos finales determinados por el algoritmo.

1.9.3. Algoritmo de MajorClust

En Levner et al. (2007) y Stein and Eissen (2004) se expone un algoritmo basado en grafos para el agrupamiento de documentos: MajorClust. Este algoritmo, puede considerarse como emparentado a los algoritmos jerárquicos pues ambos se basan en un grafo, que en el caso de los modelos jerárquicos es un dendograma. Sin embargo, a diferencia de los métodos jerárquicos, este algoritmo no radica en una jerarquía de grupos, sino que busca la maximización de la conectividad parcial pesada, Λ, de un grafo construido a partir de los vectores de documentos y de una medida de similitud. Por tanto, no cabría clasificarlo como un algoritmo jerárquico, ya que no existe ninguna jerarquía en el grafo construido por el algoritmo.

Básicamente el algoritmo propone que dado un grafo G, tal que $G = \langle V, E, \phi \rangle$ (donde V denota los nodos del grafo, que en este caso son los documentos; E, las aristas entre los documentos delimitados por una función de similitud ϕ, que puede verse como la distancia que separa los nodos), el grafo G se descompone en una colección de grupos C, tal que $C = \{C_1, ..., C_k\}$, donde C_i representan los grupos de documentos con alta similitud, entonces:

$$\Lambda(C) = \sum_{i=1}^{k} |C_i| \cdot \lambda_i$$

Donde λ_i es una *función de máxima atracción* que designa la conectividad entre las aristas de los grupos, $G(C_i)$. Esto se puede definir como el mínimo de aristas que se necesitan eliminar para desconectar el grafo $G(C_i)$, esto es:

$$\lambda_i = min \sum_{(u,v) \in E'} \phi(u, v)$$

Donde $E' \subset E$; es decir, se busca minimizar la función de similitud (o bien la distancia) entre dos aristas pertenecientes a dos nodos o documentos u y v. La descripción del algoritmo se expone claramente en Stein and Eissen (2004) y se muestra en el Algoritmo 1.

Entrada: $G = \langle V, E, \phi \rangle$
Salida: Función $c : V \rightarrow N$ que asigna un grupo a cada nodo.
n = 0, t = False
for $v \in V$ **do**
  while $t = False$ **do**
    $t = True$ **end**
    for $v \in V$ **do**
      if $max(\lambda_i)$ **then**
        $c^* = i$
      end
      if $c(v) \neq c^*$ **then**
        $c(v) = c^*$, $t = False$
      end
    end
  end

Algorithm 1: Algoritmo de MajorClust.

Figura 1.4: Imagen de asignación de nodos a grupos en un grafo por el algoritmo de MajorClust.

Figura 1.5: Reasignación de un nodo dentro de un grupo por el algoritmo de MajorClust.

Lo que hace el algoritmo es determinar un grupo por cada nodo $v \in V$ e itera la propagación de los nodos dentro de los grupos de acuerdo con el principio de máxima atracción λ_i. Gráficamente, el proceso de MajorClust puede verse en las Figuras 1.4 y 1.5. En palabras más generales, lo que el algoritmo recibe como entrada es una grafo que bien puede representarse como en las Figuras 1.4 y 1.5 donde cada punto o nodo es un documento y cada línea es una arista que está determinada por una función de similitud que puede verse como una distancia; es decir, cada documento está separado de los otros documentos por una distancia dada por la función de similitud. Así, si dos documentos son muy

similares, las aristas serán más cortas, mientras que si son muy diferentes, las aristas serán más largas. A partir de este grafo, el algoritmo itera tratando de agrupar en un mismo grupo los documentos más cercanos o con mayor similitud; en cada iteración busca los mejores grupos, hasta que determina que, al desconectar las aristas que no estén dentro de un grupo, se mantienen las relaciones de similitud.

Desde esta perspectiva, se puede elaborar una definición formal de agrupamiento, dada por Stein and Eissen (2004):

> Sea V un conjunto de objetos. Un agrupamiento $C = \{C|C \subseteq V\}$ de V es una división de V dentro de conjuntos para los cuales se presenta la siguiente condición: $\bigcup_{C_i \in C} C_i = V$ y $\forall C_i, C_j \in C$: $C_i \cap C_{j \neq i} = \emptyset$. Asimismo, si V representa los nodos del grafo $G = \langle V, E, \phi \rangle$, entonces C es el agrupamiento de G. Tanto C_i como los subgrafos $G(C_i)$ son grupos.

Esta definición no es muy diferente de las ya dadas, pero resalta el uso de grafos para la realización del agrupamiento. Así, el conjunto de documentos V son los nodos del grafo y a partir de esto se crean grupos C. Se cumple la condición de que la unión de todos los grupos resultantes C_i es el conjunto de documentos V y que no deben existir elementos repetidos dentro de los grupos. En un grafo los grupos se pueden ver como subgrafos de G, tal como se puede apreciar en la Figura 1.4.

1.10. Consideraciones sobre agrupamiento

En este capítulo se revisaron algunos algoritmos para la realización de agrupamiento de documentos. Estos son los algoritmos jerárquicos aglomerativo y divisivo, el algoritmo de $k-medias$ y el modelo de MajorClust. De los algoritmos jerárquicos se pueden señalar varias desventajas que ya nota Segaran (2007); la principal desventaja de estos algoritmos es que en el dendograma que se obtiene los documentos no se separan en grupos sino a través de un proceso que muchas veces es ambiguo, pues se pueden realizar cortes en un nivel de similitud, pero esto no garantiza que para diferentes corpus textuales el límite será siempre el mismo. Otra de las desventajas que menciona el autor es el costo computacional, pues este tipo de algoritmos es el más costoso de la lista. Por ello, el algoritmo de k-medias podría verse como una opción, pues es menos costoso que un algoritmo

jerárquico y que el algoritmo de MajorClust. Sin embargo, el gran problema es la elección de k, pues no siempre se conoce el número de grupos que se deben obtener por medio del agrupamiento; en muchas de las tareas de agrupamiento textual no se conoce k (como es el caso de esta tesis). Por tanto, el algoritmo de k-medias no parece apropiado por lo menos para la tarea que se expone aquí. Los algoritmos jerárquicos también presentan una problemática cuando se trata de comparar métodos, pues no todos los métodos pueden presentar los valores de similitud en igual rango; por esto, también se ha descartado el uso de este tipo de algoritmos. Finalmente, se ha elegido el algoritmo de MajorClust por presentar una menor problemática para los propósitos del presente trabajo.

2

Similitud textual

El concepto de similitud tiene un amplio panorama de aplicación dentro de diferentes áreas científicas y no científicas. Para el agrupamiento de documentos es absolutamente necesario asignar una medida de similitud que sea capaz de indicar qué documentos pertenecen a cada grupo. En este capítulo se presentan algunas nociones básicas sobre la similitud, posteriormente se adentra en el concepto de similitud que se tiene dentro del área de la computación, es decir, el uso de métricas. Por último se presentará la noción de similitud textual que se ha adoptado en la lingüística computacional y se presentarán varios métodos desarrollados en esta área: la relación en taxonomías, la teoría de la información y el supuesto de la hipótesis distribucional.

2.1. Nociones sobre similitud

Tversky (1977) hace notar la importancia de la similitud como principio de organización por el cual se pueden clasificar objetos, formar conceptos y hacer generalizaciones; asimismo, Goldstone (1994) reflexiona sobre el uso de la similitud como modo de categorización del mundo. Sin embargo, cabe señalar que la similitud hace referencia a diferentes rasgos e interacciones entre las características de los elementos y las necesidades que se tengan, pues dependiendo de diferentes áreas, contextos o perspectivas, las similitudes entre dos objetos pueden variar.

Si tomamos la Figura 2.1, por ejemplo, la geometría asumiría que los dos pentágonos presentados son similares pues cuentan con los mismos ángulos, con

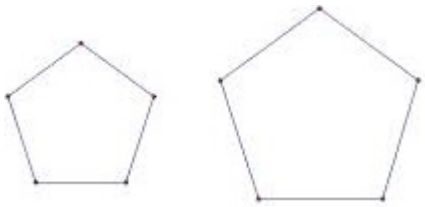

Figura 2.1: Dos polígonos similares

la única variante de la escala. Esta categorización responde a objetivos particulares; si se tratará, por ejemplo de una categorización por tamaños, ambos polígonos bien podrían ser categorizados dentro de grupos diferentes.

Gentner and Ratterman (1991) proponen tres sublclases de similitud:

- Analogía. Implica la similitud a nivel relacional y de estructura, es independiente de los objetos en que estas relaciones se presentan.
- Apariencia. Es complemento de la analogía; se basa en la comparación de la descripción de objetos comunes.
- Similitud literal. Se refiere a un alto grado de rasgos comunes. Involucra la estructura relacional y la descripción de objetos.

Una análisis de la similitud que ha tenido amplio uso en los sistemas de agrupamiento asume que la similitud se puede representar a partir de puntos en un espacio de coordenadas y que ésta se comporta como una función de distancia o métrica (Tversky, 1977). De esta forma, una función de distancia, δ, asigna a cada par de puntos en el plano un número positivo, de tal forma que se cumplan los siguientes axiomas:

1. Minimalidad: $\delta(a, b) \geq \delta(a, a) = 0$.
2. Simetría: $\delta(a, b) = \delta(b, a)$.
3. Desigualdad triangular: $\delta(a, b) + \delta(b, c) \geq \delta(a, c)$.

En este caso, el criterio de minimalidad señala que la distancia entre dos elementos distintos en el plano tiene que ser mayor o igual (si son idénticos) a la distancia del elemento consigo mismo, y que la distancia del elemento consigo mismo es igual a 0, pues es el grado máximo de similitud que se puede alcanzar; la simetría indica que la distancia entre dos elementos tiene que ser la misma en

ambos sentidos, es decir, debe existir la misma distancia entre a y b que entre b y a. Por último, la desigualdad triangular simplemente indica que si existe un tercer punto c, la distancia entre los puntos a y b más la distancia entre uno de estos puntos y c tiene que ser mayor que la distancia simple entre cualquiera de los dos primeros puntos y c.

La problemática, sin embargo, es cómo determinar la función de distancia. Ante ello, primero, deben aclararse presupuestos sobre la similitud, quizás el más importante es que la similitud es dependiente del contexto. Señala Goldstone (1994) que la afirmación «a es similar a b» sólo cobra sentido en cuanto se pueda argüir que «a es similar a b, en cuanto c»; es decir, se pueda especificar en qué contextos son similares ambos elementos. De esta forma, la similitud depende del contexto que rodea al elemento. Un ejemplo planteado por Tversky (1977) es la búsqueda de países similares a Austria; cuando estos países son Suiza, Polonia y Hungría, el que los individuos asocian con mayor similitud a Austria es Suiza; por otra parte, cuando estos países son Suiza, Noruega y Hungría, el que tiene mayor porcentaje de similitud con Austria es ahora Hungría. Esto se debe a que dado el nuevo conjunto, son otros los rasgos que resaltan; así, si se incluye Noruega, el rasgo «escandinavo» ahora es prominente, mientras que en el primer conjunto, sin Noruega, el rasgo con prominencia, según señala el autor, es «forma de gobierno».

Por tanto, para toda tarea que implique similitud se debe aclarar cuál es el contexto y cuáles son los rasgos de los objetos a los que se quiere dar prominencia. En el ámbito que tratamos en esta tesis, cabe limitar la similitud dentro de la lingüística computacional.

En el área de la lingüística computacional se han adoptado dos términos casi como sinónimos para hacer referencia a las similitudes compartidas por dos o más entidades lingüísticas: uno de estos términos es *similitud semántica* y el otro *similitud textual* (Ali, 2011; Harispe et al., 2013; Sahlgren, 2008). En general, ambos términos se refieren a encontrar relaciones que sean capaces de crear conjuntos donde los miembros de un mismo conjunto compartan características entre sí, mientras que se puedan distinguir de los miembros de otros conjuntos.

Harispe et al. (2013) define similitud semántica como la medida de comparación de las propiedades de rasgos que comparten y los rasgos específicos a cada uno de los elementos. Lo que estos autores proponen no está lejos de las otras propuestas (por ejemplo, Lin (1998)); se basa en las características propias de los elementos que se comparan, propiedades que pueden compartir o que pueden ser distintivas. En cuanto al concepto de *similitud textual*, se puede decir que es

una propiedad, la cual es posible medir entre dos o más textos, y que determina el grado de similitud entre ellos.

La similitud entre elementos lingüísticos, así entendida, busca agrupar los elementos del lenguaje, de tal forma que aquellos con un mayor número de rasgos compartidos entren dentro de un conjunto, mientras que aquellos con mayor cantidad de rasgos distintivos se agrupen en grupos distintos. El problema, entonces, radica en encontrar una medida fiable que permita realizar esta tarea. Las aproximaciones para resolverla han sido distintas. Padó and Lapata (2003) afirman que «The semantic similarity between words can be then mathematically computed by measuring the distance between points in the semantic space using a metric as cosine or Euclidean distance» (Padó and Lapata, 2003, p. 128). Dentro del área de agrupamiento no es raro que se utilicen este tipo de medidas para la detección de similitud entre documentos (Senellat and Blondel, 2003). Sin embargo, la lingüística computacional se ha encargado de abordar el problema de la similitud para encontrar medidas que sean más precisas y determinen un mejor valor cuantitativo a la similitud entre elementos lingüísticos.

En este capítulo nos enfocamos a analizar algunos de estos enfoques propuestos: desde los métodos basados en la teoría de la comunicación de Shannon (1948), hasta aquellos que se preocupan más por recuperar la semántica de las palabras a través de taxonomías o incluso por medio de métodos distribucionales.

2.2. Aproximaciones para detección de similitud textual

En un marco general, Metzler et al. (2007) distinguen dos categorías generales de medidas de similitud: *léxicas* y *probabilísticas*. Las similitudes léxicas se basan en la idea básica de encontrar patrones de coincidencia entre términos de una representación superficial, y se pueden dividir en tres tipos:

- **Exacta:** Considerando que Q es un término que se busca y C es una coincidencia encontrada, si Q y C son léxicamente equivalentes se dice que son «exactos».
- **Frase:** Cuando C es una subcadena de Q
- **Subconjunto:** Los términos en C son un subconjunto de los términos en Q.

Este tipo de medidas suelen ser precisas, pues capturan la similitud entre dos cadenas cuando contienen léxico similar; sin embargo, el nivel de aplicación de este tipo de medidas es muy pobre, pues no logran capturar todas las coincidencias que tengan similitud fuera del nivel léxico.

Un segundo tipo de medidas de similitud es la que Metzler et al. (2007) llaman probabilísticas. Se basan en la probabilidad de una palabra en un vocabulario y ocupan métodos matemáticos para calcular la similitud. De igual forma, se pueden combinar ambos métodos para formar uno híbrido que aproveche las ventajas de los métodos léxicos y de los probabilísticos.

Por su parte, se toma en cuenta tres aproximaciones para la detección de similitud: modelos de espacio vectorial, alineamiento textual y superposición de n-gramas, en donde caben la mayoría de los métodos propuestos en la literatura; sin embargo, existen otras aproximaciones que no pueden clasificarse en ninguno de estos tres grupos. A saber, los modelos basados en relaciones léxicas o taxonomías, así como el modelo basado en información de Lin (1998). Además, aquí también se toman en cuenta los modelos basados en la distribución, llamados *espacios semánticos*[1].

A continuación describimos estas aproximaciones, resaltando sus ventajas y desventajas.

2.2.1. Modelos de espacio vectorial

Dentro del área de la recuperación de información, y más precisamente del agrupamiento, se ha propuesto la representación de documentos como vectores en un espacio vectorial. En general se parte de una matriz, donde cada vector representa un documento y a partir de estos vectores se puede calcular una distancia de similitud entre ambos que determinará qué tan lejanos o cercanos están los documentos; «We define a similarity metric between patterns based on the number of overlapping argument pairs, and then use a clustering procedure based on this similarity to group patterns» (Grishman, 2010, p. 526).

En Lee (1999) y en Huang (2008) se pueden encontrar algunas de las métricas más conocidas en el área de recuperación de información. Las métricas que presenta se basan en la idea de representar los documentos dentro de un espacio euclidiano. Éstas son:

[1]Los modelos de *espacios semánticos* bien pueden caber dentro de los modelos de espacio vectorial, pues pueden verse como una especialización orientado al lenguaje natural de estos modelos. Sin embargo, por las características propias de los modelos basados en distribución, aquí se consideran como un modelo aparte.

Distancia euclidiana:

$$L_2(q,r) = \sqrt{\sum_v (q_i) - r_i)^2} \tag{2.1}$$

L_1 normalizado:

$$L_1(q,r) = \sum_v |q_i - r_i| \tag{2.2}$$

Coeficiente de Jaccard:

$$Jac(q,r) = \frac{|\{i : q_i > 0 \text{ y } r_i > 0\}|}{\{i|q_i > 0 \text{ o } r_i > 0\}} \tag{2.3}$$

Divergencia de Skew:

$$S_\alpha(q,r) = \sum_i q_i \cdot log(\frac{q_i}{\alpha \cdot q_i + (1-\alpha) \cdot r_i}) \tag{2.4}$$

Distancia de coseno:

$$\cos(q,r) = \frac{q \cdot r}{\sqrt{q \cdot q \times r \cdot r}} \tag{2.5}$$

Medida cluster:

$$cluster(q,r) = \frac{q \cdot r}{||q||_l} \tag{2.6}$$

Coseno suave:

$$soft_cosine(q,r) = \frac{\sum\sum_{i,j=1}^{N} q_{i,j} r_{i,j}}{\sqrt{\sum\sum_{i,j=1}^{N} q_{i,j}}\sqrt{\sum\sum_{i,j=1}^{N} r_{i,j}}} \tag{2.7}$$

Estas medidas también se pueden categorizar como métricas, pues cumplen los cuatro puntos planteados por Huang (2008); estos son:

1. Dado dos objetos x e y, la distancia entre ambos es positiva; es decir, $d(x,y) \geq 0$.

2. La distancia entre los dos objetos debe ser 0 si y solo si los dos objetos son idénticos, $x = y$.

3. La distancia debe ser simétrica.

4. La distancia debe satisfacer la desigualdad triangular.

Una de las distancias más simples es la distancia euclidiana, presentada en la Ecuación 2.1; se basa en un espacio euclideo, donde se trata de determinar la distancia entre dos puntos o vectores, en este caso. Un caso más simple es la Ecuación 2.2; en cuyo caso simplemente se restan los elementos del vector y el valor absoluto de la resta representa la distancia. Por su parte, Lee (1999) nota que el coeficiente de Jaccard, que se muestra en la Ecuación 2.3, difiere de las otras métricas en que es esencialmente *combinatorio*, pues se basa en los tamaños de q, r y $q \cdot r$ más que en sus distribuciones. Lee (1999) propone una nueva métrica, *Skew Divergence* (Ecuación 2.4), que también retoma Padó and Lapata (2003).

La distancia de coseno, que se puede ver en la Ecuación 2.5, se retoma más adelante; dicha métrica se basa en determinar el ángulo entre dos vectores, siendo que si los vectores cumplen con un ángulo igual a 0, la similitud es máxima y $cos(0) = 1$; mientras que si el ángulo es máximo, es decir de $\frac{\pi}{2}$, los documentos son disímiles y se tiene que $cos(\frac{\pi}{2}) = 0$. Por último, mostramos la *medida cluster* expuesta en Senellat and Blondel (2003); en este caso la semejanza se da como el producto interno de los vectores entre la suma de las magnitudes de i dado por $||i||_l$.

Una distancia basada en la fórmula de coseno es la de coseno suave (Sidorov et al., 2014). En este caso, cada $x_i j$ se define como: $x_i j = \sqrt{s_{ij}} \frac{x_i + b_j}{2}$, donde $s_{ij} = cos(e^i, e^j) = sim(f_i, f_j)$; es decir, es el coseno entre vectores base y f_i y f_j son rasgos que corresponden a estos vectores base, así $sim()$ determina una medida de similitud (como puede ser la sinonimia). Una descripción más detallada se encuentra en Sidorov et al. (2014).

2.2.2. Alineamiento textual

Los modelos basados en alineamiento textual comparan dos cadenas de caracteres para calcular las operaciones que transforman la primera de las cadenas en la segunda de ellas[2]. Toman en consideración tanto los caracteres como las palabras. Una primera aproximación es la distancia de Levenshtein, que formalmente se define como:

[2]Las cadenas pueden estar conformados por caracteres, por palabras o bien lemas. Por simplicidad, aquí se toma el caso más simple, es decir, el de los caracteres.

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & \text{si } min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases} & \text{de otra forma} \end{cases}$$

En este caso i y j representan dos cadenas a comparar. Si no hay ningún cambio en las cadenas (es decir, si son iguales), la distancias de Levenshtein es igual a 0. En otro caso, si la diferencia entre las cadenas es de una inserción, se suma 1 a la distancia, por ejemplo entre «gato» y «gatos» la distancia es igual a 1; si la diferencia es una elisión, se suma 1 a la distancia, como en «hola» y «ola»; finalmente, si lo que se presenta es un trueque de letra, también se suma 1 a la distancia, como en «hola» y «bola». Esta distancia no está normalizada y puede tener valores de entre 0 y n.

Una distancia similar a la de Levenshtein es la distancia de Jaro-Winkler (Ali, 2011), que formalmente se describe como:

$$d_j = \begin{cases} 0 & \text{si } m = 0 \\ \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}) \end{cases}$$

Donde m es el número de caracteres iguales entre la sentencia s_1 y otra sentencia, s_2, y t es la mitad del número de caracteres iguales pero en diferente orden. Así, si se toman dos palabras como «sola» y «suela» se tiene que $m = 3$, pues comparten 3 caracteres, $|s_1| = 3$ (el tamaño de la primera palabra), $|s_2| = 5$ (el tamaño de la segunda palabra) y $t = 0$. De tal forma que $d_j = \frac{1}{3}(\frac{3}{|4|} + \frac{3}{|6|} + \frac{3-0}{4}) \approx 0.7$. Esta distancia está normalizada, es decir, toma valores entre 0 y 1, a diferencia de la distancia de Levenshtein.

Una última distancia es la distancia de Ratcliff/Obershelp (Black, 1998) que toma en cuenta los caracteres compartidos de una cadena y los divide sobre el total de caracteres de la cadena más larga. Si tomamos el ejemplo anterior de «sola» y «suela» se tendría $sim = \frac{3}{5} = 0.6$, ya que se comparten 3 caracteres y la palabra con mayor número de caracteres es «suela».

Estos métodos pueden servir para detectar similitud a nivel de palabras. Su aplicación a un problema de detección temática no es justificable, pues en ellos no se captura, ni se trata de capturar, ninguna característica temática o semántica, si no que se enfoca en las características gráficas y fonológicas, aunque tales características representen a un caracter, una palabra o un lema.

Por tanto, para los fines de este trabajo este tipo de modelos quedan descartados.

2.2.3. Superposición de n-gramas

La superposición de n-gramas mide el número de n-gramas, sean estos de palabras, caracter o lema, compartidos por dos textos; es decir, las combinaciones de n número de palabras en un documento. La medida de similitud se calcula usando una métrica, como las expuestas en 2.2.1.

2.2.4. Similitud basada en taxonomías

Una taxonomía describe una jerarquía de elementos conceptuales dentro de dominios específicos o clases. El concepto de taxonomía describe una estructura en donde, como describe Tyler (1969), i) los elementos al mismo nivel contrastan entre sí; y ii) elementos en diferentes niveles se relacionan por inclusión. Se nota también, dentro del mismo concepto de jerarquía, que los elementos de mayor nivel son más inclusivos y pueden ser clases que incluyan a elementos de niveles más bajos. En general, una taxonomía se puede representar como un árbol en donde elementos de mayor concentración conceptual dependen de elementos más difusos llamados clases. Las relaciones que subyacen en los elementos de la taxonomía son de orden léxico; es decir, pueden existir relaciones de sinonimia, hermandad taxonómica, hiperonimia, hiponimia, meronimia, entre otras. Un ejemplo de la representación gráfica de una taxonomía puede apreciarse en la Figura 2.2, donde se observa que el término superior en la jerarquía es «entidad» y este comprende conceptos más concretos conforme el nivel de la jerarquía es más bajo.

Con tal estructura en mente, se han originado diferentes propuestas para la aproximación de similitud textual. Esta aproximación se basa en la idea de que si palabras o elementos lingüísticos pueden ser definidos como conceptos de una estructura jerárquica en una taxonomía, es posible entonces determinar una medida de información basada en dicha taxonomía (Wu and Palmer, 1994).

En general se pueden distinguir dos grandes clasificaciones de la aproximación taxonómica al concepto de similitud. Jiang and Conrath (1997) distinguen entre la *aproximación basada en nodos* (contenido de información) y la *aproximación basada en aristas* (distancia). A continuación se revisan estas dos propuestas.

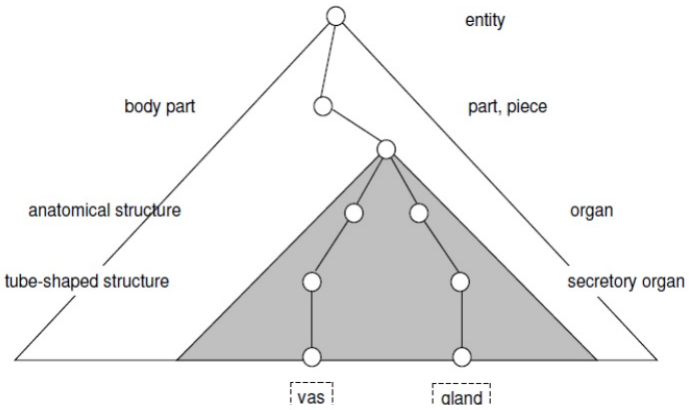

Figura 2.2: Ejemplo de una taxonomía simple (Castro et al., 2011)

Enfoque basado en aristas (distancia)

Un enfoque de este tipo estima la distancia (longitud de la arista) entre nodos que corresponden a un concepto o clase comparada. En una red semántica con relaciones del tipo *es-un*, la manera más simple de determinar la distancia entre dos nodos de conceptos elementales, A y B, es el camino más corto entre A y B. En general, dos conceptos similares serán aquellos que minimicen la distancia entre sus nodos, mientras que dos conceptos disímiles serán aquellos cuya distancia entre ellos sea máxima. Dos enfoques similares a esta aproximación son expuestos en Wu and Palmer (1994), Maynard (2000), Vivaldi et al. (2010) y Castro et al. (2011).

Wu and Palmer (1994) proponen una medida basada en la profundidad de dos conceptos dentro de una taxonomía. Formalmente, esta medida de similitud se define como:

$$ConSim(c_1, c_2) = \frac{2N_3}{N_1 + N_2 + 2 \cdot N_3}$$

Donde N_3 es la profundidad del nodo que separa los conceptos dentro de la taxonomía, y N_1 y N_2 son la profundidad de los conceptos c_1 y c_2, respectivamente. Una versión adaptada de este algoritmo se expone en Maynard (2000), y posteriormente en Vivaldi et al. (2010) y Castro et al. (2011):

$$sim(sy_1, sy_2) = \frac{2 \cdot \#CommonNodes(sy_1, sy_2)}{Depth(sy_1) + Depth(sy_2)}$$

En esta variación sólo se elimina la profundidad del nodo que separa los

conceptos N_3 en el divisor. Un ejemplo claro de la aplicación de este tipo de algoritmos puede mostrarse con los términos «vas» y «gland» de la Figura 2.2. Lo primero que se puede observar es que se encuentran en una relación de hermandad taxonómica. Lo que se busca, entonces, es encontrar una medida cuantitativa que determine la similitud entre ambos documentos; en este caso tanto N_3 como $\#commonNodes$ en los algoritmos tienen un valor de 2, pues es el número de nodos que hay entre el término genérico «entidad» y el nodo en que si bifurcan las ramas de la taxonomía para representar los conceptos de interés. Los valores de N_1 y N_2 (o bien $Depth(sy_1)$ y $Depth(sy_2)$) son ambos iguales a 5, pues es el total de nodos desde el término genérico al término de interés. De esta forma, la propuesta de Wu and Palmer (1994) sería:

$$\frac{2 \cdot 2}{5 + 5 + (2 \cdot 2)} \approx 0.286$$

La propuesta de Maynard (2000), por su parte, muestra un valor más alto al eliminar N_3, por lo que para estos autores, la similitud entre «vas» y «gland» está dada por:

$$\frac{2 \cdot 2}{5 + 5} = 0.4$$

Enfoque basado en nodos (información contenida)

Este enfoque tiene bases fuertemente ligadas a la propuesta de Shannon (1948) y puede equipararse a la propuesta de Lin (1998) expuesta en la Sección 2.2.6. La principal diferencia radica en que este enfoque utiliza una taxonomía y la información acumulada en ella para calcular la similitud, mientras que la propuesta de Lin (1998), como veremos más abajo, no requiere el uso de este recurso.

Dado un espacio multidimensional en el que un nodo represente un concepto único que consiste de cierta cantidad de información, y una arista que represente una asociación directa entre dos conceptos, la similitud entre ambos es el grado de información que comparten. La información contenida (IC) de un concepto o clase se puede representar formalmente de la siguiente forma:

$$IC(c) = log^{-1} P(c)$$

Es decir, se trata del inverso del logaritmo de la probabilidad del concepto dentro de la taxonomía. El valor logarítmico, tal como lo explica Shannon (1948), cuantifica la información de un concepto; en la Figura 2.3 se puede ver un

ejemplo de los términos de una taxonomía con un valor de información asociado. Por tanto, la forma más sencilla de calcular la similitud entre un concepto c_1 y otro concepto c_2 se puede expresar como:

$$sim(c_1, c_2) = max_{c \in Sup(c_1,c_2)}[IC(c)]$$

En otras palabras, la similitud es el valor que maximice la IC del concepto o clase c superior en la jerarquía a c_1 y c_2. Un ejemplo claro se puede dar con los términos «car» y «bicycle» en la Figura 2.3, donde el concepto superior o clase es «vehicle»; por tanto $sim(car, bicycle) = 8.3$, mientras que $sim(car, fork) = 3.53$. Se ve entonces que la información contenida es mayor en el par «car» y «bicycle».

Los mismos Jiang and Conrath (1997) proponen una medida basada en las ideas de la información contenida (IC) y en la distancia de aristas. Dado un concepto c y una clase o concepto padre p, se calcula la probabilidad del concepto iésimo dado el concepto padre:

$$P(c_i|p) = \frac{P(c_i \cap p)}{P(p)} = \frac{P(c_i)}{P(p)}$$

Esto es, la probabilidad del concepto dado el padre es simplemente la probabilidad del concepto entre la probabilidad del padre. Así, si por ejemplo la probabilidad de un concepto cualquiera c_i es de $\frac{1}{4}$, y este concepto tiene en la taxonomía un nodo padre con probabilidad $\frac{1}{2}$, entonces $P(c_i|p) = \frac{1}{8}$. El segundo paso para esta metodología es calcular la fuerza del vínculo entre el concepto y el padre o «Link Strength» (LS) de la siguiente forma:

$$LS(c_i, p) = -log(P(c_i|P)) = IC(c_i) - IC(p)$$

Es decir, LS será la información contenida de la probabilidad del concepto iésimo dado el padre, y esto será igual al IC del concepto menos la IC del padre. Por tanto, siguiendo con el ejemplo anterior, si tenemos que $P(c_i|p) = \frac{1}{8}$, entonces $LS = -log(\frac{1}{8}) \approx 2.8$.

Posteriormente, a las aristas se les asigna un peso:

$$wt(c, p) = (\beta + (1 - \beta)\frac{\overline{E}}{E(p)})(\frac{d(p) + 1}{d(p)})^{\alpha}[IC(c) - IC(p)]T(c, p)$$

Donde se definen los siguientes elementos:

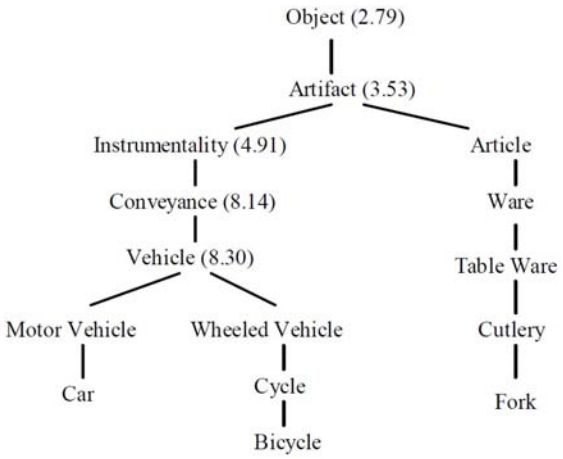

Figura 2.3: Ejemplo de taxonomía con valores de información contenida (Jiang and Conrath, 1997, p. 4)

- $d(p)$ es la profundidad del nodo p
- $E(p)$ el número de aristas que se unen al hijo (densidad local)
- $\overline{E}$ es la densidad media en toda la jerarquía
- $T(c,p)$ el factor de unión relación/tipo
- α y β controlan el grado de afectación de la densidad y la profundidad, tal que: $\alpha \geq 0$ y $0 \leq \beta \leq 1$

La distancia promedio entre dos nodos será la suma de las aristas con peso sobre el camino más corto que vincule los dos nodos a comparar:

$$Dist(w_1, w_2) = \sum_{c \in [path(c_1,c_2) - LSuper(c_1,c_2)]} wt(c, parent(c))$$

En este caso, $LSuper(c_1, c_2)$ denota el super-ordinado más bajo de c_1 y c_2.

En general, los métodos basados en taxonomías sufren el problema de depender de un recurso externo, que es difícil de elaborar. El cálculo de la similitud se ve, entonces, limitada por la precisión con la que se elabora una taxonomía, ya que si en ésta no se representan bien los términos y sus jerarquías, los resultados obtenidos por alguno de estos métodos reflejarán las deficiencias de la taxonomía. La elaboración de una taxonomía que represente todos los términos

de un dominio específico es una tarea laboriosa y requiere de buenos conocimientos sobre el dominio en que se está trabajando; además, el método se vuelve dependiente del dominio, pues cuando se quiera aplicar a un dominio distinto se tendrá que elaborar una nueva taxonomía.

2.2.5. Similitud basada en diccionarios

Lesk (1986) se plantea la problemática de la ambigüedad en las palabras y propone una solución computacional. Propone un método para decidir automáticamente cuál es el mejor sentido de una palabra en un contexto a partir del uso de diccionarios legibles por computadora. Su propuesta es simple, se basa en la idea de que palabras vecinas (o que comparten un mismo contexto) suelen tener tópicos comunes. Esto es, que teniendo una palabra objetivo t y dado un conjunto de definiciones $\{d_1, d_2, ..., d_n\}$ en una entrada D_n de diccionario, tal que $d_n \in D$, y un conjunto de palabras vecinas $\{p_1, p_2, ..., p_n\}$ que conforma el contexto $\mathbb{C}$ de t, cada una con un conjunto de definiciones asociadas, tal que $\forall p_n : \exists D_n$, la definición asociada (d_n) a t es aquella que comparta un mayor número de palabras comunes al contexto $\mathbb{C}$ de t.

Una adaptación de este algoritmo la plantea Banerjee and Pederson (2002), con base además en el uso de taxonomías, específicamente *WordNet*[3]. La propuesta de Banerjee and Pederson se basa en una jerarquía del tipo *is-a*; se propone una medida para la desambiguación de palabras basada en la información contenida en tal jerarquía. Básicamente, se trata de una adaptación del algoritmo de Lesk. Esta medida toma como entrada un ejemplo o instancia en la cual ocurra la palabra relevante y asigna un sentido a partir de *WordNet* (que en este caso funge como el diccionario), dependiendo de la información proporcionada por otras palabras inmediatas a ésta en un contexto. Este *contexto* lo definen como una ventana de n tokens a la derecha e izquierda, tal que el contexto es igual a $2n + 1$, incluyendo a la palabra objetivo.

Posteriormente se evalúan cada posible combinación de sentidos, tal que hay $\coprod_{i=1}^{N} |W_i|$ candidatos a combinación. A partir de esto, se computa un valor de combinación y la combinación con mayor valor es la que determina el sentido de la palabra objetivo.

En Pedersen et al. (2004) se propone el uso del algoritmo de Lesk adapatado para estimar la similitud entre dos palabras a partir del uso de *WordNet*. En este caso, se utiliza el algoritmo de Lesk para encontrar superposiciones entre

[3]Para más detalles puede visitarse la página: *wordnet.princeton.edu*

las entradas de los conceptos A y B; a partir de ello se crea una matriz de co-ocurrencias de cada palabra usada en las entradas de *WordNet*.

En general, el algoritmo de Lesk no es un algoritmo pensado para encontrar similitud entre elementos; sin embargo, deja ver cómo las palabras dentro de un contexto se relacionan a partir de «tópicos», o bien «temáticas», similares. Por tanto, es posible adaptar este algoritmo para encontrar similitud entre elementos lingüísticos (Pedersen et al., 2004). Lo que deja ver el algoritmo de Lesk es que hay una relación entre las palabras que pertenecen a un mismo contexto (que las hace similares), pues éstas comparten un tópico o tema. Lesk (1986) ejemplifica con *pine* y *cone*, y nota que comparten una entrada de diccionario en donde aparecen las palabras *evergreen* y *tree*. Finalmente, lo que nos deja ver es que las palabras en un mismo contexto no aparecen aleatoriamente, sino que están determinadas, en cierta medida, por el contexto. Por tanto, podemos pensar que el contexto (temático) está determinando el uso de ciertos elementos, que en este caso son términos.

2.2.6. Similitud basada en teoría de la información

La teoría de la información es un concepto que surge a partir de Shannon (1948), en donde se propone un modelo matemático de la comunicación. Antes de introducir la propuesta de similitud basada en información, conviene introducir el concepto de información en el ámbito del lenguaje. El concepto de información dentro de este marco se puede entender como la comprensión mínima de datos necesaria para transmitir un mensaje (Cover and Thomas, 2006). Dada la fuente discreta por la que se transmite el mensaje, la información se mide en dígitos binarios o *bits*. Así, Shannon (1948) señala que «a flip-flop circuit, can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $log_2 2^N = N$» (Shannon, 1948, p. 379).

Ben-Naim (2011) sintetiza lo dicho arriba con un ejemplo simple. Considérese un tablero como el que se muestra en la Figura 2.4; en una de las casillas se ha escondido una moneda y el objetivo es encontrarla con el mínimo de preguntas posibles. Entonces, la forma más eficiente de hacer esto es hacer preguntas binarias (sí y no) de la siguiente forma: *¿se encuentra la moneda en la mitad derecha del tablero?*, *¿se encuentra en la mitad superior de las casillas restantes?*, y por último *¿se encuentra en la mitad izquierda de las casillas restantes?* Se debe notar que sin importar la respuesta (sí o no) se puede obtener la ubicación de la moneda con el mismo número de preguntas. Así, el mínimo de preguntas

para obtener la respuesta correcta para un total de 8 casillas es de 3, lo que es coherente con la definición de información, es decir $log(8) = 3$. Esto quiere decir que se necesitan 3 preguntas binarias cuando tengo 8 casillas, esto es $2^3 = 8$.

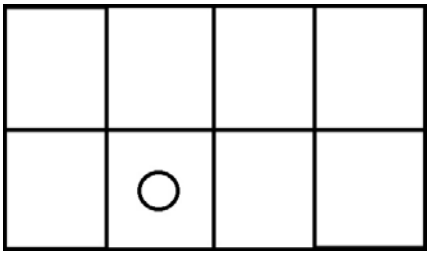

Figura 2.4: Ejemplo de información

Con base en esto, Lin (1998) y Cover and Thomas (2006) definen la información formalmente como $I = -log_2 p(i)$ [4], donde $p(i)$ es la probabilidad del elemento i. Y así Lin (1998) propone un modelo teórico de información para la definición de similitud. Su propuesta tiene dos objetivos principales:

- Universalidad. Que el modelo sea aplicable tanto como el dominio tenga un modelo probabilístico.
- Justificación teórica. Que el modelo tenga una propuesta teórica derivada de una serie de supuestos sobre el concepto de similitud.

Sobre el concepto de similitud, este autor se plantea tres intuiciones primarias:

- La similitud entre A y B se relaciona con sus rasgos comunes. Entre más rasgos comunes compartan, más similares serán.
- La similitud entre A y B se relaciona con sus diferencias. Entre más diferencias tengan, menos similares serán.
- La similitud máxima entre A y B es conseguida cuando A y B son idénticos, sin importar cuántos rasgos comunes compartan.

Conforme a estas suposiciones básicas, el autor desarrolla su propuesta de similitud, que está estrechamente relacionada con la teoría de la información.

[4] En este caso se utiliza un logaritmo negativo para que el valor de I sea positivo, ya que cuando $0 < i < 1$, $log(i) = -I$.

Entiende que la similitud entre dos elementos A y B, se puede determinar por la información compartida, definida como:

$$I(common(A, B)) = -logP(common(A, B))$$

Donde $common(A, B)$ son los rasgos comunes entre el elemento A y el elemento B. Con esta misma suposición determina la información de los rasgos diferenciales del mismo par de conjuntos como:

$$I(description(A, B)) = -logP(description(A, B))$$

En este sentido, el autor plantea que $description(A, B)$ describe que son A y B, pues son los rasgos diferenciales entre uno y otro. Con esto en mente, la similitud $sim(A, B)$ es una función de los rasgos comunes y diferencias de A y B. Finalmente, plantea una función de similitud tal que:

$$sim(A, B) = f(I(common(A, N)), I(description(A, B)))$$

De tal forma que el dominio de f será $\{(x, y)|x \geq 0, y \geq 0, y \geq x\}$; es decir, los rasgos compartidos y los rasgos similares siempre tendrán que ser mayores o iguales a 0, y los rasgos diferenciales serán mayores o iguales a los rasgos compartidos, esto para que no se convierta en una función indefinida. Así, si los documentos son idénticos, la similitud es igual a 1, pues $\forall x > 0, f(x, x) = 1$, y si son absolutamente diferentes, la similitud es igual a 0, ya que $\forall y > 0, f(0, y) = 0$.

Con estas suposiciones, se llega a las siguientes ideas:

$$\forall x_1 \leq y_1, x_2 \leq y_2 : f(x_1 + x_2, y_1 + y_2) = \frac{y_1}{y_1 + y_2} f(x_1, y_1) + \frac{y_2}{y_1 + y_2} f(x_2, y_2)$$

De tal forma que las similitud entre A y B, $sim(A, B)$, se determinará por una función de la información de rasgos comunes y la información de rasgos diferenciales que define con el siguiente algoritmo:

$$sim(A, B) = \frac{logP(common(A, B))}{logP(description(A, B))}$$

Lo que se plantea Lin (1998) es que la similitud entre dos elementos, A y B, es un cociente de la información que comparten, sobre la información que los diferencia. Se puede ver que la función de similitud de Lin (1998) es bastante

intuitiva, pues plantea que la similitud está dada por qué tanto se parecen dos elementos y qué tanto se diferencian. Lin (1998) muestra cómo su algoritmo se aplica al caso de palabras, siendo los rasgos en común los contextos que comparten, mientras que los rasgos diferenciales, los contextos en que difieren. La aplicación de este algoritmo para el caso de documentos, empero, muestra una complicación distinta, pues es difícil determinar el contexto de un documento. Quizá el modelo más simple sea a partir de las palabras que comparten y las que no, como se muestra a continuación.

Propuesta de Lin aplicada a documentos

La propuesta de Lin (1998), ya se ha señalado, no especifica cuáles son los elementos que se comparten y cuáles los diferenciales. Ya mencionado que los términos juegan un papel importante en la distinción temática de documentos, se planteó la idea de que fueran precisamente estos elementos los que fungieran como elementos de intersección y diferenciales. Es decir, dado el conjunto $D = \{t_1, t_2...t_n\}$ donde t_n equivale a los términos del documento, los elementos en común se definen como:

$$I(common(D_1, D_2)) = -logP(D_1 \cap D_2)$$

Como se puede ver, se trata de la información compartida entre un documento y otro. En el caso de los elementos diferenciales de D_1 y D_2, se entendió estos como los términos que no se comparten (la diferencia simétrica) entre ambos documentos:

$$I(description(D_1, D_2)) = -logP(D_1 \triangle D_2)$$

La adaptación del método trata de no alejarse de la propuesta de Lin (1998); así, sólo se plantea una especificación, tratando de conservar los elementos de información (el valor logarítmico). De esta forma, sustituyendo en la fórmula original de Lin (1998), se obtiene que la similitud entre dos documentos está dada de la siguiente forma:

$$sim(D_1, D_2) = \frac{\{-\sum_1^n log(P(x_n)) : x_n \in D_1 \cap D_2\}}{\{-\sum_1^n log(P(x_n)) : x_n \in D_1 \triangle D_2\}} \tag{2.8}$$

Lo que se plantea aquí es que la similitud entre un documento, D_1, y otro, D_2, está determinada por la información de los términos compartidos sobre la información de los términos diferenciales. Uno de los problemas que se puede

pensar ante esta propuesta es que, en principio, fue planeada para determinar similitud entre palabras (Lin, 1998). Sin embargo, la mayor carencia que se puede ver en la propuesta de Lin es que, al trabajar por pares de documentos, no contrasta el cómo se distribuyen los términos en las diferentes temáticas. Por tanto, se plantea contrastar esta propuesta con otras que sí tratan de capturar esta distribución, a decir, los llamados *espacios semánticos*.

2.2.7. Hipótesis distribucional y concepto de similitud

La idea general de la hipótesis distribucional se basa en la idea de que elementos del lenguaje similares aparecen en contextos similares. En general, este enfoque se ha aplicado a palabras, bajo el supuesto de que palabras similares tienden a compartir un contexto (Sahlgren, 2008). Se trata de una relación entre elementos paradigmáticos y sintagmáticos.

La hipótesis distribucional (Harris, 1954) o estructura distribucional (Sahlgren, 2008) surge del concepto de estructura distribucional propuesto por Harris (1954). Para este autor, la distribución de un elemento se identifica con la suma de todos los contextos en que dicho elemento aparece, donde el contexto se define como: «An environment of an element A is an existing array of its co-ocurrences, i.e. the other elements, each in particular position, with which A occurs to yield an utterance» (Harris, 1954, p. 3). Visto de esta forma, la estructura distribucional de un elemento lingüístico no es otra cosa que ese elemento lingüístico (fonema, morfema, palabra o elementos superiores) dentro de un arreglo que especifique los elementos que le rodean y sus posiciones determinadas.

Con respecto a esto, Sahlgren (2008) propone dos pasos a seguir para llevar a cabo lo que el llama la «hipótesis distribucional», a saber:

1. Construir perfiles distribucionales para palabras basadas en qué palabras ocurren a su alrededor.

2. Construir perfiles distribucionales basados en qué región del texto ocurre una palabra.

Para validar las propuestas basadas en la distribución de elementos lingüísticos, empero, se debe argüír, como Harris (1954) lo hace, que: i) los elementos del lenguaje no ocurren de forma arbitraria con respecto de otras, sino que cada elemento ocurre en una posición relativa con respecto de otra; ii) esta distribución de clases es persistente a través de las ocurrencias de un mismo elemento; iii) es

posible determinar la posición relativa de un elemento con respecto a otros; y iv) las restricciones en la ocurrencia relativa de un determinado elemento pueden describirse por una red de reglas interrelacionadas.

Si bien estos cuatro puntos nos dejan ver que un análisis basado en la distribución de un lenguaje es factible, todavía no se responde la relación que puede existir entre la estructura distribucional y la similitud. Sahlgren (2008) apunta que «the distributional methodology is only concerned with meaning differences, or, expressed in different terms, with *semantic similarity*» (Sahlgren, 2008, p. 4). Si bien el concepto de *similitud semántica* puede resultarnos problemático, es claro que las relaciones sintagmáticas y paradigmáticas que captura la estructura distribucional de una lengua nos pueden, en efecto, dar información sobre oposición de sentidos o incluso relaciones sinonímicas, hiponímicas, de antonimia, etc. El ejemplo más claro de relaciones de oposición puede mostrarse a nivel morfológico; en este caso, si tomamos una palabra, por ejemplo «*niñ-o*», queda claro que existe una oposición paradigmática con respecto al morfema de género en contraste con «niñ-a». La pregunta relevante para los objetivos de esta tesis es, entonces, ¿estos elementos lingüísticos, en este caso los morfemas, están determinados por el significado? La relación entre los morfemas *-o/-a*, parece marcar una diferencia de significada a nivel sintagmático. Citando a Harris (1954): «meaning is of course a determinant in these and in other choices that we make when we speak» (Harris, 1954, p. 13). En general, este autor plantea que las diferencias de significado tienen una correlación con las diferencias distribucionales.

El problema planteado en el marco de esta tesis es más complicado que el de elementos morfológicos o incluso que al tratar sólo con palabras. Se quiere saber si la distribución de ciertos elementos lingüísticos puede determinar similitud o contraste entre elementos discursivos (es decir, textos completos). Parecería que la propuesta de Harris (1954) plantea esta posibilidad. Él propone que dentro de un discurso se puede analizar la ocurrencia relativa de cualquier parte de una oración con respecto a otra dentro de la vecindad del mismo discurso; y este análisis nos puede llevar a encontrar regularidades. Pero ¿qué pasa cuando se extiende este análisis no sólo a los elementos de un discurso sino a aquellos presentes en diferentes discursos? Al respecto Harris (1954) dice que: «such regularities (and meaning) will not extend from one dicourse to another (except to another in some relevant way to the first, e.g. succesive lectures of a series)» (Harris, 1954, p. 15). Este comentario hecho por el autor resulta de primordial importancia para los objetivos de esta tesis. Es claro, conforme a lo que se men-

ciona en primera instancia, que dos elementos discursivos distintos no compartirán regularidades distribucionales; sin embargo, cabe preguntarse la relevancia de un discurso a otro dado en que se comparte una temática. En otras palabras, ¿la estructura distribucional de determinados elementos lingüísticos entre un documento y otro con temática similar al primero es regular en ambos?

Un ejemplo tomado de este mismo autor deja ver que la respuesta a esta pregunta bien puede ser afirmativa. Dice Harris (1954) que en una oración del tipo «He examinado mis ojos con el mismo *oculista* por veinte años», la palabra *oculista*, si bien puede ser remplazada por *doctor de ojos* (*eye-doctor*), la introducción de una palabra como *abogado* resultaría ciertamente anómala y poco frecuente. Se puede inducir por esto que el contexto determina, en cierta medida, el uso de ciertos elementos lingüísticos (en este caso, palabras); es decir, la estructura distribucional está ligada en gran medida con el significado (o, si se quiere, con la temática).

Con base en las ideas propuestas por Harris (1954) aquí discutidas, dentro de la lingüística computacional se ha creado un recurso que se ha dado en llamar espacios semánticos.

Espacios semánticos

La propuesta de los espacios semánticos planteadas en Brychcín and Koponík (2013), Lund and Burgess (1996), Hare et al. (2008), entre otros, consiste en la idea central de que el significado de una palabra se relaciona con el contexto en que tal palabra es usada. Bajo este supuesto, se concluye que «it is possible to compute semantic similarity of words by a statistical comparison of their contexts» (Brychcín and Koponík, 2013, p. 194). Este punto resulta altamente relevante, pues se tiene la suposición de que documentos similares se enmarcan en contextos similares y, por tanto, comparten elementos que se comportan distribucionalmente de la misma forma.

La metodología principal para los espacios semánticos consiste en la representación del lenguaje en matrices de co-ocurrencia (Lund and Burgess, 1996), en donde se comparan principalmente las palabras con los contextos en los que aparecen. Pero, como plantea Brychcín and Koponík (2013), la principal divergencia entre los modelos de espacios semánticos que se proponen radica en la selección de lo que se toma como contexto, pues éste puede ser la oración, el párrafo, el documento o algún concepto más amplio.

A continuación se presentan algunos de los modelos de espacios semánticos

propuestos con la finalidad de aclarar dicho concepto, además de comparar las metodologías.

- Hyperspace Analogue to Language (HAL) : Crea un espacio semántico a partir de la co-ocurrencia de palabras y a un grupo de palabras que se le llama «ventana». Las palabras en una ventana se pesan a partir de una distancia y entre más cercanas se encuentren, se asume un mayor impacto en las características semánticas de una palabra central.

- Correlated Ocurrence Analogue to Lexical Semantics (COALS) : Modelo de espacio semántico basado en las ideas de la Semántica Latente (LS) (Lauander et al., 1998). Se construye una matriz de forma similar a HAL, pero no se hace distinción entre palabras co-ocurrentes a una palabra central; además utiliza una Descomposición en Valores Singulares (SVD), con base a LS.

- Indexación Aleatoria: Se basa en el proceso de acumulación de vectores contextuales de una palabra que co-ocurre. Primero genera vectores contextuales a partir de los cuales que crea la matriz (Brychcín and Koponík, 2013).

- Bound Encoding of the AggreGate Language Enviroment (BEAGLE): Modelo que construye un espacio semántico de forma similar a la indexación aleatoria. En los vectores que se crean aleatoriamente se dan valores que van de acuerdo con una distribución gaussiana. El valor de la media es 0 y la varianza es igual a $\frac{1}{D}$, donde D es una dimensión (generalmente $D = 1024$) (Brychcín and Koponík, 2013).

- Energía Textual (ENERTEX). Propuesta basada en la física estadística que ve las palabras como *spines*, donde el comportamiento de una palabra se determina por las palabras vecinas a través de la fórmula $E = \frac{1}{2}(S \times S^t)^2$, donde S es una matriz de co-ocurrencia de palabras y S^t es su matriz transpuesta (Fernández et al., 2007).

Estos son sólo algunos de los modelos de espacios semánticos que se han propuesto en la literatura. En todos ellos se crea una matriz (aunque con métodos distintos) de donde se observan las relaciones que mantienen las palabras y los contextos. El planteamiento del contexto parece quedar de lado, asumiendo que un buen contexto es el documento; sin embargo, esto debe cuestionarse y,

por tanto, se pensó en utilizar más de un contexto disponible entre documentos. Para esto, se decidió tomar cada documento como un contexto, en una primera propuesta, y, en una segunda propuesta, el contexto estaba limitado por la oración en que se encuentra la palabra (véase el Capítulo 3). A continuación se muestran dos ejemplos de espacios semánticos que se utilizarán más adelante.

Uso de espacios semánticos

Bajo el supuesto de espacio semántico, se puede crear una matriz en donde se representan las co-ocurrencias de términos dentro de un contexto. Siguiendo lo mostrado en la literatura se decidió tomar como contexto el documento completo, primero; de igual forma, se plantea la implementación de la descomposición en valores singulares (SVD).

1. Contexto a nivel documento

Generalmente, se asume, como primera intuición, que al buscar similitud entre documentos, la selección del documento como el contexto es afortunada. A partir de tal idea, se creó una matriz de frecuencias de co-ocurrencias de términos dentro de los documentos. El primer paso fue extraer los términos relevantes por cada documento a partir de una ponderación con tf-idf (expuesta en la Sección 3.2.2). De esta forma, se obtuvo una matriz que relacionaba las co-ocurrencias de un término en los documentos, como se puede ver en la Figura 2.5.

$$\mathbb{X}_{m,n} = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & \cdots & d_n \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}$$

Figura 2.5: Ejemplo de matriz conformada por co-ocurrencias de términos en documentos.

Creada la matriz, el siguiente paso fue representar los documentos de tal forma que se pudiera determinar una similitud entre pares de documentos. Para esto se determinó una distancia de coseno, la cual ya se ha descrito en la Ecuación 2.5. Cada vector de la matriz, así, estuvo conformado por estos términos y su co-ocurrencia entre documentos; esta idea no está muy lejos de la propuesta

de los *espacios euclidianos*[5].

2. Matriz de documentos con SVD

Se considera que la mera representación de los términos dentro de una matriz de co-ocurrencias podría verse como un modelo bastante pobre. Se propone, también implementar un modelo de reducción en valores singulares (descrita en Lauander et al. (1998), Rohde and Plaut (2004) y Brychcín and Koponík (2013). En general, el método aquí utilizado es muy similar al que propone la *semántica latente*[6].

Se asume que una matriz $\mathbb{X}_{m,n}$ como la anterior puede presentar relaciones que contengan ruido, así como redundancia y ambigüedad. Con la descomposición en valores singulares o SVD, lo que se pretende es crear un subespacio que contenga sólo asociaciones "semánticas"significativas (Lauander et al., 1998). Para esto se hace una reducción de la dimensión de la matriz. Así:

$$\mathbb{X}_{m,n} = \sum_{i=1}^{r} D_i S_i T_i = \begin{pmatrix} d_1 \cdots d_r \end{pmatrix} \begin{pmatrix} s_1 & & \\ & \ddots & \\ & & s_r \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_r \end{pmatrix}$$

Es decir, la descomposición en valores singulares *descompone* la matriz en tres submatrices donde r es el rango de la matriz $\mathbb{X}_{m,n}$, generalmente r corresponde a $min(n, m)$. Aquí, D es un vector horizontal singular que representa los documentos de la matriz, mientras que T es un vector vertical que mejor representa los términos de la matriz original; mientras S es una matriz diagonal de valores singulares.

Dada la descomposición en valores singulares se puede reconstruir la matriz original con los valores singulares para que así los valores de la matriz original sean ahora más representativos en la matriz reconstruida. Señala Wang and Nie (2003) que si se truncan los valores singulares, manteniendo así sólo los k términos más representativos, la matriz resultante es una buena aproximación a la matriz original:

$$\mathbb{X}_{m,n} = D_r S_r T_r^t \approx D_k S_k T_k^t$$

Con esto en mente, para la reconstrucción de la matriz se tomó una dimen-

[5]De hecho se considera que un modelo como este es un *espacio euclidiano* que trata de ponderar las relaciones de los términos dentro de los documentos.

[6]El modelo que aquí se presenta sigue al pie de la letra la metodología propuesta por Lauander et al. (1998) y el cual los mismos autores llaman *semántica latente*; sin embargo, aquí se ha preferido utilizar el término más técnico de descomposición en valores singulares.

sión k situada *a priori* en 4, pero con valores variables. De tal forma que el primer paso para la reconstrucción fue reducir los valores singulares de la matriz diagonal S a una matriz de $k \times k$. Después se creó una matriz Q de tamaño $n \times m$ con valores nulos, es decir, con valores 0:

$$Q_K = \begin{pmatrix} 0_k & 0_k & \cdots & 0_k \\ 0_k & 0_k & \cdots & 0_k \\ \vdots & \vdots & \ddots & \vdots \\ 0_k & 0_k & \cdots & 0_k \end{pmatrix}_{(n \times m)}$$

Se diagonalizó la matriz original con Q dimensionada a $k \times k$:

$$Q_{(k,k)} = diag(X)$$

Y se redimensionaron las matrices D de la siguiente forma:

$$D_k = D_n \times k$$

Y se hizo lo mismo con T:

$$T_k = k \times D_m$$

Finalmente, se recontruyó una matriz $\hat{X}$, tal que:

$$\hat{X} = D_k \times Q_{(k,k)} \times T_k$$

La matriz resultante ya no conserva los valores originales. Un ejemplo de la matriz resultante del proceso de SVD y su reconstrucción es:

$$\hat{X} = \begin{array}{c} \\ \mathrm{t}_1 \\ \mathrm{t}_2 \\ \mathrm{t}_3 \\ \mathrm{t}_4 \\ \mathrm{t}_5 \\ \mathrm{t}_6 \\ \mathrm{t}_7 \\ \mathrm{t}_8 \\ \mathrm{t}_9 \\ \mathrm{t}_{10} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathrm{d}_1 & \mathrm{d}_2 & \mathrm{d}_3 & \mathrm{d}_4 & \end{array} \\ \begin{pmatrix} 0.01 & 0.01 & 0.01 & 0.01 & \cdots \\ 0.02 & 0.02 & 0.02 & 0.02 & \cdots \\ -0.01 & -0.01 & -0.01 & -0.01 & \cdots \\ -0.03 & -0.03 & -0.03 & -0.03 & \\ 0 & 0 & 0 & 0 & \cdots \\ 0.86 & 0.86 & 0.86 & 0.86 & \cdots \\ 1.07 & 1.07 & 1.07 & 1.07 & \cdots \\ 1.04 & 1.04 & 1.04 & 1.04 & \cdots \\ 0.02 & 0.02 & 0.02 & 0.02 & \cdots \\ -0.02 & -0.02 & -0.02 & -0.02 & \cdots \end{pmatrix} \end{array}$$

La descomposición en valores singulares es un método que ha cobrado popularidad dentro de la lingüística computacional (puede verse Lauander et al. (1998)). Su ventaja radica en que no sólo se queda con los términos en bruto, sino que a partir de un modelo matemático busca ponderar los términos más representativos; esto, además, lo hace tomando en cuenta todos los términos presentes en todos los documentos. En principio, parecería que este es un método que representa bien las ideas planteadas en este trabajo. Empero, se propondrá un modelo nuevo basado también en espacios semánticos para tratar de obtener mejores resultados en un proceso de agrupamiento.

2.3. Consideraciones sobre los métodos

Los métodos aquí presentados abordan desde perspectivas diferentes el problema de la similitud en el lenguaje; si bien parecen centrarse en la similitud entre palabras, es posible utilizar estos métodos para identificar similitud entre documentos. Ya se ha dicho que los modelos basados en alineamiento textual no capturan más allá de rasgos gráficos y, en menor medida, fonológicos. Los métodos computacionales basados en distancia de vectores han sido ampliamente utilizados en la literatura sobre agrupamiento; sin embargo, suelen pasar por alto el análisis de las características lingüísticas de los documentos.

Las taxonomías tratan de mostrar una buena representación de las relaciones léxico-semánticas que se generan en los documentos; sin embargo, como se ha dicho, esta perspectiva depende de un recurso que es exhaustivo y no garantiza

el reflejo de una buena categorización taxonómica. Por tanto, se ha descartado este método para la realización de los experimentos de esta tesis, pues requeriría demasiado esfuerzo realizar una taxonomías para cada una de las áreas que aquí se proponen trabajar.

El modelo de teoría de la información que propone Lin (1998) captura bastante bien las propuestas de Tversky (1977) y Goldstone (1994) sobre lo que es la similitud. En principio, parece una buena apuesta para un método automático de similitud; su defecto, empero, radica en la generalidad con la que se plantea, puesto que no se precisan los argumentos que la función requiere; así, su desempeño dependerá en gran medida de la elección que se haga de los argumentos. En el siguiente capítulo se mostrará una propuesta para este algoritmo.

Por último, la hipóstesis distribucional y los espacios semánticos han tenido gran auge en los últimos años y se han visto como buena propuesta para encontrar relaciones de similitud y relaciones léxicas (Ali, 2011; Brychcín and Koponík, 2013; Hare et al., 2008; Lund and Burgess, 1996; Padó and Lapata, 2003; Sahlgren, 2008). El modelo de la semántica latente, por ejemplo, ha sido utilizado a nivel de documento (Lauander et al., 1998). Se hace notar, además, que este modelo se acopla bastante bien al modelo de espacio vectorial que proponen los métodos de recuperación de información. En los capítulos siguientes se compararán tres métodos basados en este modelo, principalmente haciendo la distinción entre el contexto de elección: sea este el documento y la oración.

3

Metodología

Ya con el concepto de agrupamiento y de similitud textual aclarados, se determinó la siguiente secuencia para la conformación del método:

1. Adecuación de un corpus de prueba dividido temáticamente, que funcionó para realizar las pruebas necesarias que llevaron a la conformación del algoritmo propuesto.
2. Análisis y selección de los métodos de preprocesamiento (eliminación de palabras funcionales, *stemming* y pesado de términos).
3. Desarrollo de una propuesta de similitud textual con enfoque en detección de documentos temáticamente similares.
4. Implementación del algoritmo de agrupamiento de MajorClust.

3.1. Corpus de prueba

Para las pruebas se construyó un corpus con material general. Se obtuvieron diferentes documentos de Wikipedia, de los cuales se pudieran hacer grupos de diferentes temáticas. La cantidad de documentos por temática; además, debería de presentar cierta variación para que de esta forma se pudiera determinar un modelo que separara lo mejor posible los documentos en temáticas claramente diferenciadas (para posteriormente aplicarlo a corpus menos variados). Se seleccionaron diferentes categorías y se tomaron artículos que compartieran una misma temática; a estos documentos se les agregó otros obtenidos de diversas

fuentes para asegurarse que la relación de similitud no respondiera a patrones similares de redacción. Las temáticas se tomaron de forma aleatoria.

De esta forma se obtuvieron 100 documentos. Estos documentos eran de diferentes tamaños, aunque se consideraron documentos que no fueran demasiado pequeños; es decir, se trató que los documentos tuvieran más de 100 palabras. El promedio de palabras por documentos fue de 98.18. Este corpus se analizó con el sistema *Simple Concordance* para determinar su tamaño y otras características estadísticas. Se obtuvo un total de 110 887 tokens y 13 626 tipos; el promedio tipo/token fue de 0.123. El total de temáticas obtenidas fue de 30. La agrupación de estas temáticas se puede ver en el Cuadro 3.1.

Tabla 3.1: Cantidad de documentos por temática dentro del corpus

Temática	**Núm. textos**	**Temática**	**Núm. textos**
Acústica	4	Fractal	2
Amor	2	Humano	4
Espectáculos	2	IFAI	3
Bioinformática	2	Matemáticas	2
Buchnera aphidicola	2	Perros	2
Capoeira	2	Política	2
Célula	7	Helicobacter pylori	3
Ciencia histórica	3	Religiones	10
Cinematografía	3	Sacarosa	2
Coca-cola	4	Salmonella	3
Condensadores	6	Sociedad	5
Mecánica cuántica	3	Sushi	3
Deportes	3	Series de Taylor	3
Escherichia coli	3	Twitter	3
Economía	4	Xbox	3

3.2. Preprocesamiento

En esta sección se describe el preprocesamiento propuesto, mostrando el porqué de la selección de los métodos seleccionados. Se describe, en primer lugar, la validación de términos y, finalmente, se discuten dos métodos de *stemming* y se muestra como su implementación logra mejorar los resultados obtenidos.

3.2.1. Eliminación de palabras funcionales

Por considerarse innecesarias para los propósitos del trabajo, se eliminaron las palabras funcionales del corpus, es decir, las preposiciones, los artículos y formas clíticas. Este proceso también se realizó de manera automática y se hizo a partir de una lista de paro; es decir, una lista con las preposiciones, artículos y clíticos en español.

Antes de aplicar la lista de paro, se consideró pasar todos los documentos a minúsculas, para que así la computadora no tuviera problemas en determinar las apariciones de ítems léxicos iguales.

Con la eliminación de palabras funcionales, se reduce el ruido al momento de aplicar el conteo de palabras para la aplicación de los algoritmos de *stemming* y para el conteo de las co-ocurrencias de palabras entre documentos, puesto que las palabras funcionales no aportan información temática relevante.

3.2.2. Validación de términos

El determinar un peso para las palabras basado en su aportación de información al documento es un paso importante en una tarea de similitud temática entre documentos. Se asume que la temática de un documento se puede determinar en gran medida por la aparición de ciertos términos; en otras palabras, el uso terminológico dentro de un documento viene condicionado en gran parte por la temática del documento. Por tanto, es necesario determinar un modelo que permita asignar pesos a las palabras de acuerdo con su aportación de información terminológica. Sobre esto se han propuesto diferentes aproximaciones, un ejemplo se puede ver en Frantzi et al. (2000).

Para los propósitos específicos de esta tesis se eligió un modelo para medir el peso de las palabras, de tal forma que se pudiera determinar la información aportada por cada una de ellas. Un modelo sencillo para la determinación del valor cuantitativo de un término es el modelo del tf-idf, que se expone en (Manning and Schutze, 2000, p. 543). Este se puede describir formalmente como:

$$tfidf_{t_i} = 0.5 + \frac{0.5 \cdot f(t_{i,d})}{max_t[f(t_d)]} \times log_2(\frac{|N|}{f(t_{i,N})})$$

Donde t_i es el término i en el documento, $f(t_{i,d})$ es la frecuencia de ese término en el documento, $max_t[f(t_d)]$ es la palabra con mayor frecuencia en el documento, $|N|$ es el total de documentos en la colección y $f(t_{i,N})$ es la frecuencia del término dentro de la colección. Supóngase que se quiere conocer

el tf-idf de un término i dentro de un documento que pertenece a un corpus de 10 documentos, en donde dicho término aparece 5 veces. En este documento específico del corpus, el término aparece 40 veces mientras que el término con mayor frecuencia dentro de ese documento apareció 160 veces. Entonces tenemos que $f(t_{i,d}) = 40$, $max[f(t_d)] = 160$, $|N| = 10$ y $f(t_{i,N}) = 5$, su tf-idf sería:

$$tfidf_{t_i} = 0.5 + 0.5 \cdot \frac{40}{160} \times log_2(\frac{10}{5}) = \frac{5}{8}$$

Se seleccionó el algoritmo de tf-idf por ser más sencillo y menos costoso computacionalmente. Se probó en los documentos del corpus de prueba para determinar pesos a las palabras con mayor información, es decir los términos. Algunos de los resultados obtenidos con este algoritmo (sin ningún tipo de *stemming*) se muestran en la Tabla 3.2.

Tabla 3.2: Primeros 5 palabras con mayor tf-idf para 4 documentos del corpus

Temática	*Amor*	*Humano*	*Matemáticas*	*Fractal*
Término 1	Amor	Homo	Plural	Fractal
Término 2	Compasión	Sapiens	Griego	Fractales
Término 3	Egoísta	Conscientes	Razonamiento	Objeto
Término 4	Comportamiento	Pasado	Números	Estructura
Término 5	Acciones	Género	Mathematica	Autosimilar

Se observa que, en general, el término con mayor tf-idf es el que mejor representa la temática de cada grupo. Pero se notó, también, que al no realizar ningún tipo de *stemming* anterior a la validación de los términos, se pueden dar casos donde formas morfológicamente distintas de una misma palabra se tomen como dos términos distintos; tal fue el caso de la separación de plurales o de formas con morfología masculina y femenina, que se tomaron como elementos distintos; esta falta morfológica provocó que el pesado de términos tuviera ciertos desajustes. Por tanto, se consideró importante reducir las formas de las palabras a una base a partir de un algoritmo de *stemming*. En el siguiente apartado se discuten dos de estos métodos y se comparan sus resultados.

3.2.3. Stemming

En esta sección se presentan dos modelos de *stemming*, los cuales tienen como objetivo conjuntar las formas léxicas similares al reducir su morfología. En primer lugar, se muestra el algoritmo de Porter Porter (1980), que es uno de los más utilizados en la literatura. En segundo lugar, se presenta un método propuesto recientemente: el *ultra-stemming* Torres-Moreno (2012).

Algoritmo de Porter

Uno de los métodos más utilizados dentro de la lingüística computacional para obtener el radical de la palabra ha sido el algoritmo de Porter (Porter, 1980). En principio, este algoritmo está pensado para el inglés, pero se han hecho adaptaciones del mismo a diversos idiomas, entre ellos el español. Este algoritmo se basa en un lexicón de sufijos y reglas contextuales que permiten remover de una palabra sus afijos.

Lo primero que define el algoritmo de Porter son los elementos con los que van a trabajar las reglas. Define como una *consonante* todos aquellos elementos diferentes de *a*, *e*, *i*, *o* y *u* (que para el español incluiría las versiones con tilde). Con esto en mente, las palabras o partes de éstas se definirán entonces como:

CVCV... C
CVCV... V
VCVC... C
VCVC... C

Donde C representa consonantes y V vocales. Se define, además, una *medida* de cada palabra o parte de palabra que se representa por $m \in \{0, 1, 2\}$; $m = 0$ representa una forma del tipo «árb»; $m = 1$, una raíz del tipo «árbol»; y $m = 2$ una palabra del tipo «árboles». Las reglas para la eliminación de los elementos sufijales se declaran, en este algoritmo, de la forma (*condition*) $S_1 \rightarrow S_2$. Así, el algoritmo de Porter declara las reglas necesarias para encontrar los sufijos de una lengua; esto lo hace dependiente de la lengua, ya que el mismo algoritmo tiene que ser modificado para que pueda trabajar con lenguas diferentes. También representa un conocimiento previo de las reglas morfológicas que dominan la sufijalidad en determinada lengua. Una versión del algoritmo de Porter para español se encuentra en el Apéndice A.

En general, al aplicar el algoritmo de Porter, se soluciona la mayor parte de los problemas de agrupamiento de formas de palabras, por ejemplo, la agrupación de formas como *fractal* y *fractales*. Como se observa en el Cuadro 3.3, las palabras con mayor tf-idf siguen siendo representativas aunque presentan variaciones con respecto al pesado sin ningún tipo de *stemming*; de igual forma, algunas palabras siguen coincidiendo; además se nota que se agrupan diferentes formas de palabras lo que representa una ventaja para la aplicación de los algoritmos de similitud textual.

Tabla 3.3: Resultados de la aplicación de tf-idf con aplicación del algoritmo de Porter.

Temática	*Amor*	*Humano*	*Matemáticas*	*Fractal*
Término 1	Amor(-es)	Homo	Plural(-es)	Fractal(-es)
Término 2	Compasión	Sapiens	Grieg(-o/-os)	Autosimilar(-es)
Término 3	Egoísta(-s)	Aprend(-er/-iz/...)	Número(-s)	Autosimil
Término 4	Comport(-ar/-miento/...)	Pertenec(-er/-e/...)	Bourbaki	Escal(-a/-ar)
Término 5	Emocion(-ante/-es/...)	Mental(-es)	Axiom(-a/-as/-ático)	Mandelbrot

Ultra-stemming

Torres-Moreno (2012) propone un modelo de *ultra-stemming* basado en el corte de palabras por caracteres, reduciendo cada una de las palabras en un documento a sus n caracteres iniciales. En principio, no se plantea cuál es el número de caracteres que se deben de preservar para que, en la medida de lo posible, se conserve el radical de la palabra. El corte en n caracteres se puede justificar con base en las propiedades prosódicas del español. El español suele tener una gran cantidad de palabras bisilábicas, pues «una palabra ha de contener necesariamente un pie y las condiciones de buena formación requieren que ese pie sea binario (bajo análisis silábico o moraico) así pues, la palabra prosódica mínima ha de ser binaria» (Saceda, 2005, p. 13). La conformación de cada una de las sílabas exige entre 2 y 3 caracteres en español, por lo que la elección para realizar el corte recayó entre 4 y 6 caracteres.

De igual forma que con el algoritmo de Porter, con la aplicación de *ultra-stemming* las palabras con mayor información son las que mejor representan a cada temática. No se observa una variación representativa en la selección de los términos más pesados de cada documento; si bien se puede determinar que con el algoritmo de *ultra-stemming* se hace una mejor selección de éstos, aunque no en gran medida. Se nota que con *ultra-stemming* se agrupan variaciones que corresponde a derivativos de las palabras; sin embargo, esto puede tomarse como una ventaja si se trata de la agrupación temática, pues aunque en categoría de palabra sean distintas, el contenido semántico que desean comunicar, por ejemplo, dos palabras como *egoísta* y *egoísmo* es la misma.

Finalmente, se optó por utilizar el *ultra-stemming* pues éste resultó ser menos costoso en tiempo de procesamiento. Mientras que el algoritmo de Porter necesitó en promedio 10.5 segundos para procesar los documentos, *ultra-stemming* solamente tardó 0.9 segundos. Se nota, además, que los resultados entre uno y otro algoritmo no varían mucho.

En general, el pre-procesamiento consistió en que las palabras del corpus

Tabla 3.4: Resultados de la aplicación de tf-idf con aplicación de ultra-stemming

Temática	*Amor*	*Humano*	*Matemáticas*	*Fractal*
Término 1	Amor(-es)	Homo	Plural(-es)	Fractal(-es)
Término 2	Egoís(-ta/-tas/-mo)	Sapien(-s/-te)	Matemá(-tico/-ticas)	Autosi(-milar—es/-mil)
Término3	Altrui(-smo/-sta)	Géner(-o/-os)	Mathema	Escalar(-r/-res)
Término 4	Comport(-ar/-amiento/...)	Pasado(-s)	Griego(-s)	Newton(-iano)
Término 5	Emocio(-nante/-nes/...)	Homín(-ido)	Campos	Algori(-tmo—s)

fueron pasadas a minúsculas y la eliminación de las palabras funcionales para la realización de los experimentos de similitud textual, además de la validación de términos y la aplicación del algoritmo de *stemming*.

3.3. Propuesta de un espacio semántico basado en oraciones

Los métodos mencionados en el Capítulo 3 han sido probados en la literatura con diferentes resultados (Ali, 2011; Brychcín and Koponík, 2013; Lin, 1998; Lund and Burgess, 1996; Padó and Lapata, 2003; Rohde and Plaut, 2004; Torres-Moreno et al., 2010); entre ellos, los modelos basados en espacios semánticos han mostrado alta efectividad. Sin embargo, pocos han sido los modelos de espacios semánticos que se han alejado de un contexto a nivel oracional. Aquí se propone tomar en cuenta un contexto a nivel oracional[1] para capturar también relaciones de términos. Se decidió tomar el valor de contexto para la elaboración de una matriz de términos, que permitiera el cálculo de una métrica de similitud textual. Pero, de esta forma, la creación de la matriz se vuelve más compleja, puesto que ahora se necesitan comparar, no los documentos entre sí, sino las oraciones de cada documento contra las oraciones de todos los documentos del corpus. En principio, se pensó en seleccionar sólo los términos relevantes; sin embargo, al hacerlo, por razón del contexto (es decir, por el tamaño de las oraciones), se corría el riesgo de eliminar todos los elementos de un contexto y quedarse con un conjunto vacío, pues no en todas las oraciones se encontraban términos relevantes. Por tanto, se tomó la decisión de no eliminar palabras en los contextos (mas que funcionales). De esta forma se construyó una matriz que proporciona mejores relaciones entre documentos; en las Figuras 3.1 y 3.2 se puede ver una matriz más poblada cuando los documentos son similares y despoblada cuando estos no lo son, respectivamente.

[1]Entendido oración todo aquello que se encuentre entre punto y punto.

$$\mathbb{X}_{m,n} = \begin{array}{c} \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} D_1/D_2 & O_1 & O_2 & O_3 & O_4 & O_5 & \cdots \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 2 & 2 & 0 & \cdots \\ 1 & 1 & 1 & 1 & 2 & \cdots \\ 0 & 1 & 1 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

Figura 3.1: Matriz conformada por co-ocurrencias de términos en oraciones de dos documentos similares.

$$\mathbb{A}_{m,n} = \begin{array}{c} \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} D_1/D_2 & O_1 & O_2 & O_3 & O_4 & O_5 & \cdots \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

Figura 3.2: Matriz conformada por co-ocurrencias de términos en oraciones de dos documentos diferentes.

En general, las matrices de documentos que compartían una temática similar reflejaban mayor co-ocurrencia de palabras. Sin embargo, debe tomarse en cuenta que al conservar todas las palabras, se corre el riesgo de que estas palabras no sean relevantes para el contenido temático. Por tanto, es importante tomar en cuenta una forma de dar peso a los términos relevantes. Para esto, se consideró adecuado el uso de una ponderación de términos (tf-idf, expuesto en la Sección 3.2.2). Así, en lugar de utilizar valores absolutos, las matrices se rellenaron por un valor de peso representado por $n \cdot tf - idf$, donde n representa el valor absoluto. Así, las palabras con mayor contenido temático influyen más en la medida de similitud que aquellas que contienen poca información temática.

En este caso, no se podían reducir las matrices a vectores que pudieran tomar una métrica como la distancia de coseno, pues ya no se trata de una matriz de dos dimensiones, sino de tres. Por tanto, se definió una medida de similitud que representara la co-ocurrencia de las palabras dentro de las oraciones. Para esto se tomó como base la *medida cluster*, expuesta en la Ecuación 2.6, de tal forma

que se pudieran comparar los vectores de los documentos d_1 y d_2, de la siguiente forma:

$$cluster = \frac{\overrightarrow{d_1} \cdot \overrightarrow{d_2}}{||d_1||}$$

Donde $||d_1||$ es la norma del vector de palabras del primer documento. Pero dentro de la matriz se requería comparar cada documento contra todos los documentos del corpus. De esta forma, para el iésimo documento, se obtuvo:

$$sim_{d_i,d_j} = \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_2}}{||d_i||} + \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_3}}{||d_i||} + \cdots + \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_j}}{||d_i||}$$

Lo que se puede abreviar como:

$$sim_{d_i,d_j} = \sum_{j=1} \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_j}}{||d_i||}$$

Finalmente, para obtener la métrica final, se consideró promediar dividiendo por el total de documentos con los que se compara el iésimo documento; es decir, la cardinalidad la columna m de la matriz $\mathbb{X}$[2], que se puede representar como $|\mathbb{X}_m|$, tal que:

$$sim_O = \frac{\sum_{j=1} \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_j}}{||d_i||}}{|\mathbb{X}_m|}$$

En este caso, el vector documento $\overrightarrow{d}$ está formado por una matriz de oraciones dentro de la matriz de documentos $\mathbb{X}$. Con esta medida, los resultados que se obtienen no están normalizados y van de 0 a n; sin embargo, se observó una tendencia en donde los documentos más parecidos tienen un valor mucho más alto que aquellos documentos disímiles.

En general, este método busca encontrar relaciones al nivel de la oración, de tal forma que la similitud entre dos oraciones de dos documentos distintos pueda, asimismo, ser parte de la similitud entre dos documentos. Esto presenta una gran dificultad en cuestión de procesamiento computacional, pues el método requiere de más análisis que los basados en documentos, pues se parte de un contexto pequeño (la oración) para determinar un valor de similitud entre documentos. Sin embargo, se cree que es esto mismo lo que puede dar una mejor

[2]Ya que $\mathbb{X}_{m,n}$ es una matriz que compara todos los documentos en la columna m contra todos los documentos en la fila n, se tiene que $|m| = |n|$

representación de las relaciones entre términos, pues ya no se ve su distribución a un nivel amplio (como el documento) sino a un nivel de mayor precisión.

3.4. Implementación del algoritmo de agrupamiento

Ya se ha mencionado que para el agrupamiento de documentos se propone la utilización de MajorClust, el cual ya se ha expuesto en la Sección 1.9.3. De esta forma, se propone ϕ como la función de similitud expuesta en 3.3 y que se basa en un espacio semántico con contexto a nivel de oración. Así, el modelo de similitud textual propuesto en la sección anterior alimentará el algoritmo de MajorClust, para crear el grafo que determinará la cercanía o lejanía entre documentos.

En el grafo expuesto en la Figura 1.4 se pueden visualizar los documentos como los puntos o nodos separados por aristas cuya distancia estará determinada por la medida obtenida a partir del modelo de similitud textual propuesto. De esta forma, si los documentos son temáticamente similares, las aristas en el grafo serán más cortas; mientras que si son diferentes, las aristas se harán más largas. Con el grafo así determinado, el algoritmo de MajorClust buscará los grupos pertinentes de documentos y el resultado de la aplicación del algoritmo será, precisamente, grupos de documentos que temáticamente serán similares.

3.5. Diagrama del método

Expuesto lo anterior, a continuación se describe el método que se propone para el agrupamiento temático de documentos a partir del modelo de similitud textual basado en espacios semánticos y con contexto a nivel de oración; asimismo, se presenta un diagrama que resume el método.

Como se ve en la Figura 3.3, en un marco general, se requiere de un corpus de documentos (el cual, no requiere de ningún procesamiento y, en principio, puede estar en cualquier lenguaje); el corpus es preprocesado con la eliminación de formas funcionales (preposiciones, conjunciones, clíticos, etcétera), la reducción de sus formas morfológicas con *ultra-stemming* y la ponderación de términos con tf-idf. De aquí se obtiene un corpus preprocesado, el cual ingresa al algoritmo de agrupamiento, MajorClust, que se vale del modelo de similitud textual aquí propuesto. Esto devuelve como resultado los grupos de documentos

obtenidos, en los cuales se representan las temáticas expuestas en los corpus. En el próximo capítulo, se evaluará el algoritmo propuesto y se comparará el modelo de similitud textual basado en oraciones con la implementación de otros modelos de similitud propuestos en la literatura.

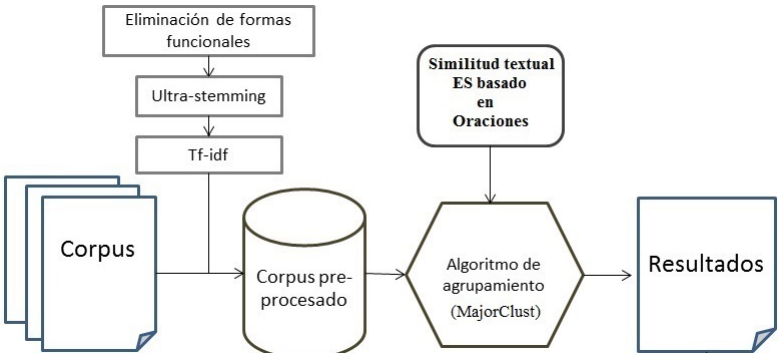

Figura 3.3: Diagrama del agrupamiento de documentos a partir de modelos de similitud textual

4

Análisis y evaluación

En este capítulo se presenta la evaluación de la metodología propuesta en el Capítulo 3 para el agrupamiento temático de documentos. Para mostrar la validez de las propuestas planteadas, se compara el modelo de similitud textual propuesto en la Sección 3.3 con la implementación de los algoritmos propuestos en la literatura, más específicamente, el modelo de Lin (1998), el uso de una distancia (distancia de coseno), y dos modelos distribucionales: uno simple basado en co-ocurrencias a nivel documento, y uno inspirado en el análisis en *semántica latente*, es decir, con aplicación de descomposición en valores singulares.

Primero se presentará el corpus de evaluación, el cual es el Corpus de las Sexualidades en México; posteriormente se procederá a mostrar los métodos de evaluación; se describirán las implementaciones hechas con las que se pretende comparar la propuesta de la tesis y finalmente se mostrarán las evaluaciones correspondientes, así como consideraciones sobre el desempeño en cuestión de tiempo computacional.

4.1. Corpus de evaluación

Dada la necesidad de un corpus para el análisis y la realización de experimentos, se especificaron determinadas características necesarias. En primer lugar, se requería que el corpus se encontrara dividido temáticamente para poder evaluar los resultados obtenidos por un método automático. En segundo lugar, se buscó que el corpus fuera especializado, es decir, que abarcara textos que pertenecieran a una misma área de investigación con diferencias temáticas propias

del área. En tercer lugar, se quería que el corpus fuera lo suficientemente grande para poder realizar un análisis y obtener resultados que fueran representativos. Para cumplir tales necesidades, se tomó el Corpus de las Sexualidades en México (CSMX), que es un corpus disponible en línea y que se describe en Sierra et al. (2009). Este corpus cuenta con 140 documentos de sexualidad dividido en 7 áreas temáticas[1]. La división temática del corpus se muestra en la Tabla 4.1.

Tabla 4.1: Cantidad de documentos por temática dentro del corpus

Temática	**Núm. textos**
Atracción sexual y relaciones entre individuos	20
Comportamiento sexual	21
Educación sexual, cultura y sexualidad social	20
Enfermedades de transmisión sexual (ETS)	20
Fundamentois biológicos de la sexualidad	19
Sexualidad variante	19
Respuesta y expresión sexual	20

Además, cabe resaltar que cada una de estas temáticas está, a su vez, subdividida dentro de 5 diferentes niveles de especialidad que van desde el académico hasta el poco especializado (obtenido de foros de discusión). Los detalles sobre la división temática y la división en niveles de especialidad se pueden ver con mayor detalle en Sierra et al. (2009).

El CSMX presenta un reto que se cree puede ser resuelto por medio de los modelos de similitud textual, especialmente en aquellos basados en espacios semánticos. Este corpus, por su especialización, puede presentar una distribución de términos bastante continua; sin embargo, se cree que hay elementos terminológicos que serán propios de cada área temática, si bien, al mismo tiempo, cada documento puede enfocarse a un subtema distinto (por ejemplo, en el caso de enfermedades de transmisión sexual, donde se puede hablar de distintas enfermedades dentro de la misma temática). A continuación se presenta un análisis preliminar de estos supuestos.

4.2. Análisis de la metodología propuesta

Para el corpus de prueba, es decir el CSMX, se siguió la metodología propuesta en el Capítulo 4. Aquí se analizan los pasos de esta metodología y los datos

[1]Las temáticas que se proponen fueron determinadas en base a diferentes criterios en la selección del corpus, estos se describen en Sierra et al. (2009).

obtenidos a partir del CSMX.

Eliminación de formas funcionales y ultra-stemming

En primer lugar, se aplicó al corpus la eliminación de formas funcionales y el *ultra-stemming* a seis letras. De esta forma, se pudo reducir notablemente la cantidad de palabras con las que debería de trabajar el sistema, lo que favoreció la rapidez del procesamiento y, asimismo, los resultados obtenidos. De las 314 679 palabras con las que contaba el CSMX, la eliminación de formas funcionales y la aplicación de *ultra-stemming* para agrupar fromas similares redujeron esta cantidad a 13 975 tipos de términos.

Tabla 4.2: Ejemplos de términos con *ultra-stemming* y las palabras que abarcan.

Términos	Palabras que abarca
sexual-	sexual, sexualidad, sexualidades, sexuales
erotis-	erotismo
enferm-	enfermedad, enfermedades, enfermos, enferma, enfermos, enfermas
especi-	especie, especies, especialidad, especialista, especial, especializado
reprod-	reproducir (y formas conjugadas), reproductivo, reproductiva, reproductivos, reproducción, reproductora, reproductividad

En la Tabla 4.2 se muestran algunos resultados obtenidos en el CSMX una vez aplicado el *ultra-stemming*. Se observa que en varios casos se agrupan palabras que están temáticamente relacionadas; sin embargo, como en el caso de *especi-* pueden presentarse agrupados a una misma forma truncada palabras que se relacionan poco temáticamente. Empero, tales circunstancias se presentaron en pocos casos.

Pesado de términos con tf-idf

Al principio de este trabajo se ha planteado la idea de que existen elementos que en gran medida indican características temáticas dentro de los documentos; estos elementos son los términos. Asimismo, se plantea que existen términos con mayor influencia temática. Se ha observado, además, que el peso de los términos puede ser capturado y cuantificado: esto se plantea con mayor detalle en la Sección 3.2.2.

Conforme a la idea de la distribución temática de los términos, aquí se presenta un análisis realizado para determinar la validez de esta hipótesis. En primer lugar, en la Tabla 4.3 se muestra la distribución de cinco términos a través de todas las temáticas; los números que se muestran son la cantidad de documentos en los que aparece cada uno de los términos dentro de cada temática.

Tabla 4.3: Distribución de términos por número de documentos en las temáticas del corpus

Temática/términos	sexual	erotis-	enferm-	especi-	reprod-
Atracción sexual y relaciones entre individuos	20	20	0	15	16
Comportamiento sexual	18	0	0	0	2
Educación sexual, cultura y sexualidad social	20	0	0	19	18
Enfermedades de transmisión sexual (ETS)	18	0	18	16	9
Fundamentois biológicos de la sexualidad	10	0	0	11	11
Sexualidad variante	19	0	0	15	4
Respuesta y expresión sexual	20	0	0	11	2

Se puede notar que, por ejemplo, el término *sexual* se distribuye uniformemente en la mayor parte del corpus (con pocos casos sólo en «Fundamentos biológicos de la sexualidad»); esto es, aparece en la mayoría de los documentos a través de las diferentes temáticas. Con términos como *erotism-* o *enferm-* no sucede lo mismo, sino que estos términos son representativos de la temática y sólo aparecen en documentos que se encuentran dentro de sus temáticas correspondientes. Por último, los términos *especi-* y *reprod-* tienen una distribución variante a través de los documentos de las diferentes temáticas.

Tabla 4.4: Términos con mayor tf-idf para documentos dentro de las temáticas

Atracción sexual y relaciones entre individuos	roland	bajtin	cinta	biogen-	apolog-
Comportamiento sexual	ego	aciert-	joint-	hombr-	avalos
Educación sexual, cultura y sexualidad social	doc	femine	follet-	sida	preoc-
Enfermedades de transmisión sexual (ETS)	signol-	orogen-	gravid-	pus	pelvia
Fundamentois biológicos de la sexualidad	links	socioh-	minell-	opcit-	robich-
Sexualidad variante	binocu-	brujul-	transb-	chimen-	reduzc-
Respuesta y expresión sexual	pesper-	futil	public-	flaca	local

Si bien, los documentos comparten varios términos (por ejemplo, *sexual*), para realizar un agrupamiento temático, se requiere dar relevancia a aquellos términos que son exclusivos de cada temática. Para tal objetivo, el pesado de términos con *tf-idf* es favorable, pues resalta los términos que son importantes para cada uno de los documentos, de tal forma que se da mayor peso a tales términos y aquellos que suelen aparecer a través de todo el corpus, como es el caso de *sexualidad*, son tomados con poca relevancia. En la Tabla 4.4, se puede ver algunos de los términos relevantes para algunos documentos de cada temática a partir de tf-idf.

Distribución de términos en la matriz de co-ocurrencias

Para la aplicación del modelo de similitud textual se necesitaba construir una matriz de co-ocurrencias de términos a través de los documentos. Es decir, en esta matriz se representa la cantidad de veces que un término aparece dentro de uno o más documentos. Se ha planteado la hipótesis de que las temáticas similares tendrán más términos en común. A continuación se presenta un ejemplo dentro del CSMX en donde se puede observar esta característica.

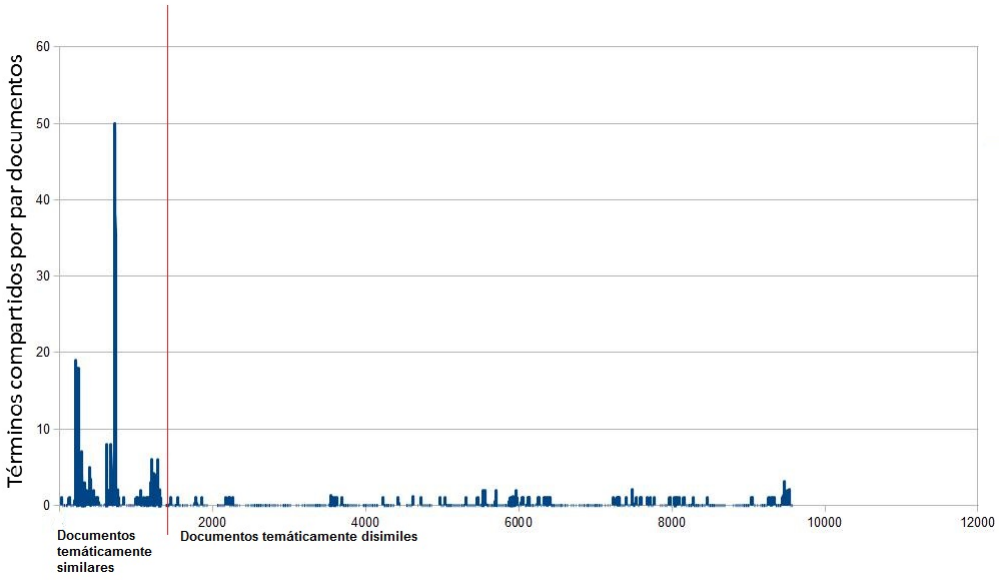

Figura 4.1: Cantidad de términos compartidos a través de documentos en función de su temática.

Si se toman las primeras 50 palabras con mayor tf-idf[2] se puede, en cierta medida, representar cada documento del corpus a partir de un arreglo con sus términos más relevantes. A partir de estos 50 términos se realizó un análisis de la distribución de dichos términos dentro de los diferentes documentos del corpus, llegando así a determinar qué cantidad de términos se comparten entre los documentos del corpus. De los 140 documentos, se pudo realizar el análisis en los 9 691 pares de documentos[3], de los cuales 1 312 pares compartían temática, mientras que los 8,279 restantes no.

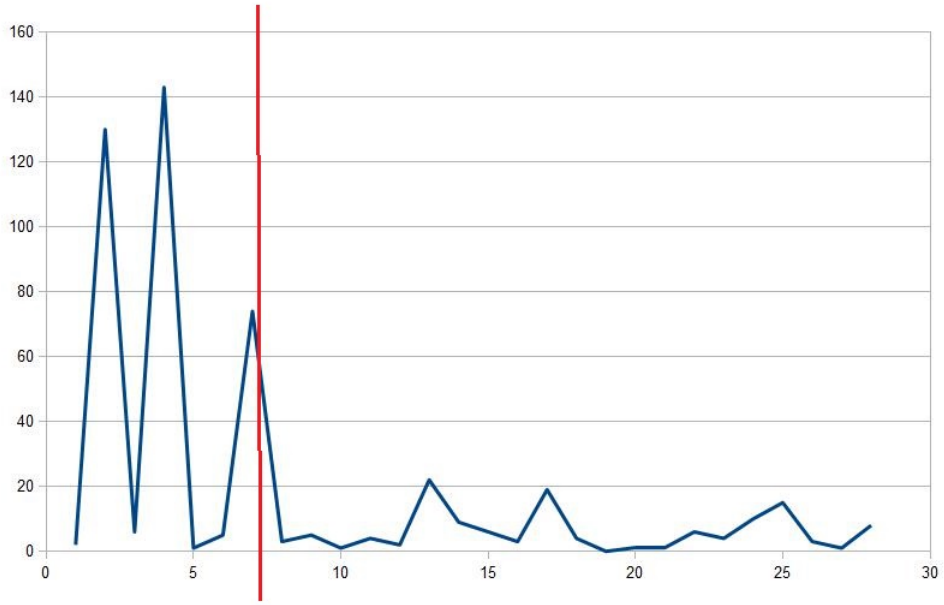

Figura 4.2: Cantidad de términos compartidos por las diferentes temáticas.

En la Figura 4.1 se pueden ver los resultados de este análisis; a la izquierda se encuentran los pares de documentos que comparten una misma temática, mientras que a la derecha los que no comparten temática. Se puede observar que dentro de los del lado izquierdo de la línea existe un mayor número de términos compartidos; mientras que en los otros documentos, los términos compartidos

[2]Se toman los primeros 50 términos por motivos prácticos, ya que, una vez aplicada la eliminación de formas funcionales, cada temática dentro del CSMX contaba con un promedio de 13 975 términos, lo que haría difícil la visualización. Para la creación de la matriz, sin embargo, no se puso esta restricción.

[3]El número de pares se puede determinar con la ecuación $C_2^n = \frac{n!}{2! \cdot (n-2)}$.

son mucho menores. Esto deja ver que los documentos que comparten temática también tenderán a compartir, en cierta medida, el uso de términos.

En la Figura 4.2 se puede ver la distribución terminológica a través de las temáticas; de igual forma que en la gráfica anterior, se determinó cuáles documentos pertenecían a la misma temática y se vieron cuántos términos comparten las temáticas; asimismo, se comparó todo el conjunto de documentos de una temática contra otra temática distinta. Del lado izquierdo de la línea vertical se encuentran las temáticas similares, mientras que del derecho las temáticas disímiles. Igual que con los documentos, las curvas son mucho más bajas cuando no se trata de la misma temática; sin embargo, si tenemos temáticas idénticas (plasmadas en diferentes discursos) los términos que se compartirán serán más altos.

Esto deja ver dos cosas: a) que, en efecto, los términos tienen determinada carga temática y, por tanto, intersectarán si se trata de textos con igual temática, mientras que esto no pasará si los textos tienen temáticas distintas; y b) que, por tanto, la temática de un área influye en el uso de los términos especializados, con lo que se pueden capturar similitudes temáticas entre uno y otro texto. Esto no está lejos de lo que apuntaba Harris (1954) al decir que ciertas regularidades (en este caso léxicas) no se extienden de un discurso a otro, a menos que dichos discursos estén relacionados de alguna forma (en este caso, a partir de la temática).

En general, el ejercicio anterior deja ver que se tendrá una mayor co-ocurrencia de términos cuando las temáticas sean similares. De esta forma, las matrices de documentos similares van a estar más pobladas. Tómese el ejemplo de los dos documentos cuya temática es «Atracción sexual y relaciones entre individuos». Parte de la matriz generada es la siguiente:

$$\hat{X} = \begin{array}{c} \\ o_1 \\ o_2 \\ o_3 \\ o_4 \\ o_5 \\ o_6 \\ o_7 \\ o_8 \\ o_9 \\ o_0 \end{array} \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\ \begin{pmatrix} 1.46 & 0.0 & 0.0 & 1.46 & 1.46 & 1.46 & \cdots \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \cdots \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \cdots \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \cdots \\ 0.13 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \cdots \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \cdots \\ 0.28 & 0.0 & 0.0 & 0.0 & 0.81 & 0.28 & \cdots \\ 0.0 & 0.0 & 1.2 & 0.0 & 2.92 & 1.46 & \cdots \\ 2.44 & 0.00 & 0.0 & 1.46. & 1.46 & 1.19 & \cdots \end{pmatrix} \end{array}$$

Cada columna de la matriz representa a un término, mientras que las filas son los contextos, esto es, las oraciones. En la matriz se compara qué tanto aparece un término en una oración y esto se multiplica por el valor de tf-idf, por tal motivo los valores resultan fraccionarios. De tal forma que se busca representar cuántos términos comparten las diferentes oraciones a través de los distintos documentos, dando relevancia a aquellos con mayor peso. Por ejemplo, en un documento de «Atracción sexual...» aparece la siguiente oración:

> La prostitución se vincula, más que con el comercio sexual, con un régimen de signos fundados en una visibilidad exacerbada...

Mientras que en un texto de «Enfermedades de transmisión sexual» se tiene:

> Las enfermedades de transmisión sexual (ETS) representan a un grupo de padecimientos infecciosos que se transmiten predominantemente por contacto sexual.

La palabra «sexual» tuvo en el corpus un tf-idf de 0.14, por lo que se tendría una matriz como la siguiente:

$$\begin{array}{c} \\ o_{1,2} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{prosti-} & \text{comerc-} & \text{sexual} & \text{enferm-} & \text{ets} & \text{contac-} & \end{array} \\ \begin{pmatrix} 0.0 & 0.0 & 0.42 & 0.0 & 0.0 & 0.0 & \cdots \end{pmatrix} \end{array}$$

Es decir, la co-ocurrencia de la palabra «sexual» entre la primera y la segunda oración es de $3 \cdot 0.14 = 0.42$, mientras que los otros términos tendrán valores

de 0. Una matriz con documentos disímiles (por ejemplo una de temática de «Atracción sexual y relaciones entre individuos» contra uno de «comportamiento sexual») está prácticamente llena con ceros, sólo algunas co-ocurrencias tienen valores que no llegan a ser mayores a 1 (por lo que no tendría caso poner aquí un ejemplo). En las Figuras 4.1 y 4.2 se trató de resumir las co-ocurrencias de las matrices.

Aplicación del modelo de similitud textual y agrupamiento

Una vez creadas las matrices con el CSMX se aplicó el modelo de similitud textual propuesto. Algunos de los resultados obtenidos se muestran en la Tabla 4.5, donde se colocan sólo algunos documentos pertenecientes a las diferentes temáticas.

Tabla 4.5: Resultados del modelo de similitud textual para algunos documentos en las diferentes temáticas.

	Atracción sexual...	Comportamiento sexual	Educación..	Enfermedades de...	Fundamentos biológicos...	Sexualidad variante	Respuesta...
Atracción sexual...	0.608	0.0	0.009	0.002	0.010	0.003	0.063
Comportamiento sexual	0.0	91.54	0.0006	0.0	1.709	0.0	0.004
Educación..	0.006	0.018	0.478	0.008	0.0	0.002	0.010
Enfermedades de...	0.003	0.0	0.101	0.421	0.003	0.006	0.004
Fundamentos biológicos...	0.0	0.0	0.22	0.007	119.77	0.0	0.0
Sexualidad variante	0.141	0.0	0.771	0.39	0.0	0.0	0.581
Respuesta...	0.007	0.0	0.069	0.009	0.0	0.066	2.014

En la tabla se muestra la comparación de un documento de una temática con un documento de cada una de las diferentes temáticas. El valor númerico es la similitud obtenida entre el par de documentos a partir del modelo de similitud textual expuesto. Como se observa, en general, las métricas más altas son para aquellos documentos que comparten temática. Para hacer este hecho más claro en la Tabla 4.6 se muestra la suma de la similitud textual de cada una de las comparaciones de documentos, dando una idea de qué tanta similitud existe entre las diferentes temáticas.

En general, se puede observar en la Tabla 4.6, que aquellas temáticas que son similares tienden a tener valores más altos de similitud. Por otro lado, las temáticas que no son iguales tienen valores más bajos. Sin embargo, se presentan algunas excepciones; tal es el caso de la temática de «Atracción sexual

Tabla 4.6: Suma de resultados del modelo de similitud textual en cada una de las diferentes combinaciones temáticas.

	Atracción sexual...	Comportamiento sexual	Educación..	Enfermedades de...	Fundamentos biológicos...	Sexualidad variante	Respuesta...
Atracción sexual...	3.462	0.589	5.434	2.208	0.683	2.617	3.376
Comportamiento sexual	0.589	1948.76	0.543	0.406	12.157	0.498	3.442
Educación..	5.434	0.543	4.259	3.111	1.061	2.469	2.873
Enfermedades de...	2.208	0.406	3.111	10.084	0.314	1.970	1.604
Fundamentos biológicos...	0.683	12.157	1.061	0.314	556.84	0.224	0.281
Sexualidad variante	2.617	0.498	2.469	1.970	0.224	3.219	0.019
Respuesta...	3.376	3.442	2.873	1.604	0.281	0.019	11.735

y relaciones entre individuos» que tiene valores más altos cuando se compara con la temática de «Comportamiento sexual». Debe tomarse en cuenta que lo que se muestra en la Tabla 4.6 es la suma de todas las métricas de pares de documentos, en donde uno pertenece a una temática y otro a la segunda temática; por tanto, pueden existir documentos (como los que se comparan en la Tabla 4.5) que tengan similitud muy baja; sin embargo, si algunos documentos, aunque sean pocos, tienen alta similitud en temáticas disímiles, se presenta ruido en el sistema y puede ocasionar que algunos documentos se agrupen mal.

Una vez aplicado el algoritmo de similitud textual, entonces se puede aplicar el algoritmo de agrupamiento que, como se ha dicho, toma la métrica propuesta como factor de cercanía entre los elementos de los grupos. En la Tabla 4.7 se condensan los grupos obtenidos por medio del algoritmo de agrupamiento.

Tabla 4.7: Suma de los grupos obtenidos con el algoritmo de agrupamiento y el modelo de similitud textual.

Temática	Grupo1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
Atracción sexual...	20	0	0	0	0	0
Comportamiento sexual	3	14	0	2	0	2
Educación sexual...	13	0	7	0	0	0
Enfermedades de...	2	0	0	18	0	0
Fundamentos biológicos...	4	8	7	0	0	0
Sexualidad variante	10	0	1	1	6	1
Respuesta y expresión...	9	0	0	0	0	11

Aquí se suman los diferentes agrupamientos realizados por cada nivel del corpus. En general, se podría hablar de 6 grupos en los que algoritmo distribuyó las

diferentes temáticas. Se observa que el primer grupo es el que mayor ruido tiene, pues, si bien agrupó a todos los documentos de «Atracción sexual...», el algoritmo solía colocar en este mismo grupo a otros documentos. De forma similar, se agruparon documentos de la temática «Eduación sexual» con «Fundamentos biológicos...» con bastante frecuencia (en la tabla se puede ver en el grupo 3). En los otros grupos, el algoritmo llegó a agrupar textos mayoritariamente de una sola temática. Los resultados obtenidos se muestran con más detalle en el Apéndice B. Para determinar la validez de los grupos, se realizó una evaluación que se presentará a continuación.

4.3. Métodos de evaluación

Para evaluar el agrupamiento realizado por el sistema se utilizó una medida de pureza y el índice de Rand (RI). Tanto la pureza y el RI son métodos comúnmente utilizados en la evaluación de agrupamiento (Manning et al., 2008). Estas medidas se exponen a continuación.

4.3.1. Pureza

La pureza evalúa qué tan puro es un grupo; es decir, toma en cuenta qué tantos elementos similares están dentro de los grupos propuestos por el algoritmo de agrupamiento (Manning et al., 2008). Formalmente se define como:

$$purity(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} max_j |\omega_k \cap c_j|$$

Donde N es el total de elementos, en este caso documentos, $\Omega = \{\omega_1, \omega_2, ..., \omega_k\}$ es el conjunto de grupos propuesto por el algoritmo de agrupamiento y $\mathbb{C} = \{c_1, c_2, ..., c_k\}$ es el conjunto de clases, que en este caso serían las temáticas de los documentos. Por ejemplo, si el algoritmo de agrupamiento arrojó 3 grupos como los de la Figura 4.3, donde se tienen 3 temáticas representadas por «x», «○» y «◇», se tiene que $N = 17$, pues es el número total de elementos agrupados en los 3 grupos. Para el grupo 1, $max_j|\omega_k \cap c_j| = 5$ ya que la clase mayoritaria es «x» y en este grupo se presentan 5 elementos «x». Lo mismo se aplicaría para el grupo 2, donde la clase mayoritaria es «○» con un total de 4 y para el grupo 3, $max_j|\omega_k \cap c_j| = 3$. Así en este caso, $purity(\Omega, \mathbb{C}) = \frac{1}{17} \cdot (5 + 4 + 3) = 0.705$.

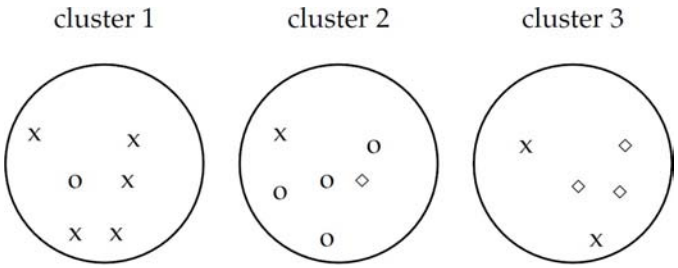

Figura 4.3: Ejemplo de grupos de documentos para su evaluación

4.3.2. Índice de Rand

Por otra parte, se tomó como medida de evaluación el índice de Rand o RI (Manning et al., 2008), que se define formalmente como[4]:

$$RI = \frac{VP + VN}{VP + FP + FN + VN}$$

En donde VP refiere a los verdaderos positivos, VN los verdaderos negativos, FP falsos positivos y FN falsos negativos. Por ejemplo, en la Tabla 4.7, se agrupan juntos documentos de «atracción sexual...» y de «educación sexual»; cada uno de los pares de documentos de estas dos temáticas es un FP. VP son los pares de documentos de «atracción sexual...» que se agruparon juntos. Por su parte, un ejemplo de VN es cada par de documentos de «educación sexual» en el grupo 1 y en el grupo 3 (aquellos documentos que deberían pertenecer al mismo grupo, pero que el sistema separó). Por último, un VN puede verse en los pares de documentos de «atracción sexual» en el grupo 1 y de «comportamiento sexual» en el grupo 2; es decir, los documentos que no pertenecían al mismo grupo y que el sistema reconoció correctamente.

Este algoritmo toma los documentos por pares y determina cuántos pares similares están agrupados juntos («verdaderos positivos») y cuántos pares disímiles se agruparon separados («verdaderos negativos»); además toma en cuenta los errores cometidos por el sistema («falsos negativos» y «falsos positivos»). Si tomamos de nuevo el ejemplo de la Figura 4.3, tenemos que se tienen 20 verdaderos positivos o VP, pues agrupa correctamente 20 pares posibles; los verdaderos negativos son 72, pues éste es el número de pares que se agrupan

[4]Se puede ver que el índice de Rand está relacionado con la medida de *accuracy* que se define como $accuracy = (VP+VN)/(VP+FP+FN+VN)$. La única diferencia entre ambas medidas es el campo de aplicación; mientras el Índice de Rand se aplica para aprendizaje no supervisado, la medida de *accuracy* se utiliza en aprendizaje supérvisado.

en grupos diferentes. Por último, los falsos negativos son 24 pares y los falsos positivos son 20, es decir que estos pares fueron mal agrupados. Así, se obtiene que $RI = \frac{92}{136} \approx 0.68$.

4.4. Evaluación de resultados

En la realización de la evaluación se utilizaron los diferentes métodos expuestos en el Capítulo 2, para compararlos con la propuesta presentada en esta tesis; se exceptuaron los modelos basados en taxonomías, pues estos representaban un enfoque diferente al que se plantea aquí, además requieren de la elaboración de una taxonomía, lo cual es una tarea de realización compleja.

A continuación se muestra la evaluación de modelo de similitud textual aquí propuesto con los que se han presentado en el Capítulo 2. En todos los casos se utilizó la misma metodología ya expuesta, con la excepción del modelo de similitud textual. Se tomaron diferentes valores de ϕ en el algoritmo de MajorClust, donde cada uno de estos valores correspondió a un modelo de similitud textual: la distancia de coseno básica expuesta en Senellat and Blondel (2003) y que es frecuentemente usada en algoritmos de agrupamiento por presentar una complejidad relativamente baja; el algoritmo propuesto por Lin (1998); el espacio semántico con contexto a nivel documento y la implementación con descomposición en valores singulares; finalmente, se utilizó el modelo de similitud textual basado en oraciones.

Cabe aclarar que, dado que el corpus era demasiado grande, la memoria del sistema era insuficiente en algunos casos, por lo que se requirió que el mismo sistema separara los documentos por nivel de especialidad, para después reunirlos una vez procesados. Es decir, el sistema de agrupamiento trabajaba los diferentes niveles de especialidad separadamente, pero al final mostraba los resultados de estos diferentes niveles en una misma salida. Para la evaluación, se aplicó este mismo proceso a todos los modelos aquí presentados.

En la Tabla 4.8 se muestran los promedios de las evaluaciones en el corpus CSMX a través de los cinco niveles de especialización. Las agrupaciones que se obtuvieron como salida del sistema se adjuntan en el Apéndice B. La evaluación se realizó en comparación con el etiquetado humano con el que ya contaba el CSMX (véase Sierra et al. (2009)). Es decir, se tomó en cuenta qué tanto coincidía la clasificación humana por la hecha con el sistema automático; de esta forma, se esperaba que el sistema automática pudiera realizar un agrupamiento

como el mostrado en la Tabla 3.1 de la Sección 4.1.

Tabla 4.8: Promedios de evaluación en diferentes funciones de similitud a partir del CSMX.

ϕ	**Pureza**	**RI**
Distancia de coseno	0.43	0.75
Lin	0.26	0.42
ES (SVD)	0.45	0.71
ES (oraciones)	**0.54**	**0.79**
ES (documentos)	0.33	0.58

Se puede observar que, con lo que respecta a la pureza, los algoritmos muestran diferentes niveles de efectividad al agrupar correctamente los documentos en grupos. Cuando se toma ϕ como distancia de coseno o semántica latente, la evaluación, se puede decir, es media, y los tiempos de ejecución varían muy poco, siendo más rápida la distancia de coseno. La función de similitud propuesta por Lin (1998) es la más baja de todas en cuanto en ambos corpus, seguida por el modelo de espacio semántico a nivel documento; además que sus tiempos de ejecución son altos comparados con la distancia de coseno o con aplicación de SVD. El modelo de espacio semántico con contexto a nivel de oración aquí propuesto presentó la evaluación más alta tanto en RI como en pureza, lo que refleja que en efecto está capturando un poco mejor las relaciones de los términos entre documentos (y más precisamente entre oraciones inter-documento).

En específico, se pueden observar las evaluaciones de cada método dentro de cada nivel de especialidad. En la Figura 4.4 se muestra el RI (barra azul) y la pureza (barra anaranjada) de cada método por nivel. Siendo el nivel 1 el que cuenta con mayor especialidad y el 5 el de menor especialidad (estos niveles van desde artículos científicos hasta foros de discusión, pasando por textos de difusión, textos creados por asociaciones dedicadas a la sexualidad y páginas web, para mayor detalle, revísese banda); en el eje y se plasma el valor de evaluación. Se nota que los diferentes métodos tienen efectividad diferente según la especificidad de los textos. En el caso de la distancia de coseno, parece capturar poco las relaciones en un nivel alto de especialidad, pero esto aumenta en los niveles más bajos, donde este método tiene un desempeño más alto que los otros. En el caso del método de Lin, los resultados son muy bajos; con SVD los resultados son buenos, resaltando en el nivel 4. En general el modelo basado en oraciones tiene un desempeño continuo, superando a los otros métodos mencionados, excepto para el nivel 5, donde su desempeño se reduce considerablemente. En el

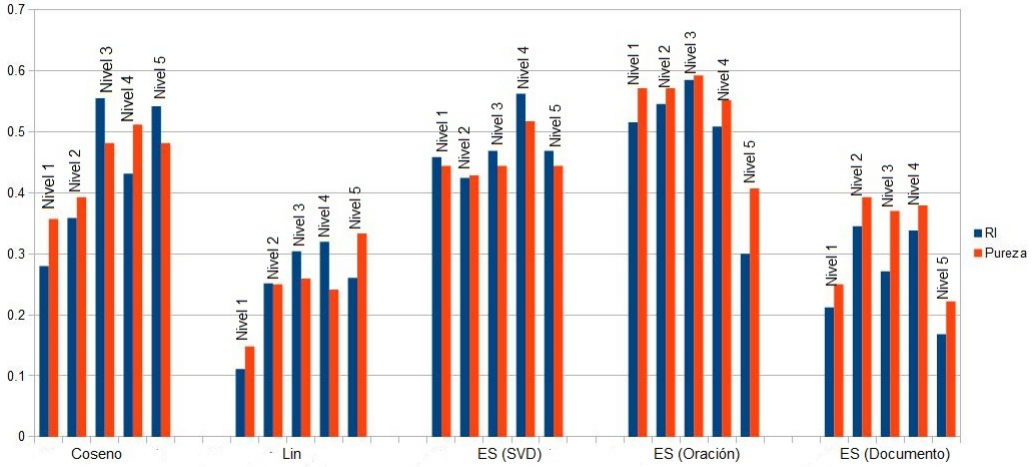

Figura 4.4: Desempeño de las propuestas de similitud textual en cuanto los niveles de especialidad en el CSMX. El nivel 1 es el más especializado (académico) y el nivel 5 el menos especializado (páginas web).

caso del uso de los documentos como contextos, el desempeño es bajo (aunque ligeramente más alto que el algoritmo de Lin), subiendo en los niveles 2 a 4.

En general, se puede notar que los modelos propuestos con espacios semánticos son una buena medida de similitud entre documentos; por tanto, son una buena herramienta para la función de similitud en agrupamiento de documentos. Estas medidas llegan a superar la distancia de coseno (excepto por el método más ingenuo de espacio semántico con contexto a nivel documento). Se puede decir que el método de mayor efectividad es el modelo basado en oraciones [5]. Esto deja ver que en un modelo de espacio semántico, las oraciones son una buena elección de contexto, pues así se puede capturar bastante bien las relaciones y características temáticas de los documentos. Es decir, en un contexto de oración se puede determinar con cierta seguridad la similitud textual entre dos documentos; sin embargo es un proceso más lento pues, a diferencia de la selección por documentos, las matrices deben hacerse entre pares de documentos. El desempeño de acuerdo con los tiempos de ejecución se puede ver en la Figura 4.5. En el eje x se muestra el tamaño del corpus en documentos y en el eje y el tiempo de ejecución del algoritmo en segundos. El modelo aquí propuesto

[5] Aunque esto en gran medida depende del objetivos que se busquen. Además, debe tomarse la cuestión de tiempos de ejecución, con el cual se preferiría el método que utiliza SVD.

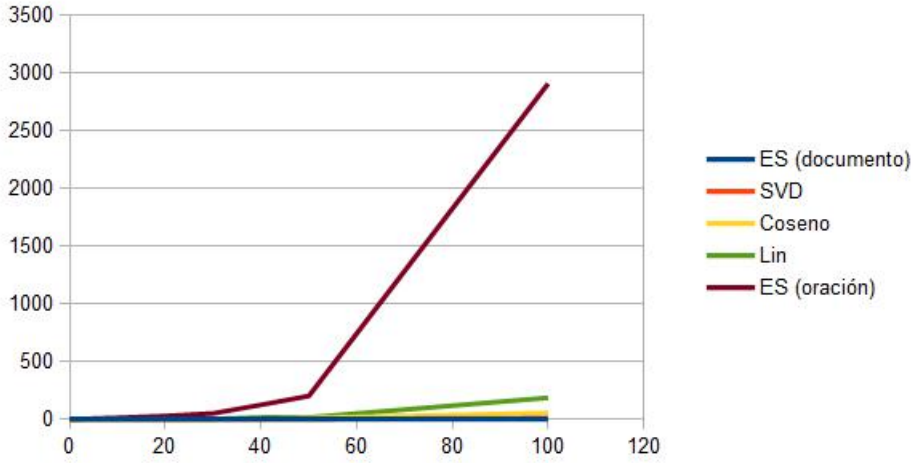

Figura 4.5: Desempeño de las propuestas de similitud textual en cuanto al tiempo de similitud

sobresale de los demás modelos que están debajo de los 500 segundos.

En cuestión de tiempo, el espacio semántico basado en oraciones está muy por arriba de las otras aproximaciones. Esto se debe a que el análisis que hace esta propuesta es mucho más minucioso; la creación de una matriz de matrices implica que la computadora tomará más tiempo en acceder a los datos; además de que la cantidad de información con la que trabaja es más grande que con los otros modelos, pues no sólo compara documento contra documento, sino que también compara cada oración de un documento contra todas las oraciones de los documentos que componen el corpus.

De esta forma, se puede concluir que el método que aquí se plantea para el agrupamiento temático de documentos creó mejores grupos que los otros algoritmos; sin embargo, el costo computacional límita el uso del algoritmo a corpus que no sean demasiado grandes, además de que requiere de mayor tiempo de procesamiento que otros modelos de agrupamiento. Por tanto, deben tomarse en cuenta los objetivos del agrupamiento para determinar si es conveniento utilizar un método como el que aquí se propone.

5

Resumen y conclusiones

5.1. Resumen de la tesis

En esta tesis se ha presentado un panorama general sobre la problemática que implica el agrupamiento temático de documentos dentro de la lingüística computacional. Se mostró que es un problema que involucra a la similitud textual, puesto que el agrupamiento requiere reconocer los textos que sean similares y separarlos de aquellos cuya similitud es mucho menor. Ante esto, se planteó la idea de que los términos podrían tener cierta carga temática que pudiera ayudar a detectar similitud entre documentos, ya que la temática parece determinar en gran medida el uso terminológico que se hace dentro de un texto.

De esta forma, se propuso cuantificar la similitud entre documentos con ayuda de los elementos terminológicos. Primero se introdujo una idea sobre lo que es el concepto de *agrupamiento*. Se trató de insertarlo dentro del área del aprendizaje de máquina que a su vez es parte de la inteligencia artificial. Dentro del aprendizaje de máquina, se vio que el agrupamiento es parte del aprendizaje no supervisado, ya que no requiere de un corpus de entrenamiento, sino que se basa en características dadas del lenguaje. Asimismo, se mostraron diversos algoritmos de agrupamiento, resaltando el método de MajorClust, que se consideró como el más adecuado para los propósitos específicos del trabajo.

Posteriormente, se abordó el tema de la similitud textual; se introdujeron conceptos básicos sobre la similitud, desde las perspectivas de autores como Tversky (1977), Goldstone (1994), Gentner and Ratterman (1991) y otros; se abordaron también las nociones formales que se han hecho sobre la similitud y

que después se retoman dentro de la similitud textual. También se mostraron varias *métricas* computacionales que han tratado de abarcar la similitud. Ya en el ámbito de la lingüística computacional, se mostraron varias perspectivas que se han abordado para tratar este problema. Se comenzó hablando de la similitud basada en modelos taxonómicos del lenguaje, en donde se representa el léxico dentro de una taxonomía y a partir de ésta se determina la cercanía de un ítem léxico con otro dentro de la red. Se mostraron varias formas de aproximar este valor de similitud y se concluyó que se trataba de un método complicado y que no garantizaba buenos resultados. Se expuso el algoritmo de Lesk (1986) y su adaptación para funcionar con un modelo taxonómico. Se pasó a hablar sobre el concepto de la información y cómo esta idea se ha aplicado a la detección de similitud textual; se expuso el modelo propuesto por Lin (1998). Se hizo notar que este algoritmo deja un panorama abierto para la inclusión de los elementos de similitud, posteriormente se trató de solventar esta ambigüedad. Posteriormente, se expuso la hipótesis distribucional expuesta en Harris (1954) y se dio un panorama de cómo ésta se ha aplicado a modelos de similitud textual. Así, se describió el concepto de *espacios semánticos* y su uso dentro de la lingüística computacional. También se describieron brevemente algunos de los modelos y se explicó cómo estos modelos adoptan la hipótesis distribucional y cómo de esta forma pueden ser aplicados para la detección de similitud textual.

En los últimos capítulos, se expuso la metodología planteada. Se describieron los corpus utilizado, tanto el de prueba, como el de evaluación (este último, el Corpus de las Sexualidades en México, CSMX). En la metodología se explicó el algoritmo, el cual constaba de los siguientes elementos: preprocesamiento (eliminación de formas funcionales, tf-idf y *ultra-stemming*), el algoritmo de agrupamiento (MajorClust) y un modelo de similitud textual, que consistía de un espacio semántico basado en oraciones.

Para la evaluación, fue necesario introducir modelos de comparación. Se expuso la aplicación de la distancia de coseno para detectar similitud y se utilizó como línea base. Aquí también se describió la adaptación que se hizo del algoritmo de Lin (1998). Se expuso las ideas de espacios semánticos que se aplicaron: con la utilización del documento como contexto y con la aplicación de descomposición en valores singulares (SVD). Finalmente, se compararon los desempeños de estos cinco métodos utilizados a partir de su evaluación por medio de *pureza* y *RI* (Indexación de Rand). Se comparó asimismo este desempeño dentro de los diferentes modelos de similitud textual planteados, y se discutieron los resultados, notando que los modelos de espacios semánticos superaban

al método de Lin (1998) y a la distancia de coseno; en este caso se observó que los mejores resultados eran dados por la delimitación de contexto a nivel de oración; es decir, por el modelo propuesto en la tesis. Se comparó, de igual forma, el tiempo de proceso por cada uno de los métodos utilizados, siendo el más costoso el basado en el modelo de espacios semánticos y con delimitación de contexto a nivel de oración.

5.2. Comentarios finales

Ya se han discutido los principales problemas del agrupamiento temático de documentos y se ha visto la necesidad que se tiene de la implementación de la similitud textual. Asimismo, se ha expuesto la metodología y los resultados obtenidos. Conforme al objetivo planteado se considera que se ha cumplido, puesto que a partir de diferentes métodos de similitud se ha realizado un agrupamiento temático de un corpus; conforme a esto, se han superado los resultados, incluso, de uno de los métodos tradicionales en la literatura sobre agrupamiento documental, es decir, la *distancia de coseno*. Por tanto, se cree que se ha logrado capturar, por lo menos en parte, las características temáticas que subyacen en los textos analizados.

Con respecto a los objetivos particulares, se cree haber cumplido con ellos, igualmente. En primer lugar, se ha estudiado la literatura y, a partir de esto, se han analizado los elementos lingüísticos que mejor permiten la realización de una clasificación temática, a decir, los términos. De esto se ha hecho un análisis cuantitativo y se ha adoptado la propuesta de *tf-idf* pues se ha visto que ésta captura bien las relaciones dentro y fuera del texto sobre el comportamiento de los términos. De igual forma, se han abarcado diferentes perspectivas dentro de la lingüística computacional sobre la similitud textual; de esta forma, se han adoptado diferentes métodos propuestos aplicándolos tal como los describe la literatura (distancia de coseno, SVD y espacio semántico con contexto a nivel de documento), adaptándolos (algoritmo de Lin) o bien proponiendo nuevos métodos (espacio semántico a nivel de oración). Esto se ha hecho analizando principalmente la distribución de los elementos temáticos.

A partir de esto, se ha tratado de describir la similitud tanto en términos cualitativos pero principalmente cuantitativos, de tal forma que se ha creado un modelo formal que pueda manejar los datos arrojados por los términos y determinar, así, un argumento de similitud. Finalmente, se ha cumplido con el

objetivo de llevar a cabo el agrupamiento temático (a nivel técnico) de textos en diferentes niveles: general y específico.

Al cumplir estos objetivos se trató también de dar crédito a los tres puntos planteados en la Introducción, a saber:

1. Se puede cuantificar la información temática de un texto,
2. existen elementos con mayor información temática (términos) y
3. la distribución de estos elementos dentro de diferentes textos está determinada, en gran medida, por la temática de cada uno de los textos.

Sobre el primer punto, se ha visto que es completamente factible y así, se ha propuesto no sólo uno sino varios métodos de cuantificar la información temática de un texto, enfocándose principalmente a la distancia de coseno, la información y su distribución en los llamados espacios semánticos. Sobre el segundo punto, se ha tratado de mostrar que los términos pueden contener información sumamente útil sobre la temática de los textos; así, se ha visto, que los términos más relevantes (en términos cuantitativos) son los que mejor representan la temática del documento en que aparecen. Gracias al método de *tf-idf* se ha podido cuantificar esta información temática y, así, se ha facilitado su procesamiento computacional.

Se mostró que el comportamiento de los términos (y, en general, de las palabras) responde también a una distribución tanto dentro como fuera del documento, que es, precisamente, lo que trata de capturar el método de *tf-idf*. Se ha tratado, también, de probar que la hipótesis distribucional, manifestada como espacios semánticos, es una buena forma de capturar el comportamiento de estos elementos temáticos. La razón principal de esta idea radica precisamente en el punto 3; es decir, la temática de un documento determina la distribución de estos elementos dentro de diferentes textos. De esta forma, como señalaba Harris (1954), las regularidades distribucionales y el significado (o bien temática) no tienden a extenderse a otros discursos a menos que estén relacionados; esto deja ver que la idea de encontrar distribuciones de términos a través de diferentes textos similares se puede capturar con la propuesta de los llamados espacios semánticos. De cierta forma, esto también explicaría por qué el modelo basado en contexto a nivel de oración puede capturar bien grupos grandes de textos (a diferencia del basado en documentos), pues con éste se capturan mejor las características distribucionales (aparición de dos o más palabras en la misma oración) de los elementos temáticos.

Lo que también se ha tratado de mostrar con esta tesis es la pertinencia de un modelo de similitud textual para el agrupamiento de documentos. Si bien los métodos de la tradición computacional pueden ser útiles, parece que dejan de lado, en mayor o menor medida, las características lingüísticas de los textos. Algunos de éstos, como la representación en espacios vectoriales, ya esbozan las ideas que posteriormente se plantearon con los espacios semánticos; sin embargo, estos modelos matemáticos no saben explicar qué está pasando con el lenguaje.

Finalmente, cabe mencionar que se ha cumplido con las ideas y los objetivos planteados; asimismo, se cree que lo expuesto en este trabajo, así como el sistema desarrollado aquí pueden ser de utilidad para su aplicación en diversas áreas. Después de todo, se ha planteado un formalismo independiente del nivel de especificidad (aunque con mejor o peor desempeño en cada uno); además, en principio, el método planteado es independiente del lenguaje y, se considera, puede trabajar por lo menos con lenguajes que tenga un comportamiento similar al que se ha discutido aquí para el español. Con esta tesis, se ha tratado de dar una aportación a una de las áreas de la lingüística computacional que ha aportado dificultades al área, en tanto que abarca cuestiones que parecen estar fuera de los niveles más bajos del lenguaje.

5.2.1. Aportaciones

Con esta tesis, se ha propuesto una aproximación nueva para el agrupamiento temático de documentos. Se propusó tomar la oración como contexto, para lo que fue necesario crear una nueva métrica para calcular las distancias entre estos. Para este objetivo, se utilizó la distancia coseno como punto de partida.

De esta forma, se mostró a través de la tesis que tomar la oración como contexto para la creación de matrices obtenía mejores resultados que si se consideraba el documento completo. Se vio que esta mejora consistía en que en la oración las coocurrencias entre palabras no se veían tan dispersas como a nivel documento.

Otra de las aportaciones de la tesis fue el mostrar el comportamiento de los términos dentro de los documentos con temáticas similares y cuantificar dicho comportamiento. Así, se pudo determinar cuantitativamente la semejanza temática entre los documentos. Esto permite ver una tendencia al uso de términos similares en textos con temáticas similares, que apoya, con argumentos cuantitativos, la propuesta de Harris. Además, en esta tesis se buscó desarrollar un algoritmo de computadora que proporcione apoyo en las tareas que requieran

de agrupamiento temático de documentos.

5.3. Trabajo a futuro

A lo largo de este trabajo también surgieron diferentes planteamientos que no pudieron ser abarcados aquí. De igual forma, se mostraron ciertos puntos que no pudieron ser trabajados los suficientemente a fondo. Por tanto, se considera que existen puntos que pueden elaborarse más a fondo en trabajos futuros.

Uno de los puntos que pueden mejorarse es el sistema mismo que se creó y utilizó aquí. Si bien para los fines específicos de la tesis, este trabajo está de más, se considera que bien puede realizarse una mejora al sistema para que pueda trabajar de manera más veloz y con mayor cantidad de textos.

También, se podrían utilizar más modelos de comparación, e inclusive se podrían utilizar otros algoritmos de agrupamiento para determinar si los resultados pueden mejorar. Sin duda, una de las cuestiones que implican más trabajo es la recolección del corpus, por tanto, de éste se puede generar una muestra más grande (lo que implicaría, también, que el sistema fuera capaz de manejar mayor cantidad de datos), pues de ésta forma se podría observar la aplicación de un sistema basado en las ideas aquí a problemas corrientes en diferentes ámbitos de la realidad: en el manejo de documentación en empresas o instituciones, en la gestión de textos de internet, etcétera.

Finalmente, se puede seguir ahondando en las ideas de la estructura distribucional y explotarlas en otros ámbito de la lingüística computacional. Estas ideas parecen ser una buena forma de manejar el lenguaje a nivel cuantitativo, puesto que permiten la representación de relaciones y distribuciones específicas entre palabras y textos. La idea de manejar el contexto a nivel oración también podría traer aportaciones a esta rama de la lingüística. En general, esta área abre una multitud de opciones para la creación de aplicaciones y conceptos teóricos dentro de la lingüística computacional.

Bibliografía

Ali, A. (2011). *Textual similarity*. Technical University of Denmark.

Banerjee, S. and Pederson, T. (2002). An adapted lesk algorithm for word sense desambiguation using wordnet. *Computational Linguistics and Intelligent Text Processing*, pages 136–145.

Ben-Naim, A. (2011). *La entropía desvelada: el mito de la segunda ley de la termodinámica y el sentido común*. Tusquets Editores.

Bijuraj, L. (2013). Clustering and its applications. *Proceedings of National Conference on New Horizons in IT*, pages 169–172.

Black, P. E. (1998). Ratcliff/obershelp pattern recognition. *Dictionary of Algorithms and Data Structures*.

Bolshakov, I. and Gelbukh, A. (2004). *Computational Linguistics. Models, Resources, Applications*. Fondo de Cultura Económica, IPN, UNAM.

Brychcín, T. and Koponík, M. (2013). Semantic spaces for improving language modeling. *Computer Speech and Language*.

Cambronero, C. and Moreno, I. (2006). Algoritmos de aprendizaje: Knn y kmeans. *Inteligencia en Redes de Comunicación, Universidad Carlos de Madrid*.

Castro, B., Sierra, G., Torres-Moreno, J.-M., and da Cunha, I. (2011). El discurso y la semántica como recursos para la detección de similitud textual. *Anais do III Workshop: A RST e os Estudos do Text*, pages 30–42.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley-Interscience.

Dougherty, R. (1994). *Natural Language Computing: An english generative grammar in Prolog*. Psychology Press.

Fernández, S., SanJuan, E., and Torres-Moreno, J.-M. (2007). Textual energy of associative memories: Performant aplications of enertex algorithm in text summarization and topic segmentation. *MICAI 2007*, pages 861–871.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3:115–130.

Friedhelm, S. and Trentin, E. (2014). Pattern clasification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, 37:4–14.

Gelbukh, A. and Sidorov, G. (2006). *Procesamiento automático del español con enfoque en recursos léxicos grandes*. Instituto Politécnico Nacional.

Gentner, D. and Ratterman, M. (1991). Language and the career of similarity. *Perspectives on language and thought: Interrelations in development*, pages 225–277.

Goldstone, R. (1994). The role of similarity in categorization. *Cognition*, 52:125–157.

Grishman, R. (2010). *Information extraction*, pages 517–530. Wiley Blackwell.

Hare, J., Samangooei, S., Lewis, P., and Nixon, M. (2008). Semantic spaces revisited: Investigating the performance of auto-annotation and semantic retrieval using semantic spaces. *CIVR'08*, 28.

Harispe, S., Ranwez, S., Janaqui, S., and Montmain, J. (2013). Semantic measures for the comparison of units of language, concepts or instances from text and knowledge representation analysis. *arXiv preprint arXiv*, 1310.

Harris, Z. (1954). Distributional structure. *Papers on Syntax*, pages 3–22.

Howland, P. and Park, H. (2003). *Clustering and Classification*, pages 2–24. Springer.

Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the sixth new zealand computer science research student conference*, pages 49–56.

Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference Research on Computational Linguistics (ROCLING).*

Jones, K. (1972). A statistical representation of term specificity and its application in retrival. *Journal of documentation*, 1:11–21.

Kaufman, L. and Rousseeuw, P. (2005). *Finding Groups in Data. An Introduction to Cluster Analysis.* Wiley.

Lauander, T., Foltz, P., and Laham, D. (1998). An intorduction to latent semantic analysis. *Bell System Technical Journal*, 27:379–423.

Lee, D., Chuang, H., and Seamons, K. (1997). Document ranking and the vector-space model. *Software, IEEE*, 2:67–75.

Lee, L. (1999). Measures of distributional similarity. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of 5th Annual International Conference on Systems Documentation*, pages 24–26.

Levner, E., Pinto, D., Rosso, A., Alcaide, D., and Sharma, R. (2007). Fuzzing clustering algorithms: The case of study of majorclust. *MNICAI 2007: Advances in Artificial intelligence*, pages 821–830.

Lin, D. (1998). An information-theoric definition of similarity. *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning.*

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-ocurrence. *Behavior Research Methods, Instruments & Computers*, 28:203–208.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1:281–297.

Manning, C. and Schutze, H. (2000). *Foundations of Statistical Natural Language Processing.* The MIT Press.

Manning, C., Schutze, H., and Raghavan, P. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Marmanis, H. and Babenko, D. (2009). *Clustering: Grouping Things Together*, pages 121–163. Manning.

Maynard, D. (2000). *Term recognition using combined knowledge sources*. Manchester Metropolitan University.

Metzler, D., Dumais, S., and Meek, C. (2007). Similarity measures for short segments of text. *Advances in information retrieval*, pages 16–27.

Michalsky, R., Carbonell, J., and Mitchell, T. (1986). *Machine Learning: An Artificial Intelligence Approach*. Morgan Kaufmann Publishers, Inc.

Mitchell, T. (1977). *Machine Learning*. McGraw-Hill.

Padó, S. and Lapata, M. (2003). Constructing semantic space models from parsed corpora. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1:128–135.

Payrató, L. (1998). *De profesión, lingüísta: Panorama de la lingüística aplicada*. Ariel.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity - measuring the relatedness of concepts. *Demonstration Papers at HLT-NAACL 2004*, pages 38–41.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 20:130–137.

Raghavan, V. and Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37:279–287.

Rohde, D. and Plaut, L. G. D. (2004). An improved method for deriving word meaning from lexical co-ocurrence. *Cognitive Psychology*, 7:573–605.

Saceda, M. (2005). *Adquisición prosódica en español peninsular septentrional: la sílaba y la palabra prosódica*. Universidad Pompeu Fabra.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20:35–54.

Sateli, B. and Witte, R. (2012). Supporting wiki users with natural language processing. *Proceedings of the 8th Annual International Symposium on Wikis and Open Collaboration*, pages 379–423.

Segaran, T. (2007). *Discovering Groups*, pages 29–53. O'Reilly.

Senellat, P. and Blondel, V. (2003). *Automatic Discovery of Similar Words*, pages 25–45. Springer.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Sidorov, G., Gelbukh, A., Gomez-Adorno, H., and Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18:491–504.

Sierra, G., Medina, A., and Lázaro, J. (2009). Determinación de la terminología básica en sexualidad a partir de la web como corpus. *Panorama de investigaciones basadas en corpus*, pages 374–385.

Stein, B. and Eissen, S. M. (2004). Document categorization with majorclust. *Proceedings of the 12th Workshop on Information Technology and Systems*.

Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. *TextMining Workshop at KDD2000*.

Torres-Moreno, J.-M. (2012). Beyond stemming and lemmatization: Ultra-stemming to improve automatic text summarization. *CoRR*, abs/1209.3126.

Torres-Moreno, J.-M., Molina, A., and Sierra, G. (2010). La energía textual como medida de distancia en agrupamiento de definiciones. *JADT 2010: 10th International Conference on Statistical Analysis of Textual Data*, pages 25–27.

Tversky, A. (1977). Features of similarity. *DPsychological Review*, 84:327–352.

Tyler, S. (1969). *Cognitive Anthropology*. Holt, Rinehart and Winston, Inc.

Vivaldi, J., Cunha, I., Torres-Moreno, J.-M., and Velázquez, P. (2010). Automatic summarization using terminological and semantic resources. *LREC*.

Wang, M. and Nie, J.-Y. (2003). A latent semantic structure model for text classification. *Matrix*, 2.

Wu, H. C., Luk, R. W. P., Wong, K. F., and Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems*, 3:13–37.

Wu, S. and Palmer, M. (1994). Verb semantics and lexical selection. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pages 133–138.

Apéndice A

Algoritmo de Porter

Definiciones

I Vocales del español: a e i o u á é í ó ú

II R1 es la región después de la primera consonante que sigue a una vocal, o la región nula al final de la palabra si no existe tal vocal.

III R2 es la región después de la primera consonante seguida por una vocal en R1, o es la región nula al final de la palabra si no existe dicha consonante.

IV RV es la región después de la tercera letra; o bien la región después de la primera vocal que no sea principio de palabra o final de palabra. Si la segunda letra es una consonante, RV es la región después de la vocal siguiente. Si las dos primeras letras son vocales, RV es la región después de la siguiente consonante. En cualquier otro caso, RV es la región después de la tercera letra.

Algoritmo

Siempre realizar los Pasos 0 y 1.
Paso 0 - Clítico agregado:

> Se buscan los siguientes sufijos:

- me se sela selo selas selos la le lo las les los nos.

> Se borran si son aparece en RV alguno de los siguiente casos:

- iéndo ándo ár ér ír
- ando iendo ar er ir
- yendo (seguido de) u

Paso 1 - Eliminación de sufijos básicos:

> Se buscan los siguientes sufijos y se lleva a cabo la acción indicada:

- anza anzas ico ica icos icas ismo ismos able ables ible ibles ista istas oso osa osos osas amiento amientos imiento imientos → se borran si aparecen en R2.
- adora ador ación adoras adores aciones ante antes ancia ancias → Se boran si aparecen en R2, o precedidos por «ic» en R2.
- logía logías → Se reemplazan por «log» si aparecen en R2.
- ución uciones → Se remplazan por «u» si aparecen en R2.
- encia encias → se reemplzan por «ente» si aparecen en R2.
- mente → Se borran si aparecen en R2.
- idad idades → Se borran si aparecen en R2 (si son precedidas por «abil» «ic» o «iv»).
- iva ivo ivas ivos → Se borran si aparecen en R2 (precedidas por «at»).

Realizar el Paso 2a si no se ha removida ningún final de palabra en el Paso 1.

Paso 2a - Sufijos verbales que comienzan por «y»:

> Buscar los siguientes sufijos en RV y borrar si son precedidos por «u»:

- ya ye yan yen yeron yendo yo yó yas yes yais yamos.

Realizar el Paso 2b si en 2a no ha removido sufijos.

Paso 2b - Otros sufijos verbales:

> Buscar los siguientes sufijos en RV y borrarlos si son precedidos por «u»:

- en es éis emos → (si son precedidos por «gu» se borra también «u»).

- arían arías arán arás aríais aría aréis aríamos aremos ará aré erían erías erán erás eríais ería eréis eríamos eremos erá eré irían irías irán irás iríais iría iréis iríamos iremos irá iré aba ada ida ía ara iera ad ed id ase iese aste iste an aban ían aran ieran asen iesen aron ieron ado ido ando iendo ió ar er ir as abas adas idas ías aras ieras ases ieses ís áis abais íais arais ierais aseis ieseis asteis isteis ados idos amos ábamos íamos imos áramos iéramos iésemos ásemos.

Siempre realizar el Paso3.

Paso 3 - Sufijos residuales:

> Buscar por los siguientes sufijos y realizar la acción indicada:

- os a o á í ó → Borrar si se encuentran en RV
- e é → Borrar si están en RV y son precedidos por «gu»; si la «u» está también en RV, ésta también se borra.

Paso final:

> Eliminar acentos agudos.

Apéndice B

Grupos realizados

Aquí se presentan los grupos realizados por medio del método de similitud textual propuesto (ES oración) aplicado al CSMX. Se muestra, además, cada uno de los grupos obtenidos con los otros modelos de similitud con los que se compara el método, es decir, la distancia de coseno, el algoritmo de Lin, un Espacio Semántico con SVD y un Espacio Semántico a nivel Documentos.

Distancia de coseno

```
---- CLUSTER 5 ----
n_3\comportamiento (1).txt
n_3\comportamiento (2).txt
n_3\comportamiento (3).txt
n_3\fundamentos (2).txt

---- CLUSTER 15 ----
n_3\atraccion (1).txt
n_3\atraccion (2).txt
n_3\atraccion (3).txt
n_3\educacion (1).txt
n_3\educacion (3).txt
n_3\educacion (4).txt
n_3\ets (1).txt
n_3\ets (2).txt
n_3\ets (4).txt
n_3\parafilias (3).txt
n_3\respuesta (4).txt

---- CLUSTER 18 ----
n_3\fundamentos (1).txt
n_3\fundamentos (3).txt

---- CLUSTER 20 ----
n_3\parafilias (1).txt
n_3\parafilias (2).txt
```

```
                                n_5\parafilias (2).txt
---- CLUSTER 23 ----
n_3\comportamiento (4).txt
n_3\educacion (2).txt           ---- CLUSTER 23 ----
n_3\ets (3).txt                 n_5\atraccion (1).txt
n_3\parafilias (4).txt          n_5\atraccion (3).txt
n_3\respuesta (1).txt           n_5\atraccion (4).txt
                                n_5\educacion (1).txt
                                n_5\educacion (2).txt
---- CLUSTER 24 ----            n_5\educacion (3).txt
n_3\atraccion (4).txt           n_5\educacion (4).txt
n_3\respuesta (2).txt           n_5\ets (1).txt
n_3\respuesta (3).txt           n_5\ets (2).txt
                                n_5\ets (4).txt
                                n_5\parafilias (3).txt
---- CLUSTER 8 ----             n_5\parafilias (4).txt
n_5\comportamiento (2).txt      n_5\respuesta (1).txt
n_5\comportamiento (2).txt      n_5\respuesta (2).txt
n_5\comportamiento (3).txt      n_5\respuesta (3).txt
n_5\comportamiento (4).txt      n_5\respuesta (4).txt


---- CLUSTER 18 ----            ---- CLUSTER 5 ----
n_5\comportamiento (1).txt      n_4\comportamiento (1).txt
n_5\fundamentos (1).txt         n_4\comportamiento (2).txt
n_5\fundamentos (2).txt         n_4\comportamiento (3).txt
n_5\fundamentos (3).txt
n_5\fundamentos (4).txt
                                ---- CLUSTER 12 ----
                                n_4\educacion (4).txt
---- CLUSTER 21 ----            n_4\ets (1).txt
n_5\ets (3).txt
n_5\parafilias (1).txt
                                ---- CLUSTER 15 ----
                                n_4\atraccion (1).txt
---- CLUSTER 22 ----            n_4\educacion (2).txt
n_5\atraccion (2).txt           n_4\educacion (3).txt
```

```
n_4\ets (2).txt
n_4\ets (3).txt
n_4\ets (4).txt
n_4\fundamentos (2).txt
n_4\parafilias (1).txt
n_4\parafilias (2).txt
n_4\parafilias (3).txt
n_4\respuesta (1).txt
n_4\respuesta (3).txt

---- CLUSTER 16 ----
n_4\comportamiento (4).txt
n_4\fundamentos (1).txt


---- CLUSTER 19 ----
n_4\atraccion (2).txt
n_4\atraccion (3).txt
n_4\atraccion (4).txt
n_4\fundamentos (3).txt
n_4\fundamentos (4).txt


---- CLUSTER 24 ----
n_4\educacion (1).txt
n_4\respuesta (2).txt
n_4\respuesta (4).txt


---- CLUSTER 17 ----
n_1\ets (1).txt
n_1\fundamentos (2).txt


---- CLUSTER 10 ----
n_1\educacion (1).txt
n_1\educacion (3).txt
n_1\fundamentos (4).txt


---- CLUSTER 27 ----
n_1\comportamiento (4).txt
n_1\respuesta (2).txt
n_1\respuesta (4).txt


---- CLUSTER 20 ----
n_1\educacion (4).txt
n_1\fundamentos (1).txt
n_1\parafilias (1).txt


---- CLUSTER 21 ----
n_1\atraccion (1).txt
n_1\atraccion (2).txt
n_1\atraccion (3).txt
n_1\atraccion (4).txt
n_1\comportamiento (1).txt
n_1\comportamiento (2).txt
n_1\comportamiento (3).txt
n_1\educacion (2).txt
n_1\ets (2).txt
n_1\ets (3).txt
n_1\ets (4).txt
n_1\fundamentos (3).txt
n_1\parafilias (2).txt
n_1\parafilias (3).txt
n_1\parafilias (4).txt
n_1\respuesta (1).txt
n_1\respuesta (3).txt


---- CLUSTER 25 ----
```

```
n_2\ets (2).txt
n_2\fundamentos (1).txt
n_2\respuesta (1).txt
n_2\respuesta (2).txt


---- CLUSTER 26 ----
n_2\atraccion (1).txt
n_2\atraccion (3).txt
n_2\educacion (1).txt
n_2\educacion (2).txt
n_2\fundamentos (3).txt
n_2\respuesta (3).txt


---- CLUSTER 21 ----
n_2\atraccion (2).txt
n_2\atraccion (4).txt
n_2\comportamiento (3).txt
n_2\educacion (3).txt
n_2\educacion (4).txt
n_2\ets (1).txt
n_2\ets (3).txt
n_2\ets (4).txt
n_2\fundamentos (2).txt
n_2\parafilias (1).txt
n_2\parafilias (2).txt
n_2\parafilias (3).txt
n_2\parafilias (4).txt
n_2\respuesta (4).txt


---- CLUSTER 7 ----
n_2\comportamiento (1).txt
n_2\comportamiento (2).txt
n_2\comportamiento (4).txt
n_2\fundamentos (4).txt
```

Algoritmo de Lin

```
---- CLUSTER 3 ----
n_1\atraccion (1).txt
n_1\atraccion (2).txt
n_1\atraccion (3).txt
n_1\atraccion (4).txt
n_1\comportamiento (1).txt
n_1\comportamiento (2).txt
n_1\comportamiento (3).txt
n_1\comportamiento (4).txt
n_1\educacion (1).txt
n_1\educacion (2).txt
n_1\educacion (3).txt
n_1\educacion (4).txt
n_1\ets (1).txt
n_1\ets (2).txt
n_1\ets (3).txt
n_1\ets (4).txt
n_1\fundamentos (1).txt
n_1\fundamentos (2).txt
n_1\fundamentos (3).txt
n_1\fundamentos (4).txt
n_1\parafilias (1).txt
n_1\parafilias (2).txt
n_1\parafilias (3).txt
n_1\parafilias (4).txt
n_1\respuesta (1).txt
n_1\respuesta (2).txt
n_1\respuesta (3).txt
```

```
n_1\respuesta (4).txt
```

```
---- CLUSTER 2 ----
n_2\atraccion (1).txt
n_2\atraccion (2).txt
n_2\atraccion (3).txt
n_2\atraccion (4).txt
n_2\comportamiento (3).txt
n_2\educacion (1).txt
n_2\educacion (2).txt
n_2\educacion (3).txt
n_2\educacion (4).txt
n_2\ets (1).txt
n_2\ets (2).txt
n_2\ets (3).txt
n_2\ets (4).txt
n_2\fundamentos (1).txt
n_2\fundamentos (2).txt
n_2\fundamentos (3).txt
n_2\parafilias (1).txt
n_2\parafilias (2).txt
n_2\parafilias (3).txt
n_2\parafilias (4).txt
n_2\respuesta (1).txt
n_2\respuesta (2).txt
n_2\respuesta (3).txt
n_2\respuesta (4).txt
```

```
---- CLUSTER 19 ----
n_2\comportamiento (1).txt
n_2\comportamiento (2).txt
n_2\comportamiento (4).txt
n_2\fundamentos (4).txt
```

```
---- CLUSTER 3 ----
n_3\atraccion (1).txt
n_3\atraccion (2).txt
n_3\atraccion (3).txt
n_3\atraccion (4).txt
n_3\comportamiento (4).txt
n_3\educacion (1).txt
n_3\educacion (2).txt
n_3\educacion (3).txt
n_3\educacion (4).txt
n_3\ets (1).txt
n_3\ets (2).txt
n_3\ets (3).txt
n_3\ets (4).txt
n_3\parafilias (1).txt
n_3\parafilias (2).txt
n_3\parafilias (3).txt
n_3\parafilias (4).txt
n_3\respuesta (1).txt
n_3\respuesta (2).txt
n_3\respuesta (3).txt
n_3\respuesta (4).txt
```

```
---- CLUSTER 6 ----
n_3\comportamiento (1).txt
n_3\comportamiento (2).txt
n_3\comportamiento (3).txt
n_3\fundamentos (1).txt
n_3\fundamentos (2).txt
n_3\fundamentos (3).txt
```

```
n_5\educacion (3).txt
n_5\educacion (4).txt
n_5\ets (1).txt
n_5\ets (2).txt
```

```
n_5\ets (3).txt
n_5\ets (4).txt
n_5\parafilias (1).txt
n_5\parafilias (2).txt
n_5\parafilias (3).txt
n_5\parafilias (4).txt
n_5\respuesta (1).txt
n_5\respuesta (2).txt
n_5\respuesta (3).txt
n_5\respuesta (4).txt


---- CLUSTER 7 ----
n_5\comportamiento (1).txt
n_5\comportamiento (2).txt
n_5\comportamiento (2).txt
n_5\comportamiento (3).txt
n_5\comportamiento (4).txt
n_5\fundamentos (1).txt
n_5\fundamentos (2).txt
n_5\fundamentos (3).txt
n_5\fundamentos (4).txt


---- CLUSTER 11 ----
n_4\atraccion (1).txt
n_4\atraccion (2).txt
n_4\atraccion (3).txt
n_4\atraccion (4).txt
n_4\comportamiento (4).txt
n_4\educacion (1).txt
n_4\educacion (2).txt
n_4\educacion (3).txt
n_4\educacion (4).txt
n_4\ets (1).txt
n_4\ets (2).txt
n_4\ets (3).txt
n_4\ets (4).txt
n_4\fundamentos (1).txt
n_4\fundamentos (2).txt
n_4\fundamentos (3).txt
n_4\fundamentos (4).txt
n_4\parafilias (1).txt
n_4\parafilias (2).txt
n_4\parafilias (3).txt
n_4\respuesta (1).txt
n_4\respuesta (2).txt
n_4\respuesta (3).txt
n_4\respuesta (4).txt


---- CLUSTER 6 ----
n_4\comportamiento (1).txt
n_4\comportamiento (2).txt
n_4\comportamiento (3).txt
```

ES (SVD)

```
---- CLUSTER 1 ----
n_1\atraccion (1).txt
n_1\atraccion (2).txt
n_1\educacion (3).txt
n_1\educacion (4).txt
n_1\ets (1).txt
n_1\ets (2).txt
n_1\ets (3).txt
n_1\ets (4).txt
n_1\fundamentos (1).txt
```

```
n_1\fundamentos (2).txt
n_1\fundamentos (3).txt
n_1\fundamentos (4).txt
n_1\parafilias (1).txt
n_1\parafilias (2).txt
n_1\parafilias (3).txt
n_1\parafilias (4).txt
n_1\respuesta (1).txt
n_1\respuesta (2).txt
n_1\respuesta (3).txt
n_1\respuesta (4).txt

---- CLUSTER 3 ----
n_1\atraccion (3).txt
n_1\atraccion (4).txt

---- CLUSTER 9 ----
n_1\educacion (1).txt
n_1\educacion (2).txt

---- CLUSTER 5 ----
n_1\comportamiento (1).txt
n_1\comportamiento (2).txt

---- CLUSTER 7 ----
n_1\comportamiento (3).txt
n_1\comportamiento (4).txt

---- CLUSTER 1 ----
n_3\atraccion (1).txt
n_3\atraccion (2).txt
n_3\educacion (3).txt
n_3\educacion (4).txt
n_3\ets (1).txt
n_3\ets (2).txt
n_3\ets (3).txt
n_3\ets (4).txt
n_3\fundamentos (1).txt
n_3\fundamentos (2).txt
n_3\fundamentos (3).txt
n_3\parafilias (1).txt
n_3\parafilias (2).txt
n_3\parafilias (3).txt
n_3\parafilias (4).txt
n_3\respuesta (1).txt
n_3\respuesta (2).txt
n_3\respuesta (3).txt
n_3\respuesta (4).txt

---- CLUSTER 3 ----
n_3\atraccion (3).txt
n_3\atraccion (4).txt

---- CLUSTER 9 ----
n_3\educacion (1).txt
n_3\educacion (2).txt

---- CLUSTER 5 ----
n_3\comportamiento (1).txt
n_3\comportamiento (2).txt

---- CLUSTER 7 ----
n_3\comportamiento (3).txt
n_3\comportamiento (4).txt
```

```
---- CLUSTER 2 ----
n_5\atraccion (2).txt
n_5\educacion (4).txt
n_5\ets (3).txt
n_5\ets (4).txt
n_5\fundamentos (1).txt
n_5\fundamentos (2).txt
n_5\fundamentos (3).txt
n_5\fundamentos (4).txt
n_5\parafilias (1).txt
n_5\parafilias (2).txt
n_5\parafilias (3).txt
n_5\parafilias (4).txt
n_5\respuesta (1).txt
n_5\respuesta (2).txt
n_5\respuesta (3).txt
n_5\respuesta (4).txt


---- CLUSTER 3 ----
n_5\atraccion (3).txt
n_5\atraccion (4).txt


---- CLUSTER 7 ----
n_5\atraccion (1).txt
n_5\comportamiento (1).txt
n_5\comportamiento (2).txt
n_5\comportamiento (2).txt
n_5\comportamiento (3).txt


---- CLUSTER 9 ----
n_5\comportamiento (4).txt
n_5\educacion (1).txt
```

```
---- CLUSTER 11 ----
n_5\educacion (2).txt
n_5\educacion (3).txt


---- CLUSTER 14 ----
n_5\ets (1).txt
n_5\ets (2).txt


---- CLUSTER 1 ----
n_4\atraccion (1).txt
n_4\atraccion (2).txt
n_4\educacion (3).txt
n_4\educacion (4).txt
n_4\ets (1).txt
n_4\ets (2).txt
n_4\ets (3).txt
n_4\ets (4).txt
n_4\fundamentos (1).txt
n_4\fundamentos (2).txt
n_4\fundamentos (3).txt
n_4\fundamentos (4).txt
n_4\parafilias (1).txt
n_4\parafilias (2).txt
n_4\parafilias (3).txt
n_4\respuesta (1).txt
n_4\respuesta (2).txt
n_4\respuesta (3).txt
n_4\respuesta (4).txt


---- CLUSTER 3 ----
n_4\atraccion (3).txt
n_4\atraccion (4).txt
```

```
---- CLUSTER 9 ----
n_4\educacion (1).txt
n_4\educacion (2).txt

---- CLUSTER 5 ----
n_4\comportamiento (1).txt
n_4\comportamiento (2).txt

---- CLUSTER 7 ----
n_4\comportamiento (3).txt
n_4\comportamiento (4).txt

---- CLUSTER 1 ----
n_2\atraccion (1).txt
n_2\atraccion (2).txt
n_2\educacion (3).txt
n_2\educacion (4).txt
n_2\ets (1).txt
n_2\ets (2).txt
n_2\ets (3).txt
n_2\ets (4).txt
n_2\fundamentos (1).txt
n_2\fundamentos (2).txt
n_2\fundamentos (3).txt
n_2\fundamentos (4).txt
n_2\parafilias (1).txt
n_2\parafilias (2).txt
n_2\parafilias (3).txt
n_2\parafilias (4).txt
n_2\respuesta (1).txt
n_2\respuesta (2).txt
n_2\respuesta (3).txt
n_2\respuesta (4).txt

---- CLUSTER 3 ----
n_2\atraccion (3).txt
n_2\atraccion (4).txt

---- CLUSTER 9 ----
n_2\educacion (1).txt
n_2\educacion (2).txt

---- CLUSTER 7 ----
n_2\comportamiento (1).txt
n_2\comportamiento (2).txt
n_2\comportamiento (3).txt
n_2\comportamiento (4).txt
```

ES (oración)

```
---- CLUSTER 26 ----
n_1\comportamiento (2).txt
n_1\parafilias (3).txt
n_1\respuesta (1).txt
n_1\respuesta (2).txt
n_1\respuesta (3).txt
n_1\respuesta (4).txt

---- CLUSTER 3 ----
n_1\atraccion (1).txt
n_1\atraccion (2).txt
```

```
n_1\atraccion (3).txt
n_1\atraccion (4).txt
n_1\comportamiento (4).txt
n_1\parafilias (1).txt
n_1\parafilias (2).txt


---- CLUSTER 10 ----
n_1\educacion (1).txt
n_1\educacion (2).txt
n_1\educacion (3).txt
n_1\educacion (4).txt
n_1\fundamentos (1).txt
n_1\fundamentos (2).txt
n_1\fundamentos (3).txt
n_1\fundamentos (4).txt
n_1\parafilias (4).txt


---- CLUSTER 15 ----
n_1\comportamiento (1).txt
n_1\comportamiento (3).txt
n_1\ets (1).txt
n_1\ets (2).txt
n_1\ets (3).txt
n_1\ets (4).txt


---- CLUSTER 8 ----
n_5\comportamiento (1).txt
n_5\comportamiento (2).txt
n_5\comportamiento (2).txt
n_5\comportamiento (3).txt
n_5\comportamiento (4).txt
n_5\fundamentos (1).txt
n_5\fundamentos (2).txt
n_5\fundamentos (3).txt
n_5\fundamentos (4).txt


---- CLUSTER 26 ----
n_5\respuesta (1).txt
n_5\respuesta (2).txt
n_5\respuesta (4).txt


---- CLUSTER 3 ----
n_5\atraccion (1).txt
n_5\atraccion (2).txt
n_5\atraccion (3).txt
n_5\atraccion (4).txt
n_5\educacion (1).txt
n_5\educacion (2).txt
n_5\educacion (3).txt
n_5\educacion (4).txt
n_5\parafilias (1).txt
n_5\parafilias (2).txt
n_5\parafilias (3).txt
n_5\parafilias (4).txt
n_5\respuesta (3).txt


---- CLUSTER 15 ----
n_5\ets (1).txt
n_5\ets (2).txt
n_5\ets (3).txt
n_5\ets (4).txt


---- CLUSTER 10 ----
n_4\atraccion (1).txt
n_4\atraccion (2).txt
n_4\atraccion (3).txt
n_4\atraccion (4).txt
```

```
n_4\comportamiento (4).txt
n_4\educacion (1).txt
n_4\educacion (2).txt
n_4\educacion (3).txt
n_4\educacion (4).txt
n_4\fundamentos (1).txt
n_4\fundamentos (2).txt
n_4\fundamentos (3).txt
n_4\fundamentos (4).txt
n_4\parafilias (1).txt
n_4\parafilias (2).txt
n_4\parafilias (3).txt
n_4\respuesta (1).txt
n_4\respuesta (2).txt
n_4\respuesta (3).txt
n_4\respuesta (4).txt

---- CLUSTER 5 ----
n_4\comportamiento (1).txt
n_4\comportamiento (2).txt
n_4\comportamiento (3).txt


---- CLUSTER 15 ----
n_4\ets (1).txt
n_4\ets (2).txt
n_4\ets (3).txt
n_4\ets (4).txt


---- CLUSTER 16 ----
n_2\educacion (1).txt
n_2\educacion (3).txt
n_2\educacion (4).txt
n_2\fundamentos (1).txt
n_2\fundamentos (2).txt
n_2\fundamentos (3).txt


---- CLUSTER 3 ----
n_2\atraccion (1).txt
n_2\atraccion (2).txt
n_2\atraccion (3).txt
n_2\atraccion (4).txt
n_2\comportamiento (3).txt
n_2\educacion (2).txt
n_2\parafilias (4).txt
n_2\respuesta (1).txt
n_2\respuesta (2).txt
n_2\respuesta (3).txt
n_2\respuesta (4).txt


---- CLUSTER 5 ----
n_2\comportamiento (1).txt
n_2\comportamiento (2).txt
n_2\comportamiento (4).txt
n_2\fundamentos (4).txt


---- CLUSTER 22 ----
n_2\parafilias (2).txt
n_2\parafilias (3).txt


---- CLUSTER 13 ----
n_2\ets (1).txt
n_2\ets (2).txt
n_2\ets (3).txt
n_2\ets (4).txt
n_2\parafilias (1).txt
```

```
---- CLUSTER 16 ----
n_3\comportamiento (1).txt
n_3\comportamiento (2).txt
n_3\comportamiento (3).txt
n_3\fundamentos (1).txt
n_3\fundamentos (2).txt
n_3\fundamentos (3).txt


---- CLUSTER 25 ----
n_3\comportamiento (4).txt
n_3\respuesta (1).txt
n_3\respuesta (2).txt
n_3\respuesta (3).txt
n_3\respuesta (4).txt


---- CLUSTER 3 ----
n_3\atraccion (1).txt
n_3\atraccion (2).txt
n_3\atraccion (3).txt
n_3\atraccion (4).txt
n_3\educacion (1).txt
n_3\educacion (2).txt
n_3\educacion (3).txt
n_3\educacion (4).txt
n_3\ets (3).txt
n_3\ets (4).txt
n_3\parafilias (3).txt


---- CLUSTER 13 ----
n_3\ets (1).txt
n_3\ets (2).txt


---- CLUSTER 22 ----
n_3\parafilias (1).txt
n_3\parafilias (2).txt
n_3\parafilias (4).txt
```

ES (documento)

```
---- CLUSTER 24 ----
n_1\atraccion.txt
n_1\atraccion.txt
n_1\atraccion.txt
n_1\atraccion.txt
n_1\comportamiento.txt
n_1\comportamiento.txt
n_1\comportamiento.txt
n_1\comportamiento.txt
n_1\educacion.txt
n_1\educacion.txt
n_1\educacion.txt
n_1\educacion.txt
n_1\ets.txt
n_1\fundamentos.txt
n_1\fundamentos.txt
n_1\fundamentos.txt
n_1\fundamentos.txt
n_1\parafilias.txt
n_1\parafilias.txt


---- CLUSTER 25 ----
n_1\ets.txt
n_1\ets.txt
n_1\ets.txt
```

```
n_1\parafilias.txt
n_1\parafilias.txt
n_1\respuesta.txt
n_1\respuesta.txt
n_1\respuesta.txt
n_1\respuesta.txt


---- CLUSTER 1 ----
n_2\atraccion.txt
n_2\atraccion.txt
n_2\atraccion.txt
n_2\educacion.txt
n_2\educacion.txt
n_2\educacion.txt
n_2\educacion.txt
n_2\ets.txt
n_2\ets.txt
n_2\fundamentos.txt
n_2\fundamentos.txt
n_2\fundamentos.txt
n_2\parafilias.txt
n_2\parafilias.txt
n_2\parafilias.txt


---- CLUSTER 26 ----
n_2\fundamentos.txt
n_2\respuesta.txt


---- CLUSTER 27 ----
n_2\atraccion.txt
n_2\comportamiento.txt
n_2\ets.txt
n_2\ets.txt
n_2\parafilias.txt
n_2\respuesta.txt
n_2\respuesta.txt
n_2\respuesta.txt


---- CLUSTER 7 ----
n_2\comportamiento.txt
n_2\comportamiento.txt
n_2\comportamiento.txt


---- CLUSTER 25 ----
n_3\atraccion.txt
n_3\atraccion.txt
n_3\comportamiento.txt
n_3\educacion.txt
n_3\educacion.txt
n_3\educacion.txt
n_3\educacion.txt
n_3\ets.txt
n_3\fundamentos.txt
n_3\fundamentos.txt
n_3\fundamentos.txt
n_3\parafilias.txt
n_3\parafilias.txt
n_3\parafilias.txt
n_3\respuesta.txt


---- CLUSTER 5 ----
n_3\comportamiento.txt
n_3\comportamiento.txt
n_3\comportamiento.txt


---- CLUSTER 23 ----
n_3\atraccion.txt
```

```
n_3\atraccion.txt
n_3\ets.txt
n_3\ets.txt
n_3\ets.txt
n_3\parafilias.txt
n_3\respuesta.txt
n_3\respuesta.txt
n_3\respuesta.txt


---- CLUSTER 1 ----
n_4\atraccion.txt
n_4\atraccion.txt
n_4\atraccion.txt
n_4\atraccion.txt
n_4\educacion.txt
n_4\educacion.txt
n_4\educacion.txt
n_4\educacion.txt
n_4\ets.txt
n_4\fundamentos.txt
n_4\fundamentos.txt
n_4\parafilias.txt
n_4\parafilias.txt
n_4\parafilias.txt
n_4\respuesta.txt
n_4\respuesta.txt


---- CLUSTER 27 ----
n_4\parafilias.txt
n_4\respuesta.txt


---- CLUSTER 14 ----
n_4\ets.txt
n_4\ets.txt
n_4\ets.txt
n_4\respuesta.txt


---- CLUSTER 7 ----
n_4\comportamiento.txt
n_4\comportamiento.txt
n_4\comportamiento.txt
n_4\comportamiento.txt
n_4\comportamiento.txt
n_4\fundamentos.txt
n_4\fundamentos.txt


---- CLUSTER 24 ----
n_5\atraccion.txt
n_5\atraccion.txt
n_5\atraccion.txt
n_5\atraccion.txt
n_5\comportamiento.txt
n_5\comportamiento.txt
n_5\comportamiento.txt
n_5\comportamiento.txt
n_5\educacion.txt
n_5\educacion.txt
n_5\educacion.txt
n_5\educacion.txt
n_5\ets.txt
n_5\ets.txt
n_5\ets.txt
n_5\fundamentos.txt
n_5\fundamentos.txt
n_5\fundamentos.txt
n_5\fundamentos.txt
n_5\parafilias.txt
n_5\parafilias.txt
n_5\parafilias.txt
```

```
n_5\respuesta.txt
n_5\respuesta.txt
```

```
---- CLUSTER 25 ----
n_5\ets.txt
n_5\respuesta.txt
n_5\respuesta.txt
```