UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
INGENIERA EN COMPUTACIÓN

P R E S E N T A:
MARÍA XIMENA GUTIÉRREZ VASQUES

DIRECTOR DE TESIS:
DR. ALFONSO MEDINA URREA

CIUDAD UNIVERSITARIA 2010

Agradecimientos

A la Universidad Nacional Autónoma de México por ser mi inspiración; a su Ciudad Universitaria por ser mi refugio. Mi paso por la Universidad ha sido una de las mejores experiencias de mi vida.

A la Facultad de Ingeniería, a su invaluable comunidad de estudiantes y profesores. Gracias a ustedes puedo culminar exitosamente esta etapa de mi vida profesional.

A mis hermanos, Emilio y Rodrigo, a mis padres, María y Emilio. Ustedes me han colmado de amor, ánimo y son mi soporte. Hago extensivo el agradecimiento a toda mi familia (incluidos los que ya no están presentes en cuerpo) de la que siempre he recibido amor y apoyo y de la cual me siento muy orgullosa.

Agradezco especialmente a mi director de tesis, Alfonso Medina Urrea, por su asesoría, preocupación y apoyo constante, gracias por introducirme al mundo del resumen automático. De igual manera, agradezco al grupo de Ingeniería Lingüística de la UNAM (GIL) por apoyarme enormemente en mi formación en el área de tecnologías del lenguaje.

A mis amigas que me han acompañado desde hace muchos años, han estado presentes en todo momento. A pesar de los años y las adversidades nos mantenemos unidas.

Quiero agradecer a los amigos que me acompañaron durante la carrera. En mi paso por la facultad encontré a personas excepcionales que me brindaron amistad, apoyo y cariño incondicional. Nunca imaginarán lo enormemente afortunada que me siento por ello.

Una disculpa si mis agradecimientos anteriores suenan un tanto impersonales, elegí no poner nombres porque odiaría no mencionar a alguien.

Agradezco a la DGDC de la UNAM y a la revista *¿Cómo Ves?* por haberme dejado ser parte de su equipo de trabajo. Sobretodo agradezco a Moni, mi tutora, por escucharme y darme la oportunidad de vivir lo que para mí fue un sueño: estar en el mundo del periodismo y la divulgación. Gracias también a mis compañeros becarios, Montse, Leo y Antonio, por llenar de alegría y diversión el trabajo en "el Universum".

Por último, aunque no por eso menos importante, un agradecimiento a todo aquellos voluntarios que dedicaron parte de su valioso tiempo para leer textos y ayudarme así con una parte muy importante de esta tesis.

¡México, Pumas, Universidad!

Índice de contenido

Índice de tablas y figuras

1. INTRODUCCIÓN

En este capítulo se presenta el planteamiento del problema a partir del cual surge este trabajo y los objetivos específicos de esta propuesta. Al final se describe capítulo por capítulo el contenido de esta tesis.

1.1 PLANTEAMIENTO DEL PROBLEMA

Existen diversos factores que han contribuido a la generación de grandes cantidades de información en los últimos años, uno de ellos es el crecimiento de la *World Wide Web* (WWW). Los servicios en línea hacen posible la consulta y disponibilidad de una enorme variedad y cantidad de documentos en formato electrónico.

Esto representa una ventaja, sin embargo, también es el origen de un problema: sobrecarga de información. Resulta difícil para las personas poder leer tanta información y procesarla con el fin de absorber lo que es relevante para ellos. Hoy en día es posible encontrar gran cantidad de documentos relacionados con un tema específico, lo cual dificulta la lectura y procesamiento de todos estos documentos y eventualmente la toma de decisiones.

Diversas áreas de estudio como la recuperación de información, la ingeniería lingüística, la minería de textos, entre otras, se han ocupado de resolver este tipo de problemas diseñando herramientas que sean útiles para la obtención, filtrado, clasificación y extracción de información. Algunos ejemplos de estas aplicaciones son los buscadores de contenido en internet (texto, música, imágenes), los sistemas de clasificación de documentos, filtrado de *spam* en los correos electrónicos, sistemas de búsqueda de respuestas, etc.

Los sistemas de resumen automático encajan dentro de esta familia de herramientas, debido a que su objetivo es producir una versión condensada de un documento que contenga sus partes más relevantes, facilitando de esta manera la obtención de información útil para el usuario.

Los sistemas de resumen automático se han convertido en un área de gran interés tanto en el campo de la investigación como en los sectores comerciales. Existen diferentes aproximaciones o técnicas para la generación de resúmenes así como diferentes tipos de resúmenes. En todo caso, es importante recordar que estos sistemas no pretenden, ni siquiera a largo plazo, emular el comportamiento humano para realizar resúmenes. La intención de estos sistemas no es reemplazar por completo a los agentes humanos sino ser una herramienta capaz de automatizar tareas redundantes y que pueda facilitar la obtención de información útil para el usuario.

El sistema de resumen propuesto en esta tesis tuvo su origen en un proyecto desarrollado en la materia de minería de textos. Este proyecto, llamado "Canario muerto", consistía en la clasificación y agrupamiento automático de correos electrónicos en diferentes lenguas. El resumen automático fue un módulo de este sistema y sirvió para presentar versiones condesadas de los correos electrónicos. También fue útil para que las tareas de clasificación y agrupamiento procesaran los resúmenes generados en lugar de los correos electrónicos enteros.

Otra motivación para el desarrollo del sistema de resumen automático presentado en este trabajo es la necesidad, dentro del marco de la compilación del Corpus Histórico del Español en México (CHEM), de proporcionarles a sus usuarios resúmenes de sus diversos documentos (libros, cartas, actas, etc.). Este corpus, elaborado gracias al apoyo DGAPA PAPIIT 400905 "Constitución del Corpus Histórico del Español en México" (2005-2007)[1], está constituido por documentos de los siglos XVI al XIX que, debido a las importantes variaciones de escritura (grafías y ortografías) a lo largo de esos tiempos, representan un reto interesante en la generación automática de resúmenes. Aparte de esta motivación, el sistema propuesto puede aplicarse a otras lenguas y a otros sistemas de escritura.

[1] De hecho, esta tesis se realizó el marco del proyecto DGAPA PAPIIT 402008 "Glutinometría y variación dialectal" que contempla este tipo de desarrollos en el CHEM.

1.2 Objetivos y estructura de la tesis

El presente trabajo tiene como objetivo proporcionar un panorama general acerca del resumen automático, el funcionamiento de este tipo de sistemas y las alternativas que existen para abordar el problema. Asimismo, la tesis tiene como objetivo proponer un sistema de resumen automático de tipo extractivo, independiente del lenguaje y de la fuente.

La tesis consta de seis capítulos. El segundo capítulo contiene el marco teórico, en donde se abordan los conceptos básicos relacionados con el tema y se establece una arquitectura que generaliza el funcionamiento de los sistemas de resumen automático, de tal manera que se pueda facilitar la comprensión e identificación de las partes esenciales que forman este tipo de sistemas. También se exponen los principios en los que se basan diferentes métodos para la generación de resúmenes automáticos. Finalmente, los últimos subcapítulos abordan conceptos y nociones que serán de utilidad para comprender el sistema de resumen propuesto en el siguiente capítulo.

En el tercer capítulo se presenta la propuesta del sistema de resumen automático. Se profundiza sobre sus características, su funcionamiento y los algoritmos elaborados para su desarrollo.

Las cuestiones técnicas sobre la implementación del sistema son abordadas en el cuarto capítulo; aspectos como el lenguaje de programación elegido, la estructura del programa, entre otros.

En el quinto capítulo se discute la cuestión de la evaluación en los sistemas de resumen automático; además se propone un esquema de evaluación para el sistema desarrollado y posteriormente se analizan los resultados obtenidos.

El último capítulo contiene las conclusiones de la tesis, que incluyen una valoración final del sistema desarrollado, sus alcances, limitaciones, ventajas y áreas para trabajo futuro.

2. MARCO TEÓRICO

En este capítulo se abordan los conceptos básicos relacionados con el tema de resumen automático y se establece una arquitectura que generaliza el funcionamiento de los sistemas de este propósito. También se exponen los principios en los que se basan diferentes métodos para la generación de resúmenes automáticos. Finalmente, en los últimos subcapítulos se abordan conceptos y nociones que serán de utilidad para comprender el sistema de resumen propuesto en el siguiente capítulo.

2.1 RESUMEN AUTOMÁTICO DE TEXTOS

"Un resumen se define como una transformación del texto fuente que se realiza por medio de la reducción del contenido. La reducción se lleva a cabo seleccionando y/o generalizando lo que es importante en la fuente" [1].

Mani y Maybury [2] proponen una arquitectura general de un sistema de resumen automático de textos (figura 1). A partir de este esquema se explican la terminología básica, las características y los procesos que regularmente intervienen en la generación de resúmenes automáticos.

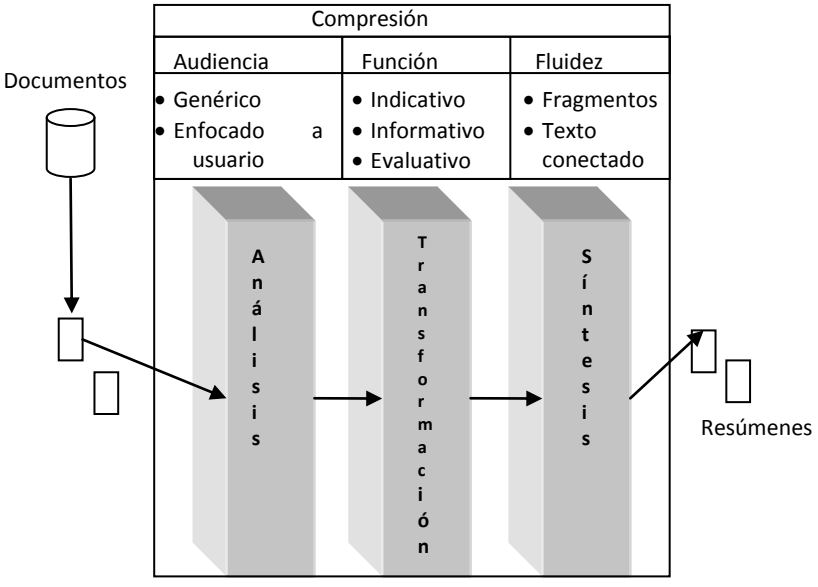


Fig. 1. Arquitectura de un sistema de resumen de textos

El primer aspecto que se debe considerar es la entrada y salida del sistema. Los sistemas de resumen automático reciben como entrada un documento o una colección de documentos. Entenderemos como documento una unidad de texto que está disponible para ser procesada por el sistema; un documento puede referirse a cualquier cosa, por ejemplo, un artículo, una noticia, un correo electrónico, fragmentos extraídos de una página Web, etc.

Los sistemas que reciben como entrada un solo documento son conocidos como sistemas de resumen mono-documental, mientras que los que reciben un conjunto de documentos son llamados sistemas de resumen multi-documental. Más adelante se abordará en qué consiste cada uno.

Por otro lado, la salida del sistema está formada por un documento que contiene una versión condensada del documento de entrada. Este documento resultante es al que llamamos resumen y puede ser un extracto de la fuente o un abstracto (el resumen de tipo extractivo está formado por fragmentos del texto original, mientras que el resumen por abstracción puede contener una nueva redacción expresando el contenido del documento fuente). Los desarrollos en el campo de resumen automático se han concentrado en los extractos más que en los abstractos. Sin embargo, es probable que este enfoque cambie dentro de pocos años conforme se vayan integrando las técnicas de procesamiento de lenguaje natural (PLN) al resumen automático.

Además de la entrada y la salida del sistema, en la figura 1 se pueden observar tres bloques que representan una generalización de las etapas que intervienen en la generación de un resumen: análisis, transformación y síntesis.

La primera etapa tiene que ver con analizar el texto fuente, procesarlo y convertirlo a una representación que sea útil para que en la siguiente etapa se pueda seleccionar la información más relevante. En la segunda etapa, se selecciona (por medio de diversas técnicas) la información considerada relevante o importante que formará al resumen. Finalmente, en la tercera etapa, a partir de la representación de resumen, se construye una forma apropiada para presentar el documento de salida.

Estas etapas describen sólo de manera general el proceso mediante el cual se genera un resumen automático. La implementación de estas etapas puede variar mucho dependiendo de las necesidades y características particulares del tipo de resumen que se planea hacer, así como de las técnicas que se utilizan para seleccionar el contenido relevante.

En la arquitectura mostrada en la figura 1 también aparecen algunas de las características esenciales que intervienen en el diseño de un sistema de resumen automático. En el siguiente apartado se explican más detalladamente estas características de acuerdo con [2] y [9].

CARACTERÍSTICAS PRINCIPALES DE UN RESUMEN

Compresión

Este parámetro se refiere a la longitud del resumen en relación a la longitud del documento fuente. Generalmente son los usuarios quienes eligen el radio de compresión, existen trabajos que sugieren que la longitud óptima de un resumen oscila entre el 20% y el 30% de la longitud del documento original [35].

La longitud del texto puede estar en función del número de palabras, número de caracteres, número de enunciados, etc.

$$radio\ de\ compresión = \frac{|resumen|}{|documento\ fuente\ |}$$

Audiencia (Resúmenes genéricos y resúmenes enfocados a usuarios)

Este aspecto tiene que ver con el alcance del resumen. Los resúmenes genéricos se construyen para una audiencia general, por lo tanto el resumen está formado por la información considerada más relevante y no responde a una necesidad particular de información del usuario. En contraste, los resúmenes enfocados al usuario, también conocidos como resúmenes enfocados a un tema o resúmenes enfocados a una consulta,

responden específicamente a la información que solicita el usuario, de manera que el resumen se construye alrededor del tema de interés para el usuario.

Función

Este aspecto se refiere a la intención o propósito del resumen que puede ser: indicativo, informativo o evaluativo. Los resúmenes indicativos tienen como objetivo mostrar qué temas aborda el texto fuente, es decir, proveen la suficiente información para que el usuario pueda saber cuál es el tema central del documento. Por su parte los resúmenes informativos pretenden cubrir o explicar los conceptos y temas del texto fuente de tal manera que el resumen pueda servir como sustituto del texto original. Finalmente, los resúmenes evaluativos, además de captar la información relevante, pueden incorporar opiniones o juicios sobre el contenido del texto fuente.

Fluidez

Este aspecto está relacionado con el formato final que tendrá el resumen. A la información relevante extraída de un documento se le pueden aplicar diferentes tratamientos con el fin de que la información extraída tenga fluidez en su nuevo contexto. Estos tratamientos pueden incluir el ordenamiento y la limpieza de la información (eliminar redundancias, resolver problemas de coherencia). Por otro lado, el resumen generado puede estar formado, por ejemplo, por un listado de enunciados o fragmentos relevantes, por fragmentos del texto conectados con lenguaje natural para darles continuidad o por un fragmento que contenga una redacción diferente a la del texto fuente, es decir, un abstracto.

Extractos vs abstractos

Esta es una dimensión crucial en la arquitectura de los sistemas de resumen de textos. Los resúmenes formados por extractos, también conocidos como resúmenes de tipo extractivo, están formados por una combinación de enunciados o fragmentos que son extraídos del documento fuente.

En contraste, los resúmenes formados por abstractos o de tipo abstractivo, son aquellos que utilizan diferentes palabras para describir el contenido del documento fuente, es decir, el resumen puede contener una nueva redacción distinta a la del documento original.

La mayoría de los sistemas actuales de resumen automático son extractivos debido a que este es el tipo de resumen más simple: obtener extractos es más fácil que generar un abstracto. La transición a sistemas abstractivos más sofisticados es una meta que persigue la investigación reciente.

Resumen monodocumental y multidocumental

Como ya se ha mencionado, los sistemas de resumen mono-documental generan el resumen de un solo documento; su objetivo es caracterizar la información de un sólo documento, por ejemplo, generar el resumen de una noticia o de un artículo científico.

Por su parte, los sistemas de resumen multi-documental producen un resumen de todo un conjunto de documentos. La meta de este tipo de sistemas es producir una versión condensada del contenido de toda la colección de documentos. Este tipo de sistemas son útiles cuando se desea obtener un resumen de varios documentos relacionados con el mismo tema, por ejemplo, una serie de noticias que abordan el mismo acontecimiento.

2.2 TÉCNICAS DE RESUMEN AUTOMÁTICO

En la elaboración de los resúmenes automáticos existen diferentes aproximaciones para la selección de contenido relevante. Usualmente estas aproximaciones se agrupan de acuerdo al nivel de procesamiento que se le aplica al texto, como se muestra en la siguiente clasificación [2]:

- Técnicas a nivel superficie del texto
- Técnicas a nivel entidades del texto
- Técnicas a nivel discursivo del texto
- Técnicas híbridas

Las técnicas a nivel superficie trabajan con características no profundas del texto, es decir, representan la información en términos de características superficiales sin realizar algún reconocimiento o análisis lingüístico de estas palabras. Por medio de estas características superficiales determinan las partes del texto que son importantes para posteriormente generar el resumen.

Comúnmente se utilizan mediciones estadísticas, por ejemplo, la frecuencia de ocurrencia de las palabras [3] y [4], la posición de las palabas dentro del texto, dentro de un enunciado o de una sección particular del texto [5] y [6]. También se considera la presencia en el texto de palabras relevantes o frases clave [7], por ejemplo, las palabras relevantes pueden ser aquellas que están contenidas en el título y subtítulos del texto, o también frases que se sabe de antemano que indican la existencia de contenido importante, frases como "en conclusión", "en resumen" o "es importante".

Por otra parte, las técnicas a nivel entidades del texto modelan al texto como un conjunto de entes lingüísticos con relaciones entre ellos [2]. Estas técnicas realizan un análisis lingüístico de las palabras, de tal manera que puedan identificar a cada palabra como un ente o unidad lingüística y posteriormente modelar las relaciones entre ellas.

Para poder realizar este reconocimiento lingüístico, es decir, identificar si una palabra es un verbo, nombre propio o pertenece a cualquier otra categoría léxica, generalmente es necesario el uso de analizadores morfológicos, desambiguadores léxicos, lematizadores[2] y bases de conocimiento léxico.

Una vez identificadas las entidades lingüísticas se pueden detectar relaciones entre ellas, por ejemplo, recurrencia de formas o lemas, relaciones semánticas, relaciones temáticas. Son precisamente estos patrones de conectividad en el texto los que permiten determinar qué es importante. Las relaciones entre las entidades se pueden representar con una topología de grafos y pueden incluir: la proximidad entre las unidades lingüísticas del

[2] Lematización es el proceso de reducir una palabra a una forma canónica llamada lema, de tal manera que se puedan agrupar diferentes variantes morfológicas de la palabra. Por ejemplo, las palabras *corrió*, *corriendo*, *corre*, son variaciones de un mismo lema: *correr*

texto, la co-ocurrencia de palabras (palabras relacionadas basadas en su ocurrencia en contextos comunes), co-referencia, relaciones sintácticas, relaciones lógicas, relaciones entre palabras como sinonimia, hiperonimia, etc. Algunos ejemplos de sistemas de resumen automático que utilizan técnicas con este nivel de procesamiento son: [26],[27],[28].

Las técnicas a nivel discursivo trabajan con características profundas del texto. Estas técnicas buscan modelar una estructura global de la argumentación contenida en el texto mediante las relaciones discursivas que lo forman [8]. Esta estructura puede incluir la estructura retórica del texto [9 p.792-3], por ejemplo, estructura narrativa o de argumentación.

Aunque existen varios ejemplos de sistemas de resumen de este tipo ([29],[30],[31],[32]) es importante mencionar que los sistemas de resumen automático que utilizan técnicas a nivel discursivo son poco comunes, debido a que requieren algunas técnicas de procesamiento del lenguaje natural que son complejas y su implementación es aún objeto de investigación. Por ejemplo, estos sistemas hacen uso de analizadores discursivos[3] que muchas veces no están disponibles para todas las lenguas pues aún no han sido desarrollados (como en el caso del español).

Finalmente, las técnicas de tipo híbrido pueden combinar procesamiento del texto a diferentes niveles para seleccionar la información importante. Entre los sistemas que adoptan una aproximación híbrida, se pueden mencionar a [33] y [34] que combinan estrategias tanto lingüísticas como estadísticas para el resumen de textos especializados del dominio médico en español.

[3] Herramientas que automáticamente construyen estructuras arbóreas para derivar la estructura discursiva de un texto.

2.3 Resúmenes de tipo extractivo

Los sistemas que generan resúmenes de tipo extractivo consideran al documento como un conjunto de enunciados para posteriormente extraer los que son considerados más relevantes, según diversos criterios, de tal manera que el resumen resultante es una selección de fragmentos del texto original.

Se profundizará en el funcionamiento de los sistemas de resumen de tipo extractivo y mono-documental puesto que son los más comunes, además de que el sistema propuesto en el tercer capítulo pertenece a este tipo.

Resumen extractivo mono-documental

Jurafsky y Martin [9] proponen una arquitectura (figura 2) para la generación de resúmenes extractivos de un sólo documento (mono-documental) en donde intervienen las siguientes etapas:

1. Selección del contenido
2. Ordenamiento de la información
3. Limpieza de los enunciados

A grandes rasgos, el proceso que los autores describen para generar el resumen consiste en: seleccionar del documento fuente los enunciados a extraer (selección de contenido), escoger el orden en que se colocarán enunciados extraídos (ordenamiento) y finalmente limpiar los enunciados, por ejemplo, resolver problemas de coherencia, fusionar varios enunciados en uno solo que contenga la información esencial, etc. (limpieza de los enunciados).

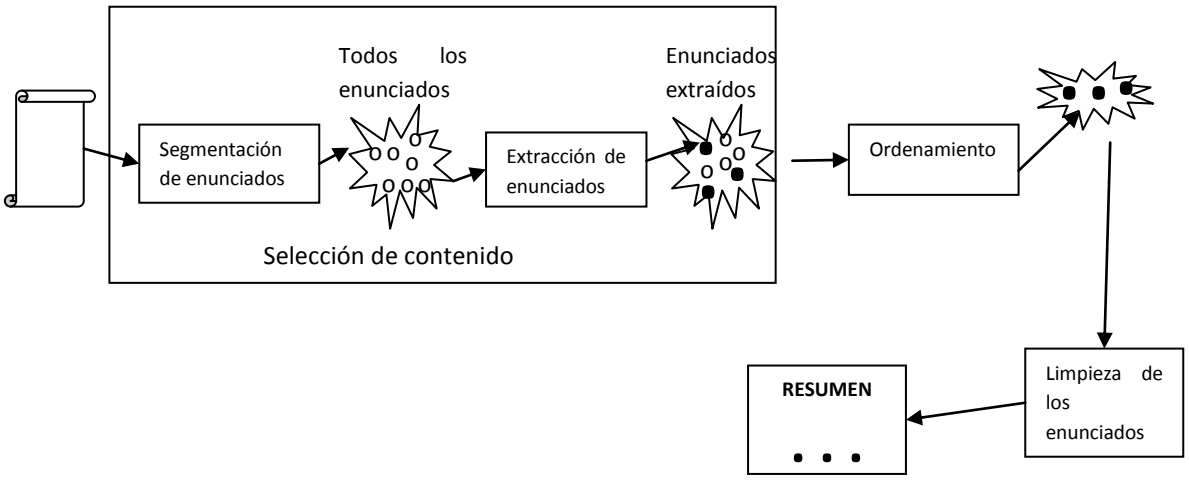


Fig.2. Arquitectura de un sistema de resumen extractivo mono-documental genérico

La etapa más importante es la de selección del contenido, que es donde se clasifica la relevancia de los enunciados que forman el documento y se seleccionan o extraen los más importantes. Como se puede ver en la figura 2, en esta etapa también interviene la segmentación del texto fuente en enunciados. Más adelante, en la sección "segmentación del texto", se hablará sobre las particularidades de este proceso.

Para determinar la importancia de los enunciados existen diversos algoritmos. Resulta conveniente hablar de una de las primeras aproximaciones hechas en el campo del resumen automático para la selección de contenido. Esta aproximación, propuesta por Luhn [3], utiliza técnicas a nivel superficie y establece que para determinar los enunciados que vale la pena extraer es necesaria una medida que permita comparar y calificar la información contenida en todos los enunciados, es decir, una medida que permita cuantificar la importancia del enunciado. Este factor de importancia se deriva de un análisis de las palabras que forman cada enunciado. Luhn propone que la frecuencia de ocurrencia de las palabras dentro del documento es una medida útil para determinar la importancia de cada palabra y por lo tanto del enunciado, es decir, los enunciados que posean un mayor número de palabras informativas o importantes serán buenos candidatos para ser extraídos y formar el resumen.

La justificación de medir la importancia de una palabra utilizando su frecuencia está basada, de acuerdo con Luhn, en el hecho de que un escritor comúnmente repite ciertas palabras conforme avanza y varía sus argumentos y también mientras desarrolla algún aspecto de un tema. Esta manera de enfatizar es tomada como un indicador de importancia. Otro factor que consideró para determinar la relevancia de los enunciados fue la posición relativa que ocupan las palabras dentro del enunciado.

Es importante mencionar que, al generar una lista de frecuencias de las palabras contenidas en un documento, no necesariamente las palabras con más altas frecuencias son las que poseen mayor contenido de información. De hecho, en teoría de la información son las más frecuentes las que menos información contienen (se hablará de esto más adelante). Así, las palabras más frecuentes tienden a ser del tipo funcional (pronombres, conjunciones, preposiciones, etc.) y al ser tan comunes no aportan información relevante del contenido del documento. Hay quien incluso considera estas palabras como "ruido" en el sistema. Luhn propuso técnicas para disminuir este supuesto ruido, por ejemplo, comparar las palabras del documento con una "lista de paro" que incluya palabras consideradas no informativas de tal manera que no se asignen valores de importancia a las palabras que coincidan con la lista de paro. Otro método consiste en establecer umbrales, de tal manera que no se incluyan las palabras con frecuencias más altas y tampoco las de frecuencias más bajas.

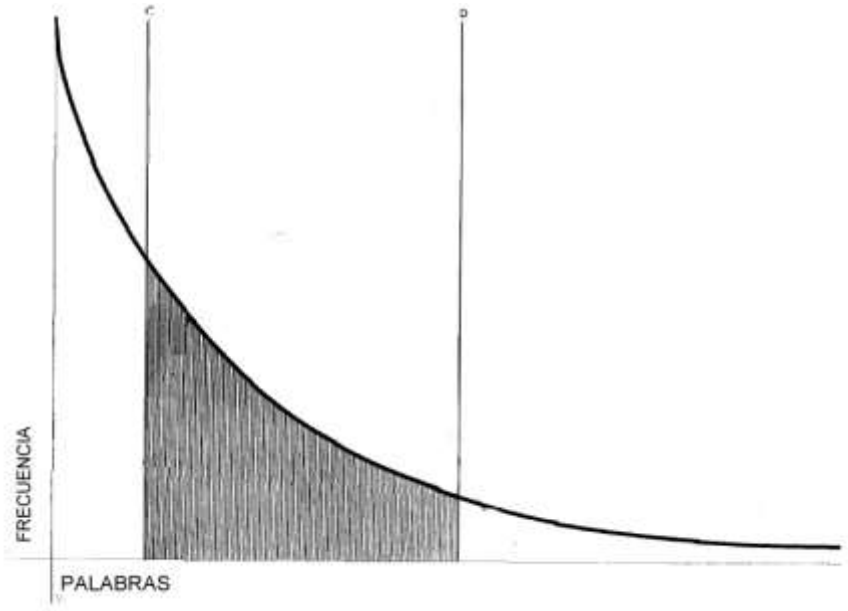


Fig. 3. Diagrama de frecuencia de las palabras

La curva mostrada en la figura 3 muestra la distribución implícita en muchos fenómenos lingüísticos y comunicativos y que se explica mediante la "Ley de Zipf" [10]. Específicamente, esta ley describe el comportamiento de las frecuencias de las palabras en los textos de prácticamente todas las lenguas. Básicamente establece que dentro de un texto sólo pocas palabras son muy frecuentes y la mayoría de palabras tienden a aparecer pocas veces. De hecho, aproximadamente la mitad son *hápax legomena*, esto es, típicamente ocurren sólo una vez [10 p.22]. En la gráfica mostrada en la figura 3, el eje de las abscisas representa las palabras únicas de un texto ordenadas por su frecuencia de aparición en el texto.

Luhn observó que las palabras muy comunes, y por lo tanto poco informativas, estarían presentes en la región de más alta frecuencia y propuso un corte para eliminar el "ruido" (línea C), de tal manera que sólo las palabras que estén a la derecha de este umbral son consideradas adecuadas para indicar relevancia. De la misma manera, la línea D representa un corte para las palabras con más baja frecuencia. Este umbral se establece porque las palabras de una sola ocurrencia suelen referirse a cuestiones tangenciales, nombres de personas u organizaciones incidentales o, muchas veces, errores de dedo [11].

Esta fue sólo una de las primeras propuestas para cuantificar la importancia de las palabras basándose en su frecuencia de ocurrencia dentro del documento. Con el tiempo se han desarrollado más modelos para la selección de contenido importante, no sólo tratando características a nivel a superficie sino también a nivel entidades del texto e incluso algunos pocos experimentos restringidos a un dominio que procesan el texto a un nivel más profundo. Sin embargo, en la última década se ha renovado el interés en las técnicas a nivel superficie que utilizan métodos estadísticos o cuantitativos.

También es importante mencionar que lo expuesto anteriormente se enfocó en la selección no supervisada del contenido, es decir, se trata de algoritmos que no requieren un conocimiento previo (datos a priori) de lo que se va a analizar. Sin embargo también existen métodos supervisados, es decir, aquellos métodos en los que es necesario entrenar al sistema con un conjunto de ejemplos, de tal manera que el sistema "aprenda" el comportamiento de los ejemplos y sea capaz posteriormente de predecir valores. Por ejemplo, hay sistemas de resumen automático en donde se entrena al sistema con un corpus formado por textos y sus respectivos resúmenes extractivos hechos por humanos [12], de tal manera que después de ser entrenado el sistema pueda clasificar qué enunciados son relevantes y generar resúmenes para otros textos.

2.4 SEGMENTACIÓN DEL TEXTO

Uno de los primeros pasos en el procesamiento de textos es dividir el texto en unidades para facilitar el análisis, procesamiento y extracción de información. Generalmente los textos son divididos en unidades más pequeñas, por ejemplo, en enunciados o palabras. A este proceso se le conoce como segmentación del texto *o tokenización*. En esta tesis le llamaremos tokenización al proceso de dividir un texto en palabras y segmentación al proceso de dividirlo en enunciados.

Dividir una cadena de caracteres es una actividad trivial para cualquier persona que esté acostumbrada a la estructura del lenguaje. Sin embargo, para un programa de computadora puede resultar difícil identificar los delimitadores de las unidades que se

quieren segmentar. De hecho, es un problema complejo para el cual no existe una solución única.

Por ejemplo, definir a qué se le considera una palabra es una cuestión controvertida en lingüística. No obstante, para fines prácticos de procesamiento de textos existe la noción de "palabra gráfica" [10 p.125-26], que se define como una cadena de caracteres alfanuméricos con un espacio en blanco en cualquiera de sus extremos y que además no incluye signos de puntuación como puntos y comas. Sin embargo, esta simplificación no resuelve todos los problemas pues no incluye casos como el de las cifras numéricas y monetarias, tampoco diferencia cuando se trata de una o dos palabras en el caso de que la "palabra" contenga guiones. Además, las palabras no siempre están rodeadas por un espacio en blanco, a veces están pegadas a un signo de puntuación como un punto, una coma, signos de exclamación o interrogación, etc. Y esto no siempre se resuelve eliminando las puntuaciones, pues a veces un punto en vez de indicar el final de un enunciado puede indicar una abreviatura, en cuyo caso, se prefiere dejar el punto para diferenciar la abreviatura de alguna otra palabra que se escriba igual pero tenga un significado distinto.

Las palabras gráficas también tienden a significar cosas juntas, por ejemplo, en grupos llamados "colocaciones" [10 p.151-155], es decir, palabras que ocurren consecutivamente de manera muy frecuente y que por lo tanto suelen ser tratadas como una sola ya que en su conjunto tienen un significado. Por ejemplo, palabras como "vino tinto", "Nueva España", "Nueva York", "no obstante", etc.

Por su parte, la delimitación de enunciados tampoco es un asunto trivial y es un problema común en el procesamiento de textos. Generalmente, para delimitar un enunciado se utilizan los signos de puntuación, en especial el punto; sin embargo, como se dijo, esto puede resultar ambiguo debido a que el punto puede indicar no solamente el final de un enunciado, sino abreviaturas, iniciales, numeraciones, etc. Existen diversos algoritmos para la detección de las fronteras o límites de un enunciado; estos algoritmos pueden

incluir expresiones regulares para detectar patrones, heurísticas, árboles de clasificación, redes neuronales, entre otros.

TOKEN Y TIPO

Como se dijo anteriormente, se le llama tokenización al proceso de segmentar una cadena de caracteres en unidades que comúnmente representan palabras, y a estas unidades se les llama *tokens*. Es importante distinguir entre dos conceptos: el tipo y el token.

Un token es la ocurrencia individual de la unidad dentro del texto mientras que el tipo es la clase a la que pertenecen esos tokens [13 p. 20]. Cada token es una instancia del tipo. Por ejemplo, si tokenizamos el siguiente enunciado:

la directora de la universidad

Resultarían cinco tokens pero solo cuatro tipos, debido a que hay dos instancias del tipo "la" que ocurre dos veces en el enunciado.

Para separar los tokens se utilizan como referencia ciertos caracteres que sirven como delimitadores. Por ejemplo, el espacio, el tabulador y el salto de línea, que generalmente actúan como delimitadores y no son contados como tokens. Sin embargo, como ya se mencionó antes, existen caracteres como los signos de puntuación que pueden ser delimitadores pero que en ciertas ocasiones también pueden ser parte de un token. Por ejemplo, un punto o una coma que se encuentre dentro de una cifra numérica no es un delimitador, más bien es parte del token.

2.5 CRITERIOS PARA ESTABLECER LA RELEVANCIA DE LOS ENUNCIADOS PARA UN EXTRACTO

Existen muchos criterios para determinar la elegibilidad de los enunciados de un texto como candidatos a formar parte de un extracto. En esta sección, se presentarán solamente las generalidades de los criterios que aplica el resumidor extractivo descrito en el siguiente capítulo: esquema TF-IDF, desviación estándar de la posición de las palabras y el contenido de información de las mismas en el documento a resumir.

ESQUEMA TF-IDF

TF-IDF es un esquema de asignación de pesos a las palabras que es común en el área de recuperación de la información y de minería de textos. Este esquema se utiliza usualmente cuando se tiene una colección de documentos y sirve para identificar las palabras representativas de cada documento con respecto a la colección en su conjunto.

El esquema TF-IDF asigna un mayor peso a las palabras que aparecen de manera frecuente en un documento pero que aparecen muy poco en toda la colección a la que pertenece dicho documento, sugiriendo que estas palabras con mayor peso son particularmente relevantes en ese documento. De esta manera, al encontrar términos que se limitan a aparecer en sólo unos cuantos documentos es posible diferenciar a esos documentos del resto de la colección.

El esquema de asignación de pesos TF-IDF es utilizado en diversas tareas de recuperación de la información por ejemplo en la identificación de palabras clave o *keywords*, en la clasificación de documentos, en buscadores de contenido, etc.

En el caso del resumen mono-documental, también es útil el esquema TF-IDF. Sin embargo, a diferencia de lo explicado previamente, en el sistema de resumen mono-documental no se tiene una colección de documentos. Más bien, se tiene un sólo documento que se puede ver como una colección de enunciados. Por lo tanto, se asignará mayor peso a las palabras que aparecen de manera frecuente en un enunciado (o en unos pocos), pero que aparecen muy poco en el resto del documento.

TF-IDF es la abreviación de *term frequency – inverse document frequency*, es decir, frecuencia del término - frecuencia inversa del documento.

El peso de las palabras en el esquema TF-IDF se obtiene a partir del siguiente producto vectorial [9 p.791]:

$$peso(w_i) = tf_{i,j} \times idf_i$$

Donde,

- $peso(w_i)$ es el peso de la palabra *i* en el enunciado que está siendo evaluado
- $tf_{i,j}$ es la frecuencia absoluta con la que ocurre un término *i* dentro del documento *j*
- idf_i es la frecuencia invertida del documento

Para un documento, la frecuencia invertida del documento (idf_i) se puede definir de la siguiente manera:

$$idf_i = \log\left(\frac{N}{n_i}\right)$$

Donde,

- N : número total de enunciados del documento
- n_i : número de enunciados que contienen al término *i*

Entre menos enunciados contengan al término *i*, mayor será el valor de su peso. El valor más bajo que puede tomar idf_i es 0 (log 1) y ese valor es asignado a los términos que aparecen en todos los enunciados.

MEDIDAS ESTADÍSTICAS DE VARIABILIDAD: DESVIACIÓN ESTÁNDAR

La desviación estándar es una medida de la variabilidad o dispersión de un conjunto de datos. Una forma de medir la variabilidad consiste en considerar las desviaciones de los valores de datos con respecto a un valor central. La varianza es una medida que toma en cuenta la distancia de cada dato con respecto a la media [10 p. 47-50], [14].

La varianza se denota mediante s^2 y se define a partir de los datos x_i.....x_n, con media $\bar{x} = \frac{\sum_{i=1}^{n} Xi}{n}$, como:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

A la raíz cuadrada positiva de la varianza se le denomina desviación estándar:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Entre más grande sea la desviación estándar, más dispersos se encuentran los datos y mientras más bajo sea el valor de desviación estándar, los datos tienden a estar cerca de la media y con una distribución uniforme. La utilidad de la desviación estándar reside en que, a diferencia de la varianza, está expresada en las mismas unidades que los datos.

Por ejemplo, si el conjunto de datos a analizar fueran las posiciones de las ocurrencias o "tokens" de una palabra dentro de un texto, la desviación estándar indicaría el promedio en que los tokens de dicha palabra se alejan de su posición media en el documento, es decir, qué tanto se alejan o se concentran en algún área del texto. Un valor alto indicaría que los tokens ocurren a lo largo de todo el texto, mientras que un valor bajo indicaría que están concentrados en un área particular del texto.

LA TEORÍA DE LA INFORMACIÓN Y EL CONTENIDO DE INFORMACIÓN.

La teoría de la información fue desarrollada por Claude E. Shannon en 1948. Se utiliza ampliamente en ingeniería para evaluar el funcionamiento de los sistemas de comunicaciones. Esta teoría suministra una medida cuantitativa de la información contenida en los mensajes y permite determinar la capacidad de un sistema de transferir esta información desde su origen hasta su destino. También ofrece las características de funcionamiento de un sistema ideal u óptimo. Esta teoría ha encontrado aplicaciones en muchas otras áreas, por ejemplo: inferencia estadística, el procesamiento del lenguaje natural, criptografía, entre otras.

La teoría de la información introduce el concepto de "contenido de información". Es importante aclarar que, desde el punto de vista de la teoría de la información, el concepto de información no es una expresión del significado o contenido semántico, más bien describe un nivel de incertidumbre. El contenido de información de un mensaje está estrechamente relacionado con su probabilidad de ocurrencia. Los mensajes que contienen noticias de gran probabilidad de ocurrencia, es decir, que indican muy poca incertidumbre en el resultado, conducen poca información. Por el contrario, los que contienen noticias con baja probabilidad de ocurrencia contienen grandes cantidades de información. Por lo tanto, el contenido de información de un mensaje es inversamente proporcional a la probabilidad de su ocurrencia [15].

Shannon propuso que el contenido de información asociado a un evento A que ocurre con una probabilidad PA se define como:

$$I_A = \log\left(\frac{1}{P_A}\right) = -\log\left(P_A\right)$$

Usualmente se utiliza el logaritmo en base dos para expresar en bits el contenido de información:

$$I_A = \log_2\left(\frac{1}{P_A}\right) = -\log_2\left(P_A\right)$$

De acuerdo con esta expresión, el evento que es menos esperado, es decir, el que tiene la probabilidad menor, deberá tener el mayor contenido de información y un evento que es certero, es decir que siempre ocurre y que por lo tanto tiene una probabilidad uno, no contendrá información alguna.

3. SISTEMA DE RESUMEN AUTOMÁTICO *RESÚMEME*

En este capítulo se presenta la propuesta del sistema de resumen automático que ocupa esta tesis. Se profundiza sobre las características que posee, su funcionamiento y los algoritmos elaborados para su desarrollo. En esencia, se propone un sistema de resumen automático de tipo extractivo y mono-documental al que se nombró "ResúmeME". Aquí se abordan los algoritmos y métodos utilizados para su desarrollo así como las características que posee.

Los parámetros que se consideraron para su diseño fueron los establecidos en la arquitectura general de un sistema de resumen de textos (fig. 1). A continuación se muestra la lista de requerimientos que se tomó en cuenta para el diseño y desarrollo del sistema *ResúmeME*.

- Requerimientos para el sistema de resumen automático

Audiencia: Genérico. El sistema de resumen automático deberá ser genérico, de tal manera que la información extraída para producir el resumen no responda a una petición o consulta en particular por parte del usuario.

Función: Indicativo. Los resúmenes generados por el sistema tendrán una función indicativa, es decir, brindarán una idea general sobre el contenido del texto o indicarán el tema central pero sin pretender profundizar exhaustivamente en los conceptos contenidos en el texto como sucede en los resúmenes informativos.

Fluidez: Fragmentos. La estructura de los resúmenes generados consistirá en fragmentos textuales extraídos del documento original. Estos fragmentos no estarán conectados entre sí mediante texto generado automáticamente.

Compresión: Para generar los resúmenes, el sistema permitirá que el radio de compresión del resumen sea definido por el usuario, con un valor por defecto del 20%.

Además, se busca que el sistema sea de tipo extractivo, es decir, que genere resúmenes formados por una combinación de enunciados seleccionados del documente fuente, así como mono-documental (que el resumen generado corresponda a un sólo documento). Finalmente, también se requiere que el sistema sea independiente del idioma y de la fuente, es decir, que las técnicas utilizadas para su desarrollo no estén orientadas a textos en una lengua en particular o que aborden un dominio temático específico.

3.1 DESARROLLO

A continuación se expone el sistema desarrollado en esta tesis, que buscó cumplir con los requerimientos antes mencionados.

El sistema de resumen automático *ResúmeME* recibe como entrada un archivo de texto plano[4]. Este archivo es el documento fuente a partir del cual se extrae la información para generar el resumen. La salida del sistema arroja un archivo de texto plano que contiene extractos del documento fuente, es decir, el resumen (fig. 4). La figura 5 muestra la arquitectura del sistema con los procesos y fases que intervienen y que se explicarán a continuación (los detalles técnicos sobre la implementación se abordarán en el marco práctico).

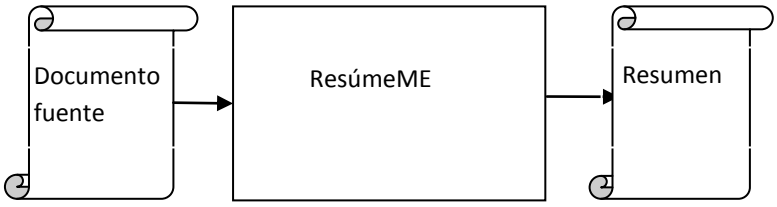


Fig. 4. Entrada y salida del sistema

[4] Los archivos que recibe el sistema son archivos de texto que no estén etiquetados con un formato especial, por ejemplo en XML.

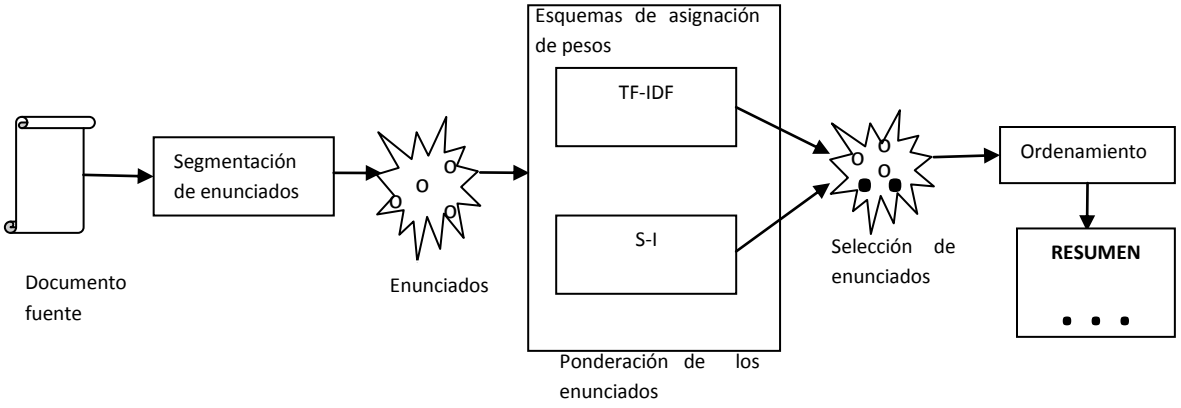


Fig. 5. Arquitectura de *ResúmeMe*

Segmentación de enunciados

En esta primera fase el documento fuente es procesado y dividido en unidades de menor longitud que representen enunciados.

Se utilizó el algoritmo para segmentación de enunciados PUNKT [16]. Se eligió este algoritmo porque es una aproximación no supervisada e independiente del lenguaje. Adicionalmente, una implementación de este algoritmo está disponible en la paquetería NLTK (*Natural Language Toolkit*) de Python que fue el lenguaje de programación utilizado para la realización de este sistema y del cual se hablará más adelante.

Este algoritmo se basa principalmente en la idea de que las abreviaturas ocasionan muchas de las ambigüedades que no permiten delimitar correctamente a los enunciados y que, por lo tanto, si se elabora un algoritmo capaz de identificar las abreviaturas, se podría mejorar significativamente la segmentación de enunciados. Para identificar las abreviaturas, este algoritmo construye un modelo que se apoya en tres criterios:

1. Las abreviaturas son similares a una colocación. En este caso la colocación estaría formada por dos elementos: una palabra truncada y el punto.

2. Las abreviaturas tienden a ser de poca longitud.

3. Las abreviaturas a veces contienen puntos internos.

Es importante mencionar que este algoritmo no hace uso de anotaciones especiales del texto, etiquetado POS o tablas pre-elaboradas con datos que ayuden a la identificación de abreviaturas y segmentación de los enunciados. La información que necesita para construir el modelo la extrae del propio corpus que va a segmentar.

Ponderación de los enunciados

Esta etapa es primordial pues tiene como objetivo clasificar los enunciados de acuerdo con su importancia, de tal manera que se pueda decidir cuáles son buenos candidatos para ser incluidos en el resumen.

Se propone un algoritmo no supervisado en donde el texto se analiza con técnicas a nivel superficie y se utilizan mediciones estadísticas para la selección del contenido relevante.

El algoritmo cuantifica la importancia de cada enunciado asignándoles un coeficiente al que se le llamó puntaje. El puntaje refleja la importancia de la información contenida en el enunciado y para calcularlo se utilizaron dos criterios: el esquema TF-IDF y el esquema S-I (este último combina la desviación estándar de la posición de las palabras y el contenido de información, se abordará más adelante).

Estos esquemas, a su vez, asignan un peso a las palabras de cada enunciado. De tal manera que el puntaje de los enunciados está en función de los pesos de sus palabras. En otras palabras, el puntaje del enunciado es un promedio de los pesos de sus términos.

Es importante mencionar que el esquema TF-IDF y el esquema S-I evalúan de manera distinta los enunciados y por lo tanto cada uno arroja una medición independiente de la otra. El puntaje final del enunciado corresponde a un promedio de las mediciones obtenidas a través de los distintos esquemas.

Este sistema de ponderación de los enunciados puede ser equiparado a un sistema de votantes en donde cada enunciado es evaluado por distintos votantes que, de acuerdo con sus criterios, juzgan la importancia del enunciado y cada uno le asigna un puntaje, de

tal manera que el puntaje total del enunciado es una combinación del valor emitido por los votantes.

Finalmente, una vez obtenidos los puntajes totales de todos los enunciados, se seleccionan aquellos que tengan los puntajes más altos (etapa de selección de enunciados). El número de enunciados seleccionados depende del radio de compresión definido por el usuario. Estos enunciados seleccionados serán los extractos que formen al resumen.

La etapa de ordenamiento simplemente consiste en acomodar los enunciados seleccionados de acuerdo con su orden de aparición en el texto original.

A continuación se definen de manera formal las mediciones utilizadas para evaluar la importancia de los enunciados del texto. Asimismo se explican los criterios en los que se basan los esquemas TF-IDF y S-I para determinar la importancia o representatividad de las palabras.

El puntaje total del enunciado se puede expresar de la siguiente manera:

$$puntaje(enunciado) = \frac{puntaje_{TF-IDF}(enunciado) + puntaje_{S-I}(enunciado)}{2}$$

Donde

$puntaje_{TF-IDF}(enunciado)$: es el puntaje del enunciado obtenido con el esquema TF-IDF

$puntaje_{S-I}(enunciado)$: es el puntaje del enunciado obtenido con el esquema S-I

3.2 ESQUEMA TF-IDF

Previamente en el marco teórico se habló del esquema TF-IDF que es ampliamente utilizado en diversas tareas de recuperación de la información y también en el resumen automático. Este esquema beneficia, es decir, asigna un mayor peso a las palabras que ocurren frecuentemente dentro de un enunciado pero que ocurren poco en el resto del documento. De manera contraria, las palabras que son frecuentes en un enunciado pero que además aparecen en los demás enunciados del texto son penalizadas puesto que tienden a ser palabras no indicativas del dominio o del tema del enunciado.

La asignación de pesos se realiza de la siguiente manera:

Sea $e \in E$ donde E es el conjunto de los enunciados que constituyen al documento fuente.

Sea $t \in T$ donde T es el conjunto de todas las palabras que aparecen en el documento.

El peso de cada palabra *t* perteneciente al enunciado *e*, utilizando el esquema TF-IDF, se obtiene a través del producto:

$$peso_{TF-IDF}(t,e) = TF_{t,e} \times \log\left(\frac{|E|}{|E_t|}\right)$$

Donde

$TF_{t,e}$: La frecuencia absoluta de la palabra *t* dentro del enunciado *e*.

$|E|$: Número total de enunciados en el documento fuente.

$|E_t|$: Número total de enunciados en donde aparece la palabra *t*.

Finalmente el puntaje del enunciado se calcula en función de los pesos de las palabras que lo forman. En este caso el puntaje del enunciado es un promedio suavizado [17] de los pesos. Se optó por este tipo de promedio porque beneficia a los enunciados largos que generalmente contienen más información, disminuyendo el impacto negativo que tiene la longitud del enunciado cuando se utiliza un promedio convencional.

$$puntaje_{TF-IDF}(e) = \frac{\sum_{t=1}^{n} TF - IDF_{t,e} * n}{\log(n)}$$

Donde

n: Número total de palabras que contiene el enunciado.

3.3 ESQUEMA S-I

La presente tesis propone un esquema de asignación de pesos al que se le denominó S-I. Este esquema combina el contenido de información de las palabras (la noción propuesta por Shannon en teoría de la información) así como su dispersión en el texto (desviación estándar). El peso asignado a cada palabra corresponde al promedio de estos dos valores, por lo tanto son beneficiadas las palabras que por su frecuencia en el texto tienen un valor alto de contenido de información (son poco frecuentes) y que además la desviación estándar de sus posiciones es alta (ocurren dispersas).

El esquema S-I se basa en la noción de que las palabras con mayor tendencia a ocurrir distribuidamente en el texto pueden ser relevantes. Sin embargo, también pueden ser palabras funcionales (artículos, pronombres, etc.) que ocurren constantemente y no aportan información del contenido propio del documento. Para contrarrestar lo anterior se agrega el criterio del contenido de información que penaliza a las palabras con frecuencias muy altas (recordemos que los eventos con probabilidad alta tienen bajo contenido de información).La asignación de pesos se realiza de la siguiente manera:

Sea $e \in E$ *donde* E *es el conjunto de los enunciados que constituyen al documento fuente.*

Sea $t \in T$ *donde T es el conjunto de todas las palabras que aparecen en el documento.*

El peso de cada palabra *t*, utilizando el esquema S-I, se obtiene promediando su contenido de información y la desviación estándar de las posiciones de esa palabra en el documento:

$$peso_{S-I}(t) = \frac{s_t + I(t)}{2}$$

Donde,

- S_t es la desviación estándar de las posiciones $x_i ... x_n$ de la palabra *t*

$$s_t = \sqrt{\frac{\sum_{t=1}^{n} (x_t - \bar{x})^2}{n-1}}$$

- $I(t)$ es el contenido de información de la palabra *t*

$$I(t) = \log_2\left(\frac{1}{p(t)}\right) = -\log_2\left(p(t)\right)$$

Finalmente el puntaje de cada enunciado es la media aritmética de los pesos de las palabras pertenecientes al enunciado.

$$puntaje_{S-I}(e) = \frac{\sum_{t=1}^{n} peso_{S-I}(t)}{n}$$

Donde:

- n : Número total de palabras que contiene el enunciado.

3.4 NORMALIZACIÓN

Para la adecuada asignación de pesos y puntajes es necesario normalizar diversos valores. Los puntajes de los enunciados bajo el esquema TF-IDF y el S-I se normalizan para que posteriormente, al promediar estos dos valores, se obtenga el puntaje final del enunciado.

Otros valores que se deben normalizar son la desviación estándar y el contenido de información que se calculan para cada palabra en el esquema S-I.

Se utilizó la técnica conocida comúnmente como min-max. Esta técnica sitúa el rango de valores a normalizar entre 0 y 1. Esta transformación se define como:

$$v' = \frac{v - m}{M - m}$$

Donde,

- v' es el nuevo valor
- v es el valor antes de ser normalizado
- m es el valor mínimo del conjunto de valores a ser normalizados
- M es el valor máximo del conjunto de valores a ser normalizados

3.5 Mecánica del funcionamiento combinado de los criterios aplicados

La aplicación conjunta de los criterios TF-IDF y S-I puede parecer contradictoria ya que, como se dijo arriba, el esquema TF-IDF privilegia aquellas palabras que ocurren en menos lugares, mientras que la desviación estándar (S) privilegia las palabras que ocurran más distribuidamente.

Sin embargo, considérese que el contenido de información (I) también privilegia aquellas palabras que ocurren poco. Y esto no constituye una contradicción, ya que el conjunto de criterios termina asignando mayores pesos a aquellas palabras que ocurren poco pero ocurren más de una vez (arriba se argumentó que los hápax tienden a representar información tangencial al contenido del documento) con alta intensidad en unos pocos enunciados (altos valores de TF-IDF) que están muy repartidos en el documento (la mayor desviación estándar posible indicaría que ocurren tanto muy al principio como muy al final del texto).

Esta mecánica será evaluada más adelante. Lo que implica que en un futuro otras mecánicas pueden incorporarse al desarrollo para comparar su funcionamiento con la seleccionada en este experimento y que constituye esta tesis.

4. IMPLEMENTACIÓN DE *RESÚMEME*

En este capítulo se abordan las cuestiones técnicas sobre la implementación del sistema. Esto es, se justifican aspectos como el lenguaje de programación elegido y la estructura del programa, entre otros.

4.1 MARCO PRÁCTICO

EL LENGUAJE DE PROGRAMACIÓN PYTHON

Python [18] es un lenguaje de programación de alto nivel y de propósito general. Fue creado a principios de los años noventa por el informático Guido van Rossum. La filosofía de su diseño hace énfasis en la legibilidad del código, *"Combine remarkable power with very clear syntax"*[5]. Es un lenguaje de programación orientado a objetos. Sin embargo, permite utilizar otros paradigmas de la programación, por ejemplo, la programación estructurada y la programación funcional.

Python es un lenguaje de programación interpretado, es decir, los programas son ejecutados por un intérprete sin pasar por una etapa de compilación. Otra característica de este lenguaje es su portabilidad: Python puede utilizarse en diversos sistemas operativos (variantes de Unix/Linux, MacOS, Windows, OS/2), además de ser extensible, pues permite integrar módulos escritos en lenguaje C o C++.

Una de las principales características de Python es que contiene una amplia librería estándar que proporciona herramientas pre-programadas útiles para diferentes tareas, por ejemplo, procesamiento de texto, protocolos de internet, ingeniería de software, sistemas operativos etc. También es posible incorporar librerías externas que abarcan muchas otras áreas.

[5] Combinar gran poder con una sintaxis muy clara.

Se eligió Python para la implementación del sistema de resumen automático *ResúmeME*, al considerar las siguientes características:

- Es software libre. Esto implica que no existe un costo por descargarlo, usarlo, o incluirlo en una aplicación, incluso en aplicaciones comerciales. Además su licencia permite modificarlo y redistribuirlo libremente.
- Es portable.
- Permite utilizar una gran gama de librerías, entre ellas la librería NLTK (*Natural Language Toolkit*) para el procesamiento del lenguaje natural, que facilita la implementación del sistema *ResúmeME*.
- Su sintaxis es simple y elegante lo que resulta en programas más fáciles de escribir y de leer.

NLTK (NATURAL LANGUAGE TOOLKIT)

NLTK [19] es un conjunto de módulos, información lingüística, tutoriales y ejercicios diseñados para el procesamiento del lenguaje natural tanto estadístico como simbólico. Este conjunto de módulos está escrito en Python y se distribuye bajo una licencia de código abierto. NLTK fue creado originalmente en el año 2001 como parte de un curso de lingüística computacional en la Universidad de Pennsylvania. Desde entonces se ha vuelto popular tanto en el campo de la enseñanza como en el de la investigación. Se ha ido expandiendo y optimizando gracias a la colaboración de desarrolladores en todo el mundo.

NLTK es una herramienta que resulta de gran utilidad para tareas de procesamiento del lenguaje como: procesamiento y filtrado del texto, etiquetado POS, clasificación, análisis sintáctico (*parsing*), métricas de evaluación, probabilidad y estimación, entre otras.

Para el desarrollo del sistema se utilizó Python 2.6.2 con las librerías NLTK 2.0b4 y Numpy 1.3.0rc2. Esta última es una librería que contiene diversas funciones matemáticas.

ESTRUCTURAS DE DATOS EN PYTHON

Resulta conveniente explicar de manera breve dos de las principales estructuras de datos disponibles en Python: las listas y los diccionarios. Estas estructuras fueron utilizadas en la implementación de *ResúmeME* y se mencionan constantemente más adelante, en el apartado de la implementación.

Listas

En Python, una lista (*list*) es una estructura de datos utilizada para agrupar valores, similar a un arreglo en otros lenguajes. Las listas no poseen un tamaño predefinido, se expanden dinámicamente mientras se agregan elementos; también son mutables, es decir, pueden ser modificadas o actualizadas una vez creadas. Los elementos de una lista no necesitan ser del mismo tipo y existen diversas operaciones y métodos que pueden ser aplicados a las listas.

Para generar una lista, los elementos se deben escribir separados por una coma y entre corchetes como se muestra en el ejemplo.

Ejemplo de una lista:

Lista= [1, “palabra”,0.2¨]

Diccionarios

Los diccionarios son estructuras de datos comúnmente conocidas en otros lenguajes como “memorias asociativas” o “arreglos asociativos”. Los diccionarios son colecciones de objetos, en donde los elementos son indexados por una llave y no por su posición relativa como sucede en las listas. Las llaves pueden ser cadenas de texto o números, en general cualquier tipo de dato que sea inmutable.

Un diccionario se puede entender como un conjunto formado por pares de la siguiente forma: *llave:valor.* Con el requisito de que las llaves deben ser únicas y no repetirse dentro del diccionario. Para crear un diccionario, se deben escribir los pares *llave:valor* separados por una coma y entre los signos {}, como se muestra en el ejemplo abajo.

Entre las principales operaciones que se pueden realizar en un diccionario se encuentran: almacenar un valor con alguna llave, extraer el valor dada la llave, borrar del diccionario un par *llave:valor* , etc.

Ejemplo de un diccionario:

D={'hola': 3, 'es': 5, 'México': 1}

4.2 IMPLEMENTACIÓN DEL SISTEMA

El sistema *ResúmeME* fue escrito en el lenguaje Python. El programa se encuentra estructurado en tres módulos: dos módulos que se encargan de calcular los puntajes de los enunciados bajo el esquema TF-IDF y el esquema S-I respectivamente y un módulo principal en donde se asignan los puntajes finales a cada enunciado para así seleccionar aquellos que formarán al resumen.

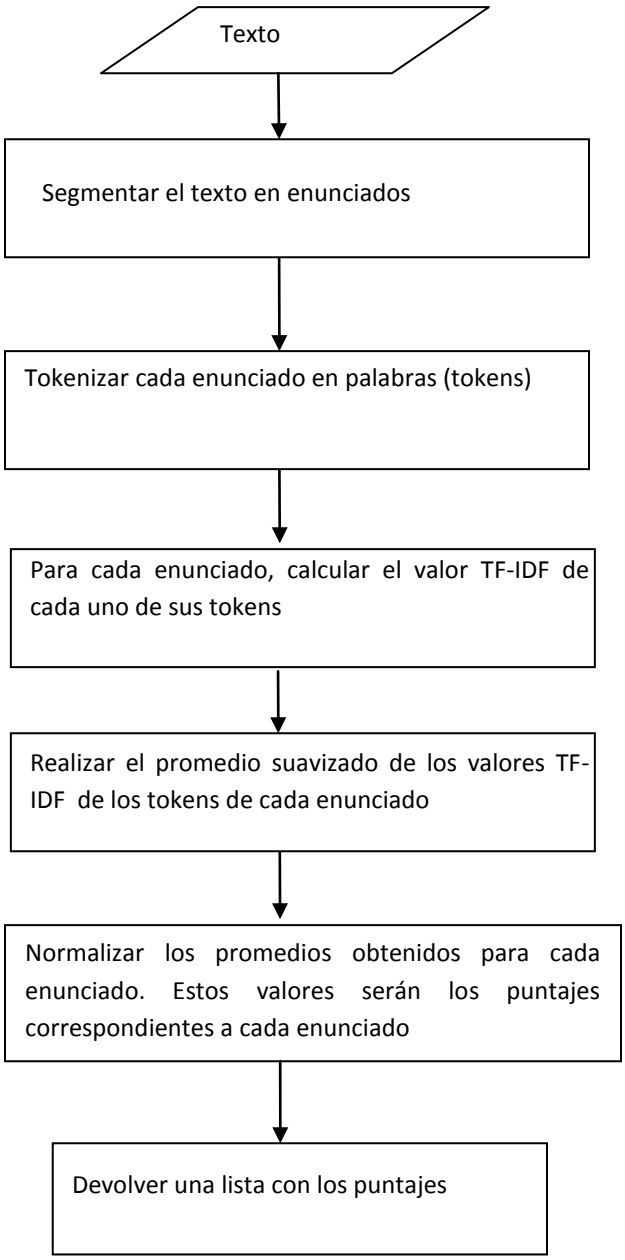
A continuación se profundiza en el funcionamiento de estos tres módulos, a los que se llamó: módulo TF-IDF, módulo S-I y módulo principal. Para cada módulo se presenta un diagrama de flujo que esquematiza las etapas involucradas. En las siguientes secciones se presentan los diagramas así como algunas especificaciones técnicas sobre su implementación.

4.3 MÓDULO TF-IDF

Entrada: Documento fuente.

Salida: Lista con puntajes correspondientes a los enunciados del documento fuente.

Diagrama:



Las siguientes consideraciones y especificaciones técnicas están relacionadas con las etapas y procesos involucrados en este módulo TF-IDF.

Segmentación de texto en enunciados

La segmentación del texto en enunciados se realiza con el algoritmo PUNKT, del que ya se habló antes. Es importante mencionar que, aunque este algoritmo es capaz de construir un modelo para detectar los límites de los enunciados a partir del texto que va a segmentar, se utilizó el algoritmo pero con un modelo pre-generado para el idioma inglés, disponible en NLTK.

Se optó por el modelo pre-generado porque todas las lenguas indoeuropeas (como el español y el inglés) comparten esencialmente el mismo sistema de puntuación[6] y este modelo al ser generado con una gran cantidad de documentos puede tener un mejor desempeño que uno generado con un sólo documento de poca longitud.

En NLTK, la implementación del algoritmo PUNKT está disponible en la clase *PunktSentenceTokenizer* (*Package tokenize :: Module punkt :: Class PunktSentence-Tokenizer*).

En las siguientes líneas de código se ilustra cómo cargar el modelo para el inglés, posteriormente se utiliza el método *tokenizer* de la clase *PunktSentenceTokenizer* para poder segmentar el texto en enunciados. Los enunciados resultantes se almacenan en la variable enunciados que es un dato tipo lista.

```
sent_tokenizer=nltk.data.load('tokenizers/punkt/english.pickle')
enunciados = sent_tokenizer.tokenize(texto)
```

[6] El sistema de puntuación de las lenguas occidentales se estableció desde el siglo XVIII, gracias a la propuesta de Aldo Manuzio un siglo antes (XVII) de utilizar los signos como marcas de las sintaxis [20].

Segmentación por palabra

Para segmentar por palabra cada enunciado se utilizó la función *nltk.tokenize.regexp_tokenize()* del módulo *regexp* (*Package tokenize :: Module regexp*). Esta función permite dividir una cadena de texto utilizando una expresión regular. Los parámetros que recibe son la cadena de texto y la expresión regular, regresando la lista de tokens que coincidan.

En las siguientes líneas de código se ilustra el proceso de segmentar por palabra a cada enunciado. Se utilizó la expresión regular *\w+* que divide al texto en sub-cadenas formadas exclusivamente por caracteres alfanuméricos (incluido el guión bajo) separadas por espacios, es decir, palabras gráficas. La expresión[7] es equivalente a [a-zA-Z0-9_]+.
Los enunciados *tokenizados* se van almacenando en una lista llamada *enunciados_tokenizados*

```
for enunciado in enunciados:
enunciados_tokenizados.append(nltk.tokenize.regexp_tokenize(enunciado.lower(), r'\w+'))
```

Cálculo del valor tf-idf

Para calcular el valor TF-IDF de cada una de los tokens se utilizó el método *tf_idf* de la clase *textcollection* (*Module text :: Class TextCollection*). Este método recibe como parámetros el token y el enunciado al que pertenece, devolviendo el valor TF-IDF del token dentro de ese enunciado. Para utilizar este método se debe crear un objeto de clase *TextCollection*; dicho objeto se inicializa pasándole como parámetro la lista de todos los enunciados tokenizados que forman al texto.

```
coleccion_enunciados=nltk.TextCollection(enunciados_tokenizados)
puntuacion=puntuacion+coleccion_enunciados.tf_idf(palabra,enuntok)
```

[7] Los caracteres alfanuméricos de la expresión regular incluyen caracteres como las letras acentuadas, la ñ, y otros caracteres que se consideren alfanuméricos de acuerdo con la configuración local en donde se encuentre el programa.

Normalización

Una vez obtenido el puntaje para cada enunciado se realiza una normalización de los valores de tal manera que el rango de puntajes se sitúe entre 0 y 1. Se utilizó la técnica conocida como min-max y que ya se expuso previamente.

Lista de puntajes

Finalmente, este módulo devuelve una estructura de datos de tipo lista que contiene el conjunto ordenado de puntajes correspondientes a cada enunciado. Por ejemplo en la lista *puntuacionesTFIDF* que se muestra, el primer elemento sería el puntaje correspondiente al primer enunciado del texto, el segundo elemento sería el puntaje del segundo enunciado y así sucesivamente.

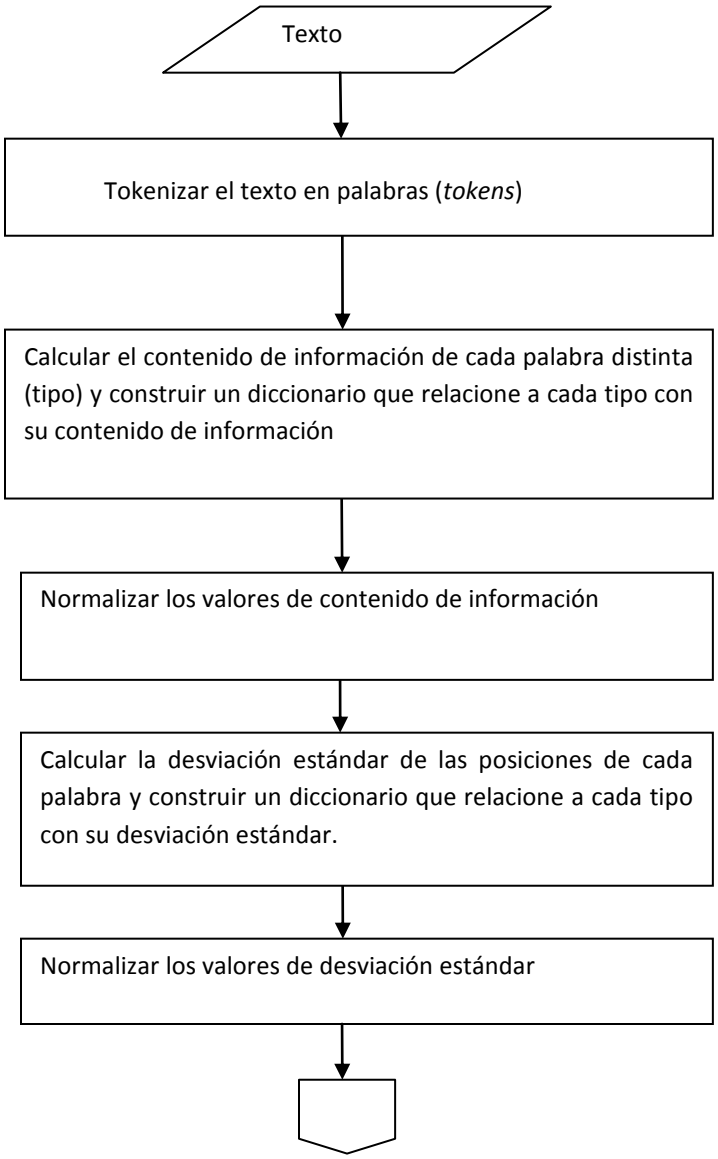
```
puntuacionesTFIDF=[0.5,0.9,0.61…]
```

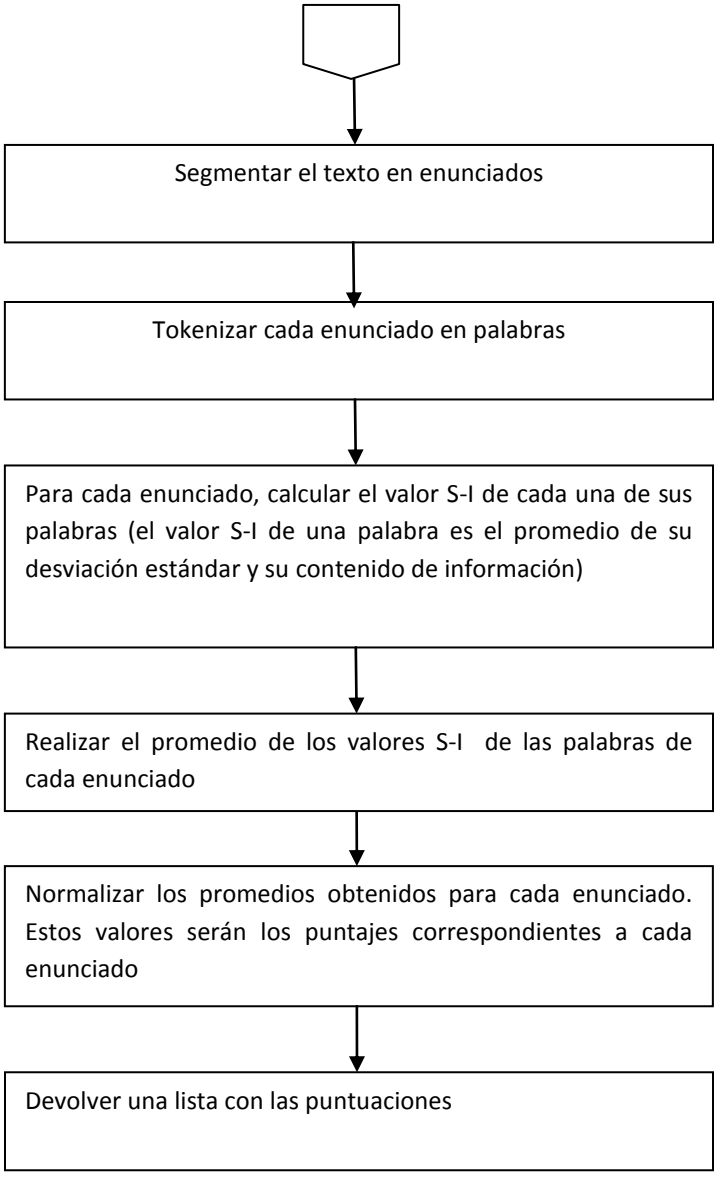
4.4 MÓDULO S-I

Entrada: Documento fuente.

Salida: Lista con puntajes correspondientes a los enunciados del documento fuente.

Diagrama:





A continuación se mencionan algunas consideraciones y especificaciones técnicas relacionadas con las etapas y procesos involucrados en este módulo. S-I

Segmentación por palabra

La segmentación del texto palabra por palabra se lleva a cabo con la función *nltk.tokenize.regexp_tokenize()* y la expresión regular *\w+* de la misma manera que en el módulo TF-IDF.

Cálculo del contenido de información de las palabras y construcción de un diccionario

Para calcular el contenido de información de las palabras, es necesario calcular primero su frecuencia y posteriormente su probabilidad de ocurrencia en el texto. Utilizando la clase *FreqDist* de nltk se construyó un diccionario que almacenó la probabilidad de cada tipo, es decir, un diccionario que tiene como llave el tipo y como valor la probabilidad (tipo:probabilidad).

Una vez obtenidos estos valores, se calculó para cada tipo su contenido de información ($\log_2\left(\frac{1}{P_A}\right)$) y se construyó el diccionario que relacionara a cada tipo con su contenido de información (tipo:contenido de información).

Normalización

Todos los procesos de normalización en el sistema se llevan a cabo con el algoritmo *min-max* explicado previamente.

Cálculo de la desviación estándar de las posiciones de cada palabra y construcción de un diccionario

Para calcular la desviación estándar es necesario conocer el conjunto de posiciones dentro del texto en las que ocurre cada palabra distinta. Se construyó un diccionario que relacionara a cada tipo de palabra con las posiciones en las que ocurren sus tokens. El siguiente es un ejemplo de un diccionario en donde las llaves son los tipos, y el valor para cada llave es una lista que contiene las posiciones de sus tokens dentro del texto.

{'lograban': [3], 'es': [1,6,10], 'no': [4,8]}

Una vez obtenidas las posiciones, se calcula la desviación estándar para cada conjunto y así se construye finalmente un diccionario que relaciona cada tipo con la desviación estándar de las posiciones de sus tokens.

Segmentación del texto en enunciados

Al igual que en el módulo anterior, la segmentación del texto en enunciados se realiza con la implementación del algoritmo PUNKT disponible en nltk y que ya ha sido explicado.

Cálculo del valor S-I para cada una de las palabras y asignación de pesos a los enunciados.

El esquema S-I, como ya se vio antes, asigna un peso a cada palabra combinando el valor del contenido de información y de la desviación estándar. El valor S-I corresponde al promedio de estos dos valores. Utilizando los dos diccionarios previamente construidos se calcula para cada tipo su valor S-I. Posteriormente, el puntaje de cada enunciado será el promedio del valor S-I de cada una de sus palabras.

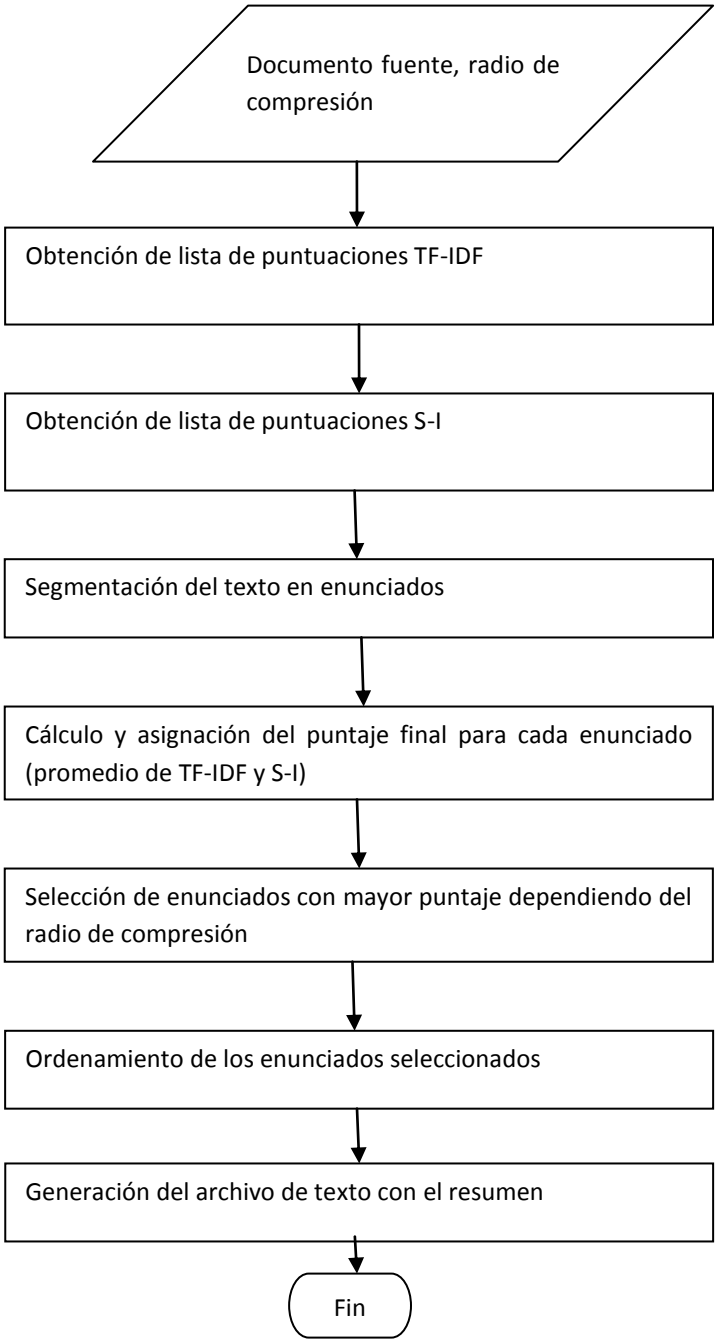
Finalmente, el módulo principal se muestra a continuación; en esencia, integra los dos módulos explicados hasta ahora.

4.5 MÓDULO PRINCIPAL

Entrada: Documento fuente, radio de compresión

Salida: Archivo de texto con el resumen.

Diagrama:



El diagrama de flujo de este módulo principal ilustra el funcionamiento general del programa, es aquí donde se integran los dos módulos anteriores (módulo TF-IDF y módulo S-I) y donde se genera el resumen.

A continuación se mencionan algunas consideraciones y especificaciones técnicas sobre las diversas etapas y procesos involucrados en este módulo.

Obtención de listas de puntajes TF-IDF y S-I

Los dos módulos que se explicaron previamente son en realidad funciones definidas por el usuario que se importan en el programa principal. Estas funciones se mandan llamar con el nombre de la función y los argumentos que recibe. En este caso las dos funciones reciben como único argumento el texto y devuelven una lista con puntajes. Estas listas son utilizadas posteriormente para la ponderación final de los enunciados. Ejemplo:

```
Lista1=TFIDF(texto)
Lista2=SI(texto)
```

Cálculo y asignación del puntaje final para cada enunciado

El puntaje final de cada enunciado se obtiene promediando dos valores: el puntaje del enunciado utilizando el esquema TF-IDF y el puntaje utilizando el esquema S-I. Las listas obtenidas en las etapas anteriores contienen estos dos valores de manera ordenada, de tal forma que si se quiere calcular el puntaje del primer enunciado sólo es necesario promediar el primer elemento de cada lista y así sucesivamente.

Selección de enunciados con mayor puntaje y ordenamiento

Una vez obtenidos los puntajes finales para cada enunciado, se deben seleccionar aquellos que hayan resultado con el puntaje más alto. El número de enunciados seleccionados depende del radio de compresión. Por ejemplo, si el radio de compresión es del 20%, se seleccionan el número de enunciados con puntajes más altos que representen 20% del total de enunciados.

Finalmente, los enunciados seleccionados se ordenan de tal manera que se preserve el orden en el que aparecen en el texto original.

Generación del archivo de texto con el resumen

En esta última etapa se genera el resumen que será presentado al usuario final. El resumen consiste en un listado de los enunciados o extractos seleccionados previamente. Se crea un archivo de texto plano que contiene a los enunciados. Este archivo conserva el nombre del documento fuente pero con la extensión *".resumen"*.

También se crean dos archivos adicionales, uno que muestra, además de lo enunciados seleccionados, el puntaje que obtuvo cada uno de ellos y otro aún más detallado que muestra el puntaje final así como el puntaje S-I y TF-IDF para todos los enunciados del texto. Estos archivos tienen la extensión *".detallado1"* y *".detallado2"* respectivamente.

Es importante mencionar que en un inicio la primera versión del sistema de resumen automático era parte del proyecto "Canario muerto", en donde era un módulo que se mandaba a llamar y se ejecutaba en un servidor GNU/Linux y posteriormente se integraban los archivos de texto generados al sistema principal. Debido a que la ejecución del módulo era a nivel consola no fue necesario desarrollar una aplicación de escritorio. Incluso, los resúmenes generados en esta tesis para la evaluación (capítulo 5) fueron hechos con la aplicación a nivel consola. Sin embargo, para facilitar la interacción con el usuario y que el sistema estuviera disponible en línea, se desarrolló también una interfaz web sencilla en donde se integró el sistema hecho en Python (ver Apéndice A. para una impresión de pantalla de la interfaz web).

5. EVALUACIÓN

En este capítulo se examina la cuestión de la evaluación en los sistemas de resumen automático y se propone un esquema para evaluar el sistema desarrollado. Al final, se analizan los resultados obtenidos.

5.1 EL PROBLEMA DE LA EVALUACIÓN EN LOS SISTEMAS DE RESUMEN AUTOMÁTICO

La evaluación de los sistemas de resumen automático ha sido un aspecto de gran interés en esta área, sobretodo porque hay una falta de consenso acerca de los métodos y tipos de evaluación que resultan apropiados.

Los sistemas de resumen automático se evalúan con el fin de conocer mejor su utilidad y desempeño. Es común que las evaluaciones sean diseñadas por las propias personas que elaboraron el sistema y por lo tanto están enfocadas a evaluar de manera específica al tipo de resúmenes que genera dicho sistema.

Evaluar un sistema de resumen automático es una tarea que enfrenta diversos retos, entre los que se encuentran [21]:

- Los sistemas de resumen automático producen como salida un resumen y resulta difícil determinar si esta salida es la correcta o no, a diferencia de otro tipo de sistemas. Por ejemplo, en casos donde la salida del sistema es la respuesta a una pregunta se tiene la noción de lo que podría calificarse como respuesta correcta. Sin embargo en un resumen automático siempre existe la posibilidad de que el sistema haya generado un buen resumen y que este sea muy diferente de cualquier resumen hecho por un humano que se utilice para aproximar o calificar lo que representaría una salida correcta. Es decir, la subjetividad es un problema inherente a la evaluación pues no hay un consenso absoluto acerca de cuál es el mejor resumen para un documento.

- El hecho de que los humanos sean requeridos para juzgar la salida del sistema implica un mayor gasto en la evaluación. De allí que se hayan diseñado evaluaciones automáticas que prescindan en buena medida de los jueces, utilizando algún programa de computadora que permita que el experimento sea fácilmente repetible.
- Un factor que interviene en la generación de un resumen es el radio de compresión así que es importante poder evaluar los resúmenes a diferentes radios. Esto incrementa la escala y complejidad de la evaluación.

Existen numerosos métodos para la evaluación de los resúmenes. Sin embargo, estos se pueden englobar en dos categorías principales [22]. La primera categoría es la evaluación intrínseca que juzga la calidad del resumen generado y toma en cuenta aspectos como: cobertura del resumen de las ideas esenciales, la coherencia o fluidez, la similitud del resumen generado con un "resumen ideal", etc. Por su parte, la evaluación extrínseca juzga qué tan útil es el resumen generado para la realización de alguna tarea posterior, por ejemplo, comprender las ideas principales del texto basándose únicamente en la lectura del resumen, poder contestar preguntas o recrear el documento fuente.

Uno de los criterios más usados para evaluar la calidad del resumen (en los métodos de evaluación intrínseca) es comparar el resumen generado por el sistema con un resumen de referencia. Este resumen de referencia, también llamado resumen ideal o *golden standard* es generalmente creado por un humano. El problema con este método radica en el hecho de que establecer un resumen ideal puede ser difícil y subjetivo, puesto que hay evidencia de que los humanos difieren al seleccionar los enunciados que consideran más importantes, además de que puede haber más de un resumen que represente de manera efectiva el contenido del documento.

Los sistemas de evaluación automática comparan, por medio de un programa de computadora, el resumen a evaluar con un conjunto de resúmenes "ideales" elaborados por humanos. Para determinar qué tanto se "parece" el resumen evaluado a los resúmenes ideales, estos sistemas automáticos se valen de diferentes técnicas y

mediciones, por ejemplo, el sistema de evaluación ROUGE [36] mide la co-ocurrencia de n-gramas entre los resúmenes a ser evaluados y los resúmenes ideales. ROUGE es uno de los sistemas de evaluación más populares y suele utilizarse en las competencias internacionales de resumen automático, por ejemplo en DUC (*The Document Understanding Conference*) [37] y en *Text Summarization Challenge* [38]. Algunos otros sistemas de evaluación automática son el *Basic Elements* [39] y el *Pyramid Method* [40].

Los sistemas automáticos de evaluación antes mencionados no prescinden completamente de jueces humanos puesto que necesitan a jueces para crear los resúmenes "ideales". Sin embargo, existen sistemas que comparan los resúmenes a ser evaluados con resúmenes generados por computadora. Algunos ejemplos de estos sistemas totalmente automáticos sin modelos humanos son los siguientes: [41],[42].

Es importante recordar que así como hay métodos automáticos, existen métodos "manuales" que también comparan el resumen generado por el sistema con un resumen generado por un humano o por un sistema de cómputo o incluso pueden compararlo con el documento fuente para determinar qué tanta información de la fuente está presente en el resumen. Sin embargo, la diferencia de este tipo de evaluación es que la comparación es realizada directamente por jueces humanos.

De hecho, los primeros métodos propuestos para la evaluación de resúmenes automáticos fueron de tipo manual. Por ejemplo, la evaluación presentada en [4] utiliza jueces humanos para evaluar al resumen generado automáticamente comparándolo con un resumen "ideal" creado por un humano. Otros ejemplos son [43] en donde los jueces comparan el resumen resultante con el documento original. En [35] los jueces contestan una serie de preguntas relacionadas con el documento fuente basándose únicamente en la lectura del resumen, posteriormente se comparan sus resultados con el de personas que sí hayan leído el documento fuente entero.

5.2 EVALUACIÓN DE *RESÚMEME*

El esquema de evaluación que se diseñó para *ResúmeME* prescinde de un resumen ideal o de referencia. En vez de esto se optó por evaluar la calidad del resumen comparándolo con el documento fuente. De manera general, el experimento consistió en proporcionarle a jueces humanos tanto el documento fuente como el resumen generado con *ResúmeME*. Esto fue posible gracias a la disponibilidad de voluntarios dispuestos a participar sin conocer la naturaleza exacta del experimento.

Ya que la duda de una evaluación objetiva es central en este trabajo, ellos juzgaron de manera subjetiva la calidad del resumen basándose en qué tantas de las ideas clave o esenciales del documento fuente habían sido abarcadas según cada uno de ellos. De esta manera, la apuesta en la evaluación no descansa en un ideal dudable, sino en un consenso de subjetividades.

Así, se les pidió a trece voluntarios leer un total de 11 textos cada uno. El conjunto de textos fue el mismo para cada voluntario excepto por el orden. Cada texto tenía anexado su resumen correspondiente.

Los resúmenes, formados por un conjunto de fragmentos extraídos del documento fuente, se generaron con un radio de compresión que osciló entre el 20% y 25% de la longitud del texto fuente[8]. Los once textos fueron seleccionados de manera aleatoria, eran textos en español, principalmente noticias y artículos de periódico, fragmentos de obras, biografías y artículos informativos de enciclopedia (ver Apéndice B. en donde se presentan algunos[9] de estos textos junto con los extractos generados). Los textos tenían una longitud no menor a una cuartilla (aprox. 2,100 caracteres) y no mayor a 3 cuartillas (aprox. 8,200 caracteres).

[8] Radio de compresión medido en función de la cantidad de caracteres del texto.

[9] Debido a la cantidad de textos y con el fin de economizar el espacio, en el apéndice sólo se muestra un subconjunto con los textos que resultaron mejor y peor evaluados para el idioma inglés y español. Así como los tres textos en alemán, chuj y tarahumara respectivamente.

Adicionalmente, con el fin de complementar la evaluación, se realizó la misma dinámica para otras lenguas, es decir, se tuvieron jueces para diferentes lenguas[10] que leyeron textos con características similares a las antes mencionadas.

El número de jueces que participaron para cada lengua y la cantidad de textos que leyeron se muestran en la Tabla 1.

Tabla1: Número de jueces y textos para cada lengua.

Lengua	Número de jueces	Número de textos
español	13	11
inglés	5	5
alemán	2	1
chuj	1	1
tarahumara	1	1

La evaluación por parte de los jueces consistió en contestar dos preguntas una vez leído el documento fuente y el resumen correspondiente. Las preguntas en todos los casos fueron las siguientes

Pregunta 1:

Consideras que los fragmentos extraídos:

a) *Todos capturan ideas relevantes del texto original*
b) *Muchos capturan ideas relevantes del texto original*
c) *Algunos capturan ideas relevantes del texto original*
d) *Ninguno captura alguna idea relevante del texto original*

[10] Los jueces no eran hablantes de lengua materna de estas lenguas, a excepción del español.

Pregunta 2:

Consideras que las ideas principales del texto original:

a) *Fueron completamente omitidas por los fragmentos*
b) *Muchas fueron omitidas por los fragmentos*
c) *Sólo algunas fueron omitidas por los fragmentos*
d) *Todas las ideas principales están contenidas en los fragmentos*

Se realizaron estas preguntas con el fin de medir dos parámetros de la calidad del resumen: la precisión y la exhaustividad (esta última también es conocida como recall o cobertura).

En el área de recuperación de la información, la precisión y la exhaustividad son medidas evaluativas que se usan típicamente para tener una idea de qué tan efectiva o eficiente es una búsqueda o consulta [23]. Cuando se hace una consulta, la precisión indica la proporción de documentos recuperados que fueron relevantes. Mientras que la exhaustividad indica la proporción de documentos relevantes que fueron recuperados del total de documentos relevantes en toda la base de datos. Expresado de otra manera:

$$precisión = \frac{número\ de\ documentos\ relevantes\ recuperados}{número\ de\ documentos\ recuperados}$$

$$exhaustividad = \frac{número\ de\ documentos\ relevantes\ recuperados}{número\ de\ documentos\ relevantes\ en\ la\ BD}$$

En la evaluación, la pregunta 1 se utilizó para medir la precisión. En este experimento la precisión indica qué fracción de los fragmentos o enunciados extraídos por el sistema *ResúmeME* son realmente importantes (en términos de consenso de subjetividades). Los jueces evalúan este aspecto seleccionando un inciso; cada inciso equivale a una puntuación que va del 0 al 3, donde 0 puntos significa que ninguno de los fragmentos

contiene ideas relevantes (inciso d) y 3 puntos significa que todos los fragmentos contienen ideas relevantes del documento fuente (inciso a).

La precisión del resumen de cada texto se calcula promediando los puntos asignados por todos los jueces, esto es, sumando los puntos asignados por los jueces y dividiendo esta sumatoria entre el máximo valor posible de puntos:

$$precisión = \frac{puntos\ obtenidos\ pregunta1}{máximo\ posible\ de\ puntos}$$

Se puede observar que si todos los jueces califican al resumen con la máxima puntuación, es decir, si todos seleccionan el inciso a), el cociente será igual a 1. Una precisión con valor de 1 indica que todos los enunciados extraídos por el sistema fueron considerados importantes; en el otro extremo, una precisión de valor 0 indica que ninguno de los enunciados extraídos poseía contenido relevante de acuerdo con los jueces.

La pregunta 2 se utilizó para medir la exhaustividad. En este experimento la exhaustividad índica qué fracción del conjunto de todos los enunciados importantes del texto fue extraída por el sistema. Como antes, los jueces evalúan este aspecto seleccionando un inciso, cada inciso equivale a una puntuación que va del 0 al 3, donde 0 puntos significa que ninguno de los enunciados importantes del texto está presente en el resumen (inciso a) y 3 puntos significa que todos los enunciados importantes del texto fueron extraídos para formar el resumen (inciso d).

La exhaustividad del resumen de cada texto se calcula promediando los puntos asignados por cada juez, esto es, sumando los puntos asignados por todos los jueces y dividiendo esta sumatoria entre el máximo valor posible de puntos.

$$exhaustividad = \frac{puntos\ obtenidos\ en\ pregunta2}{máximo\ posible\ de\ puntos}$$

Se puede observar que si todos los jueces califican al resumen con la máxima puntuación, es decir, si todos seleccionan el inciso d), el cociente será igual a 1. Una exhaustividad con valor de 1 indica que el sistema fue capaz, según el juez, de extraer todos los fragmentos

importantes del texto. En el otro extremo, una exhaustividad de valor 0 indica que el sistema no fue capaz de extraer ninguno de los enunciados relevantes del documento fuente, según dicho juez.

Adicionalmente se calculó la medida F o *f-score* [24], también usada en el área de recuperación de la información, que combina ambos valores de precisión y exhaustividad mediante la siguiente fórmula[11]:

$$F = \frac{2 * \text{precisión} * \text{exhaustividad}}{\text{precisión} + \text{exhaustividad}}$$

A continuación se muestran los resultados globales para cada lengua. Estos valores se obtuvieron promediando los respectivos valores de precisión, exhaustividad y F de los textos de cada lengua[12]:

Tabla 2: Resultados de la evaluación para cada lengua

Lengua	Precisión	Exhaustividad	Medida F
Español	0.606061	0.517483	0.55828
Inglés	0.746667	0.653333	0.696889
Alemán	0.833333	0.833333	0.833333
Chuj	1.0	1.0	1.0
tarahumara	1.0	0.666666	0.79999

[11] La medida F contempla un factor β para pesar la importancia relativa de la precisión y la exhaustividad en la fórmula. La fórmula que se utilizó es la simplificación obtenida con β=1 en donde las medidas de precisión y exhaustividad se pesan igual.

[12] El detalle de los valores obtenidos de precisión y exhaustividad para cada texto de cada lengua se muestra en el apéndice C.

5.3 ANÁLISIS DE RESULTADOS

El experimento de evaluación diseñado para *ResúmeME* utiliza métodos sencillos y está lejos de ser el esquema perfecto de evaluación. De hecho, es un lujo tener tantos jueces entusiastas evaluando los extractos, lo que en otros contextos sería prohibitivo. Sin embargo, ofrece información sobre el desempeño del sistema que es útil para futuras mejoras.

La evaluación arroja dos valores: la precisión y la exhaustividad. Se puede observar que en todos los casos la precisión resultó mayor que la exhaustividad. Aunque en el experimento el valor de F iguala la importancia de ambas medidas, en realidad conviene enfatizar la importancia de una precisión alta más que la de una exhaustividad alta, porque es preferible incluir en el resumen información pertinente, que le permita al lector decidir si necesita leer todo el documento, que incluir toda la información posible.

La exhaustividad representa la capacidad del sistema de extraer la totalidad de fragmentos relevantes del documento fuente. Por lo tanto, el valor de exhaustividad puede verse afectado por la compresión del resumen; por ejemplo, un resumen con un radio de compresión muy bajo difícilmente contendrá todos los enunciados relevantes. En este sentido, sería necesario evaluar los resúmenes a diferentes niveles de compresión para tener una interpretación más completa del valor de exhaustividad.

Por su parte, los valores obtenidos de precisión fueron favorables. En los resultados globales de todas las lenguas (Tabla 2) la precisión fue mayor a 0.6 lo que se traduce en que más del 60% de los enunciados que extrae *ResúmeME* contiene ideas relevantes del texto fuente, según el consenso de los jueces.

Es importante mencionar que, desde un inicio, *ResúmeME* se planteó como un sistema independiente del idioma. Por esta razón el experimento se extendió a diferentes lenguas. Sin embargo, las lenguas tarahumara y chuj son casos especiales pues, por su carácter de idiomas indígenas de México, fue difícil obtener documentos y jueces, por lo que únicamente se utilizaron un texto y un juez para cada una de ellas, es decir, constituyen

experimentos de valor simbólico, con muestras muy pequeñas (incluso también para el alemán), por lo que sería manifiestamente apresurado adelantar una conclusión sobre el desempeño del *ResúmeME* con todas estas lenguas.

Para el español e inglés, en cambio, hubo más jueces y textos, lo que permitió realizar un mejor análisis. Como se ve en la Tabla 2, el inglés obtuvo resultados superiores al español. Una posible explicación es que el español es una lengua morfológicamente más compleja que el inglés, sobre todo en sus morfologías flexivas; por lo que incluir una etapa de *stemming*[13] *(*truncamiento morfológico de palabras*)* o lematización en el sistema podría derivar en un mejor desempeño con textos en español y en general con lenguas más ricas morfológicamente que el inglés.

En general, la evaluación del sistema *ResúmeME* arrojó resultados alentadores y se pudo obtener información útil para una optimización posterior del sistema. Sin embargo, el método de evaluación también presenta desventajas, la principal es que no ofrece un marco de comparación con otros sistemas de resumen automático. La evaluación propuesta en este trabajo se diseñó especialmente para ResúmeME y debido a que en un inicio se buscó hacer una evaluación en donde se prescindiera de resúmenes "ideales" no se utilizaron sistemas como el ROUGE.

Sin duda es una desventaja no conocer cómo se desempeña *ResúmeME* en comparación con otros sistemas de resumen automático, en un futuro sería favorable extender la evaluación, ya sea utilizando alguno de los sistemas ampliamente conocidos y aceptados a nivel mundial o aplicando la evaluación propuesta en esta tesis a otros sistemas de resumen automático para comparar los resultados. Esto último se hubiera podido hacer generando resúmenes de los mismos textos tanto con *ResúmeME* como con otros sistemas extractivos, sin embargo esta tarea hubiera implicado que, por lo menos, se duplicara la cantidad de resúmenes que cada juez hubiera tenido que leer y juzgar. Este

[13] Al igual que la lematización el *stemming* se usa para agrupar a las palabras en una forma estándar. Sin embargo en el stemming se hace por medio del trucamiento. Por ejemplo, las palabras "correr", "corrió", "corriendo" y "corres" se agruparían como "corr".

factor fue prohibitivo debido al tiempo con el que disponían los voluntarios que fungieron como jueces.

Finalmente, al ser esta una evaluación manual tiene la gran limitante de que es un experimento costoso, en el sentido de que se necesita una cantidad considerable de jueces humanos y textos para tener una muestra representativa de juicios que permita conocer plenamente el desempeño del sistema. El hecho de que sea un experimento costoso también dificulta que se extienda fácilmente, por ejemplo, evaluar los resúmenes a diferentes niveles de compresión o evaluar resúmenes generados por otros sistemas, etc. A pesar de esto, las evaluaciones manuales tienen una ventaja y esa es que los jueces humanos toman en cuenta información de tipo sintáctico o semántico que en una evaluación automática generalmente pasa desapercibida.

6. CONCLUSIONES

En este último capítulo se presentan las conclusiones de la tesis, que esencialmente incluyen una valoración final del sistema desarrollado, sus alcances, limitaciones, ventajas y áreas para trabajo futuro.

En el presente trabajo se desarrolló un sistema de resumen automático de tipo extractivo. Este tipo de sistemas tiene como objetivo la extracción de enunciados que contengan información relacionada con las ideas relevantes de un texto para así obtener una versión condensada del mismo.

Como se pudo ver a lo largo de la tesis, la extracción o selección de los enunciados relevantes está determinada por esquemas de asignación de pesos que cuantifican la importancia de cada palabra dentro del texto para posteriormente calcular la importancia de un enunciado basándose en los pesos de las palabras que lo forman.

Los criterios en los que se basan los esquemas de asignación de pesos pueden ser muy variados y repercuten de manera importante en los resultados de un resumidor. Para el sistema *ResúmeME* se eligieron dos esquemas estadísticos, uno ampliamente conocido en los ámbitos de recuperación de información (TF-IDF) y otro propuesto en esta tesis (S-I).

La evaluación que se mostró en el capítulo IV fue un esfuerzo por medir el desempeño y efectividad de las técnicas utilizadas para generar resúmenes. Sin embargo, más allá de la elección de esquemas efectivos, existen otros retos que resolver para obtener un buen resumen automático. Uno de ellos es la adecuada segmentación de enunciados y palabras; al ser éste un sistema que extrae enunciados y que utiliza esquemas de asignación de pesos, es de gran importancia poder delimitar de manera precisa los enunciados y las palabras.

La *tokenización* es un problema complejo que ya se discutió previamente en la tesis. En el sistema *ResúmeME* la tokenización de palabras se limitó a separar las palabras gráficas

mientras que la segmentación de enunciados se realizó con un algoritmo que busca identificar abreviaturas para delimitar correctamente los enunciados.

Aunque se observó que la segmentación de palabras y enunciados de los textos utilizados para la evaluación fue favorable en la mayoría de los casos, en el futuro se podrían realizar mejoras al sistema, incluyendo métodos que realicen una mejor delimitación y un análisis más exhaustivo; por ejemplo, en el caso de las palabras se podrían elaborar algoritmos capaces de identificar no sólo las palabras individuales sino también las colocaciones. En el caso de los enunciados, con el fin de detectar ambigüedades y delimitar las fronteras, se podrían combinar tanto métodos heurísticos como estadísticos, incluso algoritmos de aprendizaje que involucren entrenar redes neuronales con textos pre-etiquetados en donde se marque el inicio y final de los enunciados. En todos los casos es importante recordar que se debe procurar que los métodos mantengan independencia de la lengua.

Otro aspecto del pre-procesamiento que es importante considerar para la optimización del sistema es incluir una etapa de stemming o lematización, que como ya se ha mencionado antes es el proceso de convertir los tokens a una forma estándar. Añadir una etapa de stemming reduciría el número de palabras distintas o tipos dentro del texto. Sería interesante evaluar en un futuro qué beneficios significativos proporciona incluir una etapa de stemming en el sistema. En lo que respecta a esta tesis, la evaluación realizada en el capítulo IV arroja indicios de que podrían beneficiarse los resúmenes de textos escritos en lenguas morfológicamente más complejas que el inglés, como el español.

En el diseño del sistema no se optó por una etapa de stemming porque muchos de los métodos más populares son dependientes del lenguaje, es decir, se basan en heurísticas o diccionarios específicos para cada lengua (utilizando el conocido algoritmo de Porter [9],[10],[13]). Se tenía especial interés en métodos que fueran independientes de la lengua y que además utilizaran mediciones estadísticas para agrupar las variantes morfológicas de una palabra. En este sentido conviene mencionar que en un inicio sí se

utilizó un método de stemming estadístico[14]. Sin embargo, su implementación en *ResúmeME* aumentó considerablemente el tiempo de procesamiento para generar los resúmenes por lo que finalmente se decidió excluirlo.

Otro de los grandes retos a resolver en un sistema de resumen automático es la coherencia o cohesión textual. Los resúmenes formados por fragmentos textuales del documento fuente, como es el caso de los resúmenes que genera *ResúmeME*, pueden tener el inconveniente de contener referencias pronominales a sustantivos que no se encuentran en el texto seleccionado (anáforas y catáforas) provocando que algunas partes del resumen parezcan incoherentes o den la impresión de "texto cortado". Las soluciones para lograr una mejor conexión de los fragmentos tienen que ver con la identificación y resolución de anáforas que es una de las metas del PLN y que se irá perfeccionando conforme se profundice en su estudio.

Además de lo antes mencionado, un aspecto que se debe mejorar es la evaluación del sistema. Como se comentó en el capítulo anterior, la principal debilidad de la evaluación presentada en esta tesis es que no ofrece un marco de comparación con otros sistemas de resumen automático. Así que queda pendiente como trabajo futuro ampliar la evaluación propuesta o adoptar alguno de los sistemas de evaluación más conocidos y aceptados a nivel mundial.

Como se dijo al inicio, en el planteamiento del problema, el sistema de resumen propuesto en esta tesis no pretende reemplazar por completo a los agentes humanos. De hecho, una vez concluido el trabajo y habiendo analizado los resultados, son evidentes las limitaciones de *ResúmeME* (muchas de ellas discutidas previamente). A pesar de esto, *ResúmeME* será de gran utilidad para generar los resúmenes del número creciente de documentos de los corpus que se desarrollan en el Instituto de Ingeniería (Corpus de las Sexualidades en México CONACyT 105711, Corpus Histórico del Español en México DGAPA PAPIIT 400905 y 402008, Corpus Lingüístico en Ingeniería CONACyT R3774-A y los que se

[14] Método expuesto por el Dr. Juan Manuel Torres Moreno de la Universidad de Avignon durante su estancia en el Grupo de Ingeniería Lingüística, 2009.

compilen en un futuro) porque puede brindar versiones condensadas de los documentos que sirvan como una vista previa o una visión general de los temas abordados y así ayudar al usuario o a un sistema a tomar decisiones sin que necesariamente tenga que leer/procesar todo el documento.

Entre las ventajas que posee el sistema se encuentran su independencia de la fuente y del lenguaje, esto es, que no está diseñado para textos con una estructura o dominio en particular ni para una lengua específica. Y aunque existen numerosos aspectos que optimizar, la evaluación sugiere que algoritmos superficiales y relativamente sencillos son capaces de capturar información relevante contenida en el documento fuente. En este sentido, se puede continuar el trabajo del sistema *ResúmeME* añadiendo más esquemas de asignación de pesos y explorando más a fondo de qué manera influyen aspectos como la distribución de las palabras o su contenido de información, entre otros, para determinar su relevancia dentro del texto.

Finalmente, si se considera que para obtener resúmenes legibles y apropiados un sistema de resumen automático debe [25]: elaborar o representar (‘entender’) el contenido del texto a un nivel lo suficientemente profundo, ser capaz de determinar la importancia del material y generar una salida coherente. Cualquiera de estas tareas son metas que también persigue el PLN y que en muchos casos aún pertenecen al campo de la investigación. En vista de lo anterior, los métodos estadísticos representan una buena alternativa pues aunque su tratamiento del texto sea superficial y no partan de representaciones más elaboradas, son lo suficientemente robustos para obtener resultados deseables.

APÉNDICES

APÉNDICE A. INTERFAZ WEB DEL SISTEMA *RESÚMEME*

APÉNDICE B. TEXTOS PARA LA EVALUACIÓN

(Únicamente se presentan algunos de los textos utilizados para las evaluaciones en inglés y español y los textos utilizados para el alemán, chuj y tarahumara)

Texto 2. (español)

Resultados obtenidos en la evaluación. Precisión: 0.846154 exhaustividad: 0.641026

José Luis Cuevas

José Luis Cuevas (1934-), dibujante, grabador, ilustrador, escritor y conferencista mexicano de formación esencialmente autodidacta, nacido en la ciudad de México. Considerado uno de los grandes representantes del neofiguративismo, forma parte también de la corriente interiorista, a favor de un arte más inclinado hacia la problemática existencial, con una visión neohumanista.

La figura más extrovertida de la Generación de la Ruptura, cobró notoriedad gracias a un manifiesto conocido como La cortina del nopal (publicado en el suplemento México en la Cultura del periódico Novedades entre 1958 y 1959). En esta serie de artículos dio a conocer su ideario estético, que participa de la búsqueda de la libertad de expresión tanto formal como temática entre los artistas jóvenes que querían alejarse de la escuela mexicana de pintura, particularmente del muralismo y su ya entonces anacrónico mensaje de contenido político-social. Esta postura anticonformista culminó con la realización de un Mural efímero (término contradictorio con el que quiso satirizar las pretensiones de continuidad de ese movimiento). El acto tuvo lugar en una concurrida esquina de la Zona Rosa, conocido barrio de la capital mexicana, que en sus inicios fuera punto de reunión de intelectuales y artistas, bautizado por él con ese nombre. Dicho happening (1967) reveló sus enormes dotes autopublicitarias pues provocó gran expectación entre los asistentes y fue ampliamente comentado por la prensa nacional y extranjera.

Su obra en su primera etapa es expresionista por influencia de José Clemente Orozco así como por una ferocidad gestual que nos recuerda la estatuaria prehispánica (particularmente mexica). Hacia la década de 1960 se revela como uno de los más prestigiosos litógrafos contemporáneos gracias a las series realizadas en Estados Unidos: Recollections of Childhood (1962) y Cuevas Charenton (1964), realizadas ambas en Los Ángeles; Crime by Cuevas (1968), en Nueva York; Homage to Quevedo (1969), en San Francisco. La literatura ha penetrado su obra desde que ilustró Los mundos de Kafka y Cuevas (Filadelfia, Falcon Press, 1957), libro considerado como una joya y una valiosa aportación a la bibliofilia contemporánea. En la década de 1970 consolida su presencia como una de las personalidades más originales y polémicas de la cultura contemporánea mexicana. No obstante, al sentirse incomprendido en su país, se autoexilia en Francia, donde afianza su prestigio con la gran retrospectiva de dibujo que le dedica el Museo de Arte Moderno de París (1976).

Una serie de temas (como el recuerdo de sus ancestros catalanes por vía materna, una quema de brujas en la España del siglo XVII, así como sus correrías por los barrios gótico y chino de Barcelona y el de Malasaña en Madrid) le sirvieron de inspiración para las suites gráficas Catalana (1981), Vasca, llamada también Intolerancia, (1983) y Madrileña (1987). En ellas, su afición por lo grotesco cobra forma mediante el tenebrismo y en especial lo esperpéntico, tan propios de la tradición española. La década de 1990 presenció un notable incremento de su actividad escultórica, con bronces de diversos tamaños, entre los que

sobresale La giganta (1991), espectacular obra que ejemplifica la dualidad masculino-femenina, realizada especialmente para el patio del Museo José Luis Cuevas (1992), en el que se aloja la colección particular de arte latinoamericano contemporáneo y europeo (fototipos de dibujos originales de Rembrandt y dos series gráficas de Pablo Picasso) donada por el artista y su esposa para ser exhibida en una vieja casona conventual del siglo XVII, perteneciente al centro histórico de la ciudad de México. Retomará posteriormente la influencia española para realizar los dibujos de la Suite Andaluza (1993) durante una estadía de varios meses en Sevilla, ciudad que junto a París y la yucateca Mérida (tierra natal de su madre) figura entre sus preferidas.

José Luis Cuevas se da a conocer internacionalmente muy joven con exposiciones en la ciudad de Washington (1954), en París (1955) y en Nueva York (1957). Inicia así una amplísima trayectoria que comprende hasta ahora cientos de exposiciones individuales y colectivas en galerías, museos y ferias de arte de las principales ciudades del mundo.

Ha sido merecedor de importantes galardones y distinciones, como el primer premio internacional de dibujo de la V Bienal de São Paulo (1959) y el Premio Nacional de Ciencias y Artes de México (1981). En 1991 fue nombrado Caballero de las Artes y de las Letras por el Ministerio de Cultura francés. En 1993 ingresó en el Sistema Nacional de Creadores como Creador Emérito. En 1997 recibió en España de manos de la reina doña Sofía el premio Tomás Francisco Prieto durante el acto inaugural de la muestra retrospectiva gráfica que le dedicó la Casa de la Moneda de Madrid. Con ocasión de su nombramiento en 1997 como Artista de la Ciudad por el Gobierno del Distrito Federal, se realizaron diversas actividades en torno a su figura, que culminaron con la inauguración en su museo de la exposición José Luis Cuevas rebasa su tiempo. A principios de 1998 el Museo Nacional Centro de Arte Reina Sofía le dedicó una gran retrospectiva de dibujo a la que el artista agregó la serie de pequeñas esculturas llamada Animales impuros inspirada en el poemario del mismo nombre, original del escritor vanguardista español José Miguel Ullán.

Vigoroso polemista, ha dado numerosas conferencias en las que vierte su vasta cultura plástica, literaria y cinematográfica. Cultiva el periodismo autobiográfico en el suplemento dominical El Búho, del Excélsior, con la columna Cuevario (1985-1998), que a partir de 1999 aparece en el periódico El Universal. Es autor de varios libros autobiográficos, como el titulado Gato macho (1994).

Mediante su inconfundible y devastadora línea, este gran dibujante desnuda no sólo las almas, sino también los cuerpos de sus personajes, reflejando en ellos las constantes que han marcado su estilo: enfermedad, vejez y muerte por lo que toca a la descomposición carnal; otra podredumbre (de índole moral) la retrata recurriendo a la prostitución y al despotismo. Para entender toda esta magnificación de la degradación humana habría que remitirse a la infancia del artista, que transcurre en un entorno de seres marginados, habitantes del viejo centro de su ciudad natal; personajes que buscará reiteradamente en su deambular por los barrios de Madrid, Tánger, Nueva York, Hamburgo y tantas otras urbes. Cuevas es un relator gráfico de la soledad y angustia que acompañan al hombre en los grandes conglomerados urbanos. El hilo conductor de su obra, que nos permite apreciarla en todo su significado, es la carne: la de hospital, de morgue, de burdel, pero también la del goce íntimo que el erotismo introduce en su obra como una afirmación de vida.

Cuevas otorga al autorretrato un lugar preeminente, al convertir su práctica en un ritual cotidiano en el que subyace una profunda meditación sobre el paso del tiempo y la muerte.

Fragmentos extraídos por *ResúmeME*:

- José Luis Cuevas
  José Luis Cuevas (1934-), dibujante, grabador, ilustrador, escritor y conferencista mexicano de formación esencialmente autodidacta, nacido en la ciudad de México.
  puntaje: 0.61679466009

- Hacia la década de 1960 se revela como uno de los más prestigiosos litógrafos contemporáneos gracias a las series realizadas en Estados Unidos: Recollections of Childhood (1962) y Cuevas Charenton (1964), realizadas ambas en Los Ángeles; Crime by Cuevas (1968), en Nueva York; Homage to Quevedo (1969), en San Francisco.
  puntaje:0.629754271923

- La década de 1990 presenció un notable incremento de su actividad escultórica, con bronces de diversos tamaños, entre los que sobresale La giganta (1991), espectacular obra que ejemplifica la dualidad masculino-femenina, realizada especialmente para el patio del Museo José Luis Cuevas (1992), en el que se aloja la colección particular de arte latinoamericano contemporáneo y europeo (fototipos de dibujos originales de Rembrandt y dos series gráficas de Pablo Picasso) donada por el artista y su esposa para ser exhibida en una vieja casona conventual del siglo XVII, perteneciente al centro histórico de la ciudad de México.
  puntaje:0.692847173433

- Mediante su inconfundible y devastadora línea, este gran dibujante desnuda no sólo las almas, sino también los cuerpos de sus personajes, reflejando en ellos las constantes que han marcado su estilo: enfermedad, vejez y muerte por lo que toca a la descomposición carnal; otra podredumbre (de índole moral) la retrata recurriendo a la prostitución y al despotismo.
  puntaje:0.664886205026

Texto 3. (español)

Resultados obtenidos en la evaluación. Precisión: 0.461538 Exhaustividad: 0.435897

Desde tiempos remotos, las manchas solares causaban estragos en todo tipo de sistema electrónico. El avance de la tecnología había logrado que en los últimos sesenta años no hubieran ocurrido catástrofes como esa. De todos modos, nada evitó que ese día, durante la plenaria de aquel congreso, la ciudad entera se quedara sin energía y, lo que es peor, tantos aparatos se dañaran irremediablemente, precisamente cuando Lucía iniciaba la lectura de su ponencia en ese congreso internacional de su especialidad, que se llevaba a cabo en aquel gigantesco centro de convenciones, de aquella calurosa ciudad junto al mar. Poco antes de empezar a leer su informe se acomodó el pelo largo y rubio tras su delgado torso y miró a la audiencia con sus ojos tristes y azules. Desde el estrado pudo ver las caras morenas y apuestas de los asistentes, sus cuerpos rechonchos y sus ojos oscuros y amables. Tal vez al presentir lo que estaba por ocurrir, recordó en una fracción de segundo sus años de adolescencia. Le vino a la memoria el acoso sexual que a menudo sufría por sus características físicas. Todos la deseaban, por considerarla un ejemplo extremo del máximo paradigma de belleza reinante.

En realidad, su angustia por la apariencia física tenía muchos motivos. El principal no era su temperamento siempre tímido, sino las maravillas tecnológicas de su época. Todo empezó en los tiempos de la comercialización a gran escala de la cosmetología genética. Primero se hizo posible cambiar el color de la piel y del pelo con sólo tomar una pastilla, luego se hizo posible modificar las narices aguileñas a narices de escultura griega. Finalmente, se modificaron las complexiones gruesas a físicos delgadísimos sin importar la cantidad en el consumo de alimentos. En pocos años la población del mundo modificó su aspecto hasta conseguir que la mayoría se pareciera a los modelos de los medios de comunicación de la época. Pero a largo plazo, esos cambios causaron problemas de salud insospechados, reacciones secundarias que por ambición e imprudencia nadie había querido reconocer, males crónicos que se volvieron enfermedades universales. Lo peor es que más modificaciones genéticas tenían resultados todavía más inesperados, así que dejaron de aplicarse mientras no se mejoraran estos procedimientos cosméticos.

En esa época se dio el cambio más abrupto que jamás se dio en los patrones de belleza que enarbolan los medios de comunicación mundial. De repente, las mujeres más bellas de los medios dejaron de ser delgadas y rubias, los hombres más atractivos ya no eran altos y fornidos. Ya nadie quería reproducirse con personas con esas características porque esos rasgos ahora eran huellas de alteraciones genéticas que denunciaban peligros insondables. La probabilidad de que alguien tuviera esos rasgos de manera natural se hizo bajísima. Y como casi todos las tenían, habían entrado a la reserva de genes de la humanidad y la única manera de sacarlos parecía ser la selección natural. Por eso los hombres regordetes y calvos se convirtieron en preciados símbolos sexuales y las mujeres celulíticas de pelo hirsuto se hicieron los ejemplos de belleza más atractivos.

Los hologramadores surgieron algunos años después. Se comercializaron rápidamente porque permitían a todos aparentar lo que en la época era considerado belleza. Al principio, sólo los ricos podían adquirir esos aparatos, pero pronto también las clases medias pudieron gozar de sus beneficios. El mundo se llenó de personas otrora consideradas feas, pero sólo en apariencia. Apretar un botón del cinturón hologramador bastaba para esconder las realidades pálidas y blondas tras una saludable apariencia flácida y prieta.

Por eso cundió la alarma en aquel salón del centro de convenciones cuando las perturbaciones solares dañaron los finos hologramadores de sus asistentes. Desde el podio, Lucía podía ver los rostros

desencajados de los congresistas. Si antes lucían morenos y rechonchos, ahora exhibían pieles pálidas y complexiones desencarnadas. Lo peor eran los intensos ojos azules de la mayoría. Por si fuera poco, aquella ciudad con playas era tan calurosa y el clima artificial del centro de convenciones tan agradable que la mayoría había prescindido de ponerse ropas verdaderas. Por eso pudo ver los desagradables cuerpos delgados, musculosos y blancos de la audiencia. Lo peor de todo para Lucía era que vieran su verdadero yo, ese que había ocultado celosamente desde que terminara la adolescencia. Todos la miraban sorprendidos. A nadie se le había ocurrido imaginar que también ella, la brillante y famosa eminencia, utilizara un hologramador. Que detrás de esa apariencia de rubia hambrienta, de piel de seda nívea, evidente disfraz de quien quiere ser juzgado por sus capacidades intelectuales y no por su apariencia física, se ocultara una mujer sublime, de piel oscura y brillante, de cara de india del antiguo Nuevo Mundo, de silueta amplia y pubis negro como zapote, de senos jugosos y colgantes, como verdadera diosa de la fertilidad.

Fragmentos extraídos por *ResúmeME*:

- De todos modos, nada evitó que ese día, durante la plenaria de aquel congreso, la ciudad entera se quedara sin energía y, lo que es peor, tantos aparatos se dañaran irremediablemente, precisamente cuando Lucía iniciaba la lectura de su ponencia en ese congreso internacional de su especialidad, que se llevaba a cabo en aquel gigantesco centro de convenciones, de aquella calurosa ciudad junto al mar.
  puntaje: 0.855245170546

- Poco antes de empezar a leer su informe se acomodó el pelo largo y rubio tras su delgado torso y miró a la audiencia con sus ojos tristes y azules.
  puntaje: 0. 621645475647

- Desde el estrado pudo ver las caras morenas y apuestas de los asistentes, sus cuerpos rechonchos y sus ojos oscuros y amables.
  puntaje: 0. 637504751902

- Por si fuera poco, aquella ciudad con playas era tan calurosa y el clima artificial del centro de convenciones tan agradable que la mayoría había prescindido de ponerse ropas verdaderas.
  puntaje: 0. 633736597446

- Lo peor de todo para Lucía era que vieran su verdadero yo, ese que había ocultado celosamente desde que terminara la adolescencia.
  puntaje: 0.623348823792

- Que detrás de esa apariencia de rubia hambrienta, de piel de seda nívea, evidente disfraz de quien quiere ser juzgado por sus capacidades intelectuales y no por su apariencia física, se ocultara una mujer sublime, de piel oscura y brillante, de cara de india del antiguo Nuevo Mundo, de silueta amplia y pubis negro como zapote, de senos jugosos y colgantes, como verdadera diosa de la fertilidad.
  puntaje: 0.734304502463

Texto 4. (español)

Resultados obtenidos en la evaluación: Precisión: 0.43589 Exhaustividad: 0.43589

Polémica sobre el canibalismo del Diccionario de la evolución.
De Richard Milner.

Por más sorprendente que parezca, los descubridores de casi cualquier criatura antropoide fósil se han apresurado siempre a anunciar el hallazgo de pruebas concomitantes de «canibalismo». Luego, en la mayoría de los casos, los colegas del descubridor declaran insuficiente la demostración, que queda así fuera de debate.

Los simios antropomorfos africanos (australopitecinos), el hombre de Pekín (Homo erectus), el del Neandertal y el del Cromañón fueron considerados por quienes los descubrieron aficionados a la carne de sus prójimos. Se ha discutido durante años si nuestros antepasados se comían unos a otros o si la espeluznante interpretación habitual arroja más luz sobre la mente de los antropólogos que sobre el canibalismo prehistórico.

Robert Broom y Raymond Dart, los paleontólogos surafricanos descubridores de muchos fósiles de australopitecos, pensaban que los huesos magullados y los cráneos perforados demostraban que descendemos de un simio predador que no se detenía ante los miembros de su propia especie.

Algunos años más tarde, el profesor Franz Weidenreich colaboró en la excavación de los restos del Homo erectus (el hombre de Pekín) en una cueva en China y observó que muchos cráneos estaban magullados por la base. Concluyó que aquella gente se comía el cerebro de sus compañeros; pero, luego, cambió de idea. En diversos momentos, otros expertos han pintado con el mismo color negro tanto al hombre del Neandertal como al primitivo Homo sapiens.

Los rasguños que se aprecian en los huesos fósiles de homínidos pueden tener otras causas distintas de las del llamado canibalismo gourmet; los huesos podrían haber sido roídos y raspados por hienas u otros animales carroñeros. El Homo erectus fue quizá la presa y no el verdugo. En el yacimiento del hombre de Pekín sólo se encontraron cráneos, lo cual podría significar que las cabezas fueron transportadas a la gruta para la celebración de algún rito. (Muchos pueblos tribales contemporáneos utilizan los cráneos de los parientes muertos en el culto a los antepasados.)

Los antropólogos siguen debatiendo todavía la naturaleza y difusión del canibalismo entre pueblos tribales contemporáneos. El profesor W. Arens provocó un escándalo entre los expertos al hacer una crítica general de la idea en su libro The Man-Eating Mith (El mito de la antropofagia) (1979). Como muchos antropólogos, Arens había dado siempre por supuesto que los exploradores del siglo XIX habían visitado tribus caníbales en Africa, Nueva Guinea y Suramérica. Pero, cuando cribó la masiva bibliografía sobre el tema, no pudo encontrar un relato satisfactorio de primera mano sobre la práctica del canibalismo como costumbre socialmente aprobada en alguna parte del mundo.

Cuando Arens marchó a Africa para recoger información sobre el canibalismo tribal se llevó la sorpresa de su vida. Los aldeanos azanda, objeto de sus estudios, habían decidido que él era un «chupasangres», una

especie de vampiro. Aunque llevaba viviendo allí año y medio, el profesor Arens nunca consiguió convencerlos de que no se alimentaba en secreto de sangre humana durante la noche.

Mientras los colonizadores europeos daban pábulo a su miedo a los caníbales africanos, nunca advirtieron que los africanos albergaban las mismas sospechas sobre ellos. Además, los africanos disponían de pruebas. Algunos años antes, durante una guerra, ciertos europeos habían intentado persuadir a los nativos de que donaran sangre para sus soldados heridos. Los campesinos temían todavía ser llamados al hospital, donde se les desangraría. Su recuerdo de las urgentes peticiones de sangre se habían convertido en la convicción de que los europeos necesitaban beber sangre africana para mantenerse vivos.

Al principio, Arens contempló esta creencia con actitud de superioridad, pero más tarde se disgustó consigo mismo por no haber captado la metáfora política subyacente. Los africanos, como es natural, consideraban perfectamente razonable sentir que los europeos les estaban robando la vitalidad y consumiéndoles la sangre que les daba vida.

Arens constató que a lo largo de toda la historia se han lanzado acusaciones de canibalismo con el fin de aunar al grupo acusador como pueblo dotado de eticidad y situar al grupo enemigo al margen de los sentimientos humanos. Los colonialistas europeos justificaron desde principios del siglo XVII el sometimiento de los pueblos tribales basándose en que se trataba de caníbales sin civilizar. Los coreanos pensaban que los chinos eran caníbales y los chinos creían lo mismo de los coreanos. Arens comenzó a sospechar que las acusaciones y creencias en torno al canibalismo están mucho más extendidas que la práctica real de la antropofagia.

Arens concluía que nunca ha habido referencias fidedignas de prácticas extendidas de canibalismo gourmet. Según él, se trataba de un mito de los antropólogos; así pues, desafió a sus colegas a que demostraran lo contrario. Al poco tiempo de la aparición de su libro, varios investigadores de campo se prestaron a presentar sus pruebas.

George Morren, de la Universidad de Rutgers, realizó en los últimos años de la década de 1960 trabajos de campo en Nueva Guinea, donde los ancianos de la tribu de los miyanmin que habían participado en actividades caníbales le ofrecieron informaciones detalladas. Morren confrontó sus complejas descripciones con varios informantes y estudió asimismo las actas de los tribunales de un juicio celebrado en 1959 contra más de treinta miyanmin acusados de asesinar y comerse a 16 personas de una tribu vecina.

Cuando Morren instó de manera particular a los miyanmin para que explicaran el posible significado religioso o simbólico del incidente, ellos insistieron en que no lo había. «No; simplemente buscábamos carne.» Se trata de un relato de canibalismo culturalmente sancionado, de la máxima autenticidad y documentación posibles.

Entretanto, las razones de Arens no han persuadido a otros antropólogos para abandonar la mayoría de la bibliografía «caníbal» por considerarla sesgada y de segunda mano. Por otra parte, un conjunto de datos cada vez más abundante, procedente de escritos y obras de arte recientemente descifradas, muestra que los antiguos mayas y aztecas practicaban sacrificios cruentos y ritos caníbales a gran escala. El debate sobre un posible pasado caníbal de la especie humana sigue su curso.

Los observadores del comportamiento de los primates han contribuido asimismo a esta fascinante polémica. Tras pasar una década observando chimpancés en las selvas del Zaire, Jane Goodall había llegado a la conclusión de que eran vegetarianos amables, pacíficos, sociales y a veces bufonescos. Pero entretanto, los ha visto cazar y matar a otros animales para conseguir carne, asesinar deliberadamente crías de chimpancés de grupos vecinos y hasta matar y comerse bebés de su propia comunidad.

En cierta ocasión, dos chimpancés, un equipo formado por madre e hija, iniciaron repentinamente una serie de infanticidios caníbales. Una distraía a alguna madre reciente, mientas la otra se llevaba la cría; luego, ambas la mataban y se la comían. Jane Goodall sintió tristeza y desilusión. Admitió haber pensado que los chimpancés eran «mejores» que los seres humanos, pero ahora se daba cuenta de que el corazón de un chimpancé también esconde oscuros secretos.

Fragmentos extraídos por *ResúmeME*:

- Se ha discutido durante años si nuestros antepasados se comían unos a otros o si la espeluznante interpretación habitual arroja más luz sobre la mente de los antropólogos que sobre el canibalismo prehistórico.
  puntaje:0.850955145863

- Los rasguños que se aprecian en los huesos fósiles de homínidos pueden tener otras causas distintas de las del llamado canibalismo gourmet; los huesos podrían haber sido roídos y raspados por hienas u otros animales carroñeros.
  puntaje:0.733005263404

- Pero, cuando cribó la masiva bibliografía sobre el tema, no pudo encontrar un relato satisfactorio de primera mano sobre la práctica del canibalismo como costumbre socialmente aprobada en alguna parte del mundo.
  puntaje:0.871960071764

- Al principio, Arens contempló esta creencia con actitud de superioridad, pero más tarde se disgustó consigo mismo por no haber captado la metáfora política subyacente.
  puntaje: 0.760793024649

- Por otra parte, un conjunto de datos cada vez más abundante, procedente de escritos y obras de arte recientemente descifradas, muestra que los antiguos mayas y aztecas practicaban sacrificios cruentos y ritos caníbales a gran escala
  puntaje: 0.705002103515

- Pero entretanto, los ha visto cazar y matar a otros animales para conseguir carne, asesinar deliberadamente crías de chimpancés de grupos vecinos y hasta matar y comerse bebés de su propia comunidad.
  puntaje: 0.828748394056

- Una distraía a alguna madre reciente, mientas la otra se llevaba la cría; luego, ambas la mataban y se la comían.
  puntaje: 0.746177105836

Texto 7. (español)

Resultados obtenidos en la evaluación. Precisión: 0.28205 Exhaustividad: 0.2564

La Fábrica de Warhol

JAVIER ARANDA LUNA

Una de las grandes lecciones de Andy Warhol fue mostrarnos que lo efímero, como la vida, permanece. Sólo así puedo explicarme que sus obras congreguen multitudes y hayan asombrado a los jóvenes de los años 60, de los 90 y aún a quienes desde hace unos días acuden con curiosidad sostenida al Museo de Arte del Banco de la República de Bogotá a mirar la exposición Andy Warhol Mr. América.

¿Qué hilos tocó, toca la obra de Warhol para mantener su vigencia?

En una sociedad donde la cultura de masas es una realidad y la aldea global algo más que una hipótesis, ¿por qué siguen asombrando sus retratos de Marilyn Monroe, de Jackie Kennedy hechos en serie o sus imágenes de autos chocados, de la silla eléctrica, de las latas de sopa Campbell's o de las botellas de Coca-Cola? ¿Por qué sus desplantes contra el arte serio no han pasado a ser sólo un par de fechas en la historia del arte contemporáneo? ¿Por qué siguen reproduciéndose sus películas de 16 milímetros en blanco y negro sin sonido, donde el propio artista solamente aparece comiendo una hamburguesa o vemos caminar a sus amigos y a quienes tal vez no lo eran frente a una cámara fija que sólo registró fragmentos de un día cualquiera, en esta época en la que los programas Big Brother han fomentado el voyeurismo hi tech?

Las instantáneas de Andy Warhol parecen guardar con sus chillantes colores fragmentos de vida, momentos que son el sueño americano. Pocos artistas han logrado concentrar en algunas imágenes las luces y las sombras de un imperio. Subversivo y cínico Warhol reconstruye con ironía asombrosa la decadencia de un modelo de vida que, curiosamente, es el modelo con el que aún sueñan millones de personas en todo el mundo y con el que soñó el propio artista: la idea del progreso como un continuum inevitable, el consumo sólo por consumir, las estrellas de Hollywood y de la televisión como iconos laicos de la modernidad.

Como Salvador Novo, Warhol amaba la buena vida de la elite y disfrutaba formar parte de su corte. Su arte, nos dice Phillip Larrat-Smith, es una celebración descarada y una afirmación de la elite; un magnífico close up –agrego yo– de las emociones de los hombres de poder.

No es improbable que la crisis financiera global que se inició en Estados Unidos cambie un poco la imagen del trabajo de Warhol. Su crítica y su ironía podrán leerse posiblemente de otra manera. Pero dudo, sin embargo, que el cambio sea sustancial, porque sus imágenes están cargadas de emociones y sus emociones de nuevos referentes entre la juventud. Marilyn Monroe, por lo demás, continuará siendo un icono delstar system; la silla eléctrica un método de corrección final y Warhol un artista de cabellos lo suficientemente grises para pensar en su edad. La Internet y las nuevas herramientas de la computación actualmente permiten no sólo multiplicar masivamente la obra de Warhol –como él hubiera querido–, sino imitarla sin rubor, como ocurre en cualquier sistema industrial que se respete. Si la producción industrial es una tarea colectiva, cualquiera de nosotros puede con un poco de iniciativa continuar el impulso iniciado por Warhol. Los jóvenes de hoy y de mañana, como los de los años 50, seguirán alimentando la Fábrica de ese genio con IQ de 60, como lo definió con tino y sorna Gore Vidal.

Fragmentos extraídos por *ResúmeME*:

- Sólo así puedo explicarme que sus obras congreguen multitudes y hayan asombrado a los jóvenes de los años 60, de los 90 y aún a quienes desde hace unos días acuden con curiosidad sostenida al Museo de Arte del Banco de la República de Bogotá a mirar la exposición Andy Warhol Mr. América.
  ¿Qué hilos tocó, toca la obra de Warhol para mantener su vigencia?
  puntaje: 0.814291473496

- ¿Por qué siguen reproduciéndose sus películas de 16 milímetros en blanco y negro sin sonido, donde el propio artista solamente aparece comiendo una hamburguesa o vemos caminar a sus amigos y a quienes tal vez no lo eran frente a una cámara fija que sólo registró fragmentos de un día cualquiera, en esta época en la que los programas Big Brother han fomentado el voyeurismo hi tech?
  puntaje: 0.710978630397

- Los jóvenes de hoy y de mañana, como los de los años 50, seguirán alimentando la Fábrica de ese genio con IQ de 60, como lo definió con tino y sorna Gore Vidal.
  puntaje: 0.675879176023

Texto 8. (español)

Resultados obtenidos en la evaluación. Precisión: 0.79487 Exhaustividad: 0.53846

Alfonso Caso y la arqueología mexicana

Uno de los pilares indiscutibles de la llamada época dorada de la arqueología mexicana fue el doctor Alfonso Caso y Andrade, ilustre arqueólogo cuya sabiduría, dedicación y ética en el desempeño de sus investigaciones, tanto en el campo como en el laboratorio, dejaron un acervo de primer orden. Entre sus grandes descubrimientos sobresalen la ciudad prehispánica de Monte Albán, con su magnífica Tumba 7, y varios sitios en la Mixteca, como Yucuita, Yucuñidahui y Monte Negro, en Tilantongo. Producto de esos descubrimientos fue una gran cantidad de libros, artículos, reportes, conferencias y literatura popular, que aún son necesarios para el estudio de las culturas mesoamericanas, sobre todo de la zapoteca, la mixteca y la mexica.

Don Alfonso Caso fue especialmente importante en las investigaciones del área cultural de Oaxaca; a partir de 1931, y por más de veinte años, se dedicó al estudio de Monte Albán, sitio al que encontró convertido en terrenos de cultivo, con mogotes llenos de vegetación añosa. Gracias a su laboriosa actuación, en la que recibió la ayuda no sólo de otros arqueólogos sino de muchos técnicos y particularmente de jornaleros que vivían y aún viven alrededor de este majestuoso lugar, pudo descubrir completamente más de veinte de los cientos de edificios y la más monumental de las plazas que configuran los restos de esta enorme ciudad prehispánica. Igualmente importantes son las 176 tumbas que exploró, pues mediante su estudio logró descifrar el sistema de vida de los pueblos zapoteco y mixteco, esto sin contar con los innumerables edificios de otros sitios hacia los que extendió su proyecto central, en el área mixteca y la zona arqueológica de Mitla en el Valle de Oaxaca.

El doctor Caso es considerado el representante de una corriente de pensamiento llamada escuela mexicana de arqueología, que significa el conocimiento de las altas culturas mesoamericanas a través del estudio sistemático de sus diferentes manifestaciones culturales, como son la arqueología, la lingüística, la etnografía, la historia y el estudio de las poblaciones, todas integradas para entender la profundidad de las raíces culturales. Esta escuela creyó en el valor de la reconstrucción de la arquitectura monumental de aquellas culturas, con el objetivo de conocer a profundidad y hacer evidente la historia de nuestros antepasados, especialmente ante los ojos de los jóvenes modernos. Para ello se basó en estudios serios de diferentes expresiones, como la arquitectura de templos, palacios y tumbas, la cerámica, los restos humanos, los libros sagrados, los mapas, los objetos de piedra y otros materiales, a los que Caso llegó a interpretar después de muchos años de estudio.

Una de sus aportaciones más importantes fue el desciframiento del sistema de escritura de las culturas prehispánicas de Oaxaca, llegando a comprender los jeroglíficos que usaron los zapotecos desde el año 500 antes de nuestra era, para nombrar a las personas, para contar el tiempo y para narrar sus conquistas, en complicados textos tallados en grandes piedras. Tiempo después, hacia el año 600 de nuestra era, con ese sistema de escritura contaban sobre todo sus violentas incursiones en los pueblos, sacrificando a algunos y tomando cautivos a sus dirigentes, todo ello para asegurar la supremacía del pueblo zapoteco, cuya capital era Monte Albán.

Asimismo, interpretó el sistema de escritura mixteca, cuyos pueblos plasmaron en libros hechos con piel de venado y pintados con colores brillantes, para narrar los mitos sobre sus orígenes, su procedencia de la tierra y de las nubes, de los árboles y de las rocas, y complicadas biografías –entre reales y míticas– de los personajes importantes, como sacerdotes, gobernantes y guerreros de esos pueblos. Uno de los primeros textos en descifrarse fue el Mapa de Teozacoalco, a partir del cual el doctor Caso logró establecer correlaciones entre el calendario antiguo y el de uso cotidiano de nuestra cultura, además le permitió ubicar geográficamente la región que habitaron los mixtecos o ñuusavi, los hombres de las nubes.

No sólo Oaxaca ocupó la atención académica de Caso, también estudió la cultura y la religión de los aztecas y se convirtió en uno de sus principales peritos. Descifró muchas de las famosas piedras grabadas que representaban a las deidades del México central, como la Piedra del Sol, que había sido la preocupación de muchos otros estudiosos de las épocas anteriores. Caso encontró que también se trataba de un sistema calendárico, parte de la cultura mexica en cuya raíz se encuentran sus mitos de origen. También descifró límites de territorios y una gran cantidad de eventos que involucraban a los dioses de lo que él llamaba el Pueblo del Sol, el pueblo mexica, que controló en gran medida los destinos de los demás pueblos mesoamericanos en una época cercana a la conquista hispana.

La arqueología de México le debe mucho a don Alfonso Caso, ya que, como el gran visionario que fue, fundó las instituciones que aseguraban la continuidad de los estudios arqueológicos, como la Escuela Nacional de Antropología, en la que formó a una gran cantidad de estudiantes, entre los que se cuentan los nombres de arqueólogos y antropólogos de la talla de Ignacio Bernal, Jorge R. Acosta, Wigberto Jiménez Moreno, Arturo Romano, Román Piña Chan y Barbro Dahlgren, sólo por mencionar algunos; y la Sociedad Mexicana de Antropología, orientada a propiciar el intercambio constante de ideas entre los científicos enfocados al estudio del hombre. Caso también fundó aquellas instituciones que aseguraban la protección del patrimonio arqueológico de los mexicanos, como el Instituto Nacional de Antropología e Historia y el Museo Nacional de Antropología. Sus estudios de las antiguas culturas le hicieron valorar a los indígenas actuales que luchan por su reconocimiento en el México de hoy. Para su apoyo, fundó el Instituto Nacional Indigenista, organismo que aún dirigía poco antes de morir en 1970, en su afán de revalorar, como él decía, “al indio vivo, a través del conocimiento del indio muerto”. En nuestros días, las instituciones que Caso fundó aún persisten en el centro de la política cultural nacional, como una muestra de la visión extraordinaria de este científico, cuya única misión, como él mismo reconocía, era la búsqueda de la verdad.

Fragmentos extraídos por *ResúmeME*:

- Alfonso Caso y la arqueología mexicana
  Uno de los pilares indiscutibles de la llamada época dorada de la arqueología mexicana fue el doctor Alfonso Caso y Andrade, ilustre arqueólogo cuya sabiduría, dedicación y ética en el desempeño de sus investigaciones, tanto en el campo como en el laboratorio, dejaron un acervo de primer orden.
  puntaje: 0.588757505145

- Don Alfonso Caso fue especialmente importante en las investigaciones del área cultural de Oaxaca; a partir de 1931, y por más de veinte años, se dedicó al estudio de Monte Albán, sitio al que encontró convertido en terrenos de cultivo, con mogotes llenos de vegetación añosa.
  puntaje: 0.640742200617

- La arqueología de México le debe mucho a don Alfonso Caso, ya que, como el gran visionario que fue, fundó las instituciones que aseguraban la continuidad de los estudios arqueológicos, como la Escuela Nacional de Antropología, en la que formó a una gran cantidad de estudiantes, entre los que se cuentan los nombres de arqueólogos y antropólogos de la talla de Ignacio Bernal, Jorge R. Acosta, Wigberto Jiménez Moreno, Arturo Romano, Román Piña Chan y Barbro Dahlgren, sólo por mencionar algunos; y la Sociedad Mexicana de Antropología, orientada a propiciar el intercambio constante de ideas entre los científicos enfocados al estudio del hombre.
  puntaje: 0.749621137829

- Para su apoyo, fundó el Instituto Nacional Indigenista, organismo que aún dirigía poco antes de morir en 1970, en su afán de revalorar, como él decía, “al indio vivo, a través del conocimiento del indio muerto”.
  puntaje: 0.612304985264

Texto 9. (español)

Resultados obtenidos en la evaluación. Precisión: 0.79487 Exhaustividad: 0.61538

Secuencian en el Cinvestav más de 35 mil genes del agave

La investigación, desarrollada en la Unidad Irapuato de ese centro con la especie tequilana Weber variedad azul, busca nuevas aplicaciones alimentarias y medicinales de la planta

DE LA REDACCIÓN

Científicos del Centro de Investigación y de Estudios Avanzados (Cinvestav) Unidad Irapuato, del Instituto Politécnico Nacional (IPN), secuenciaron más de 35 mil genes del agave, de los cuales han identificado aquellos que pueden estar involucrados en el proceso de maduración y síntesis de inulinas (frútanos), con la finalidad de que en un futuro se pueda acortar el tiempo de producción en el sector.

El grupo encabezado por June Simpson Williamson, del departamento de Ingeniería Genética y del Laboratorio Nacional de Genómica para la Biodiversidad (Langebio), trabaja desde hace seis años en el estudio del genoma funcional (transcriptoma) del Agave tequilana Weber variedad azul, con el fin de encontrar métodos más eficaces para la producción de la planta.

El equipo de Simpson Williamson ha identificado genes encargados del control de procesos de floración, de los cuales entre cuatro y cinco son los que pueden ser útiles para manipular el proceso de maduración del agave.

El proyecto, que se trabaja en colaboración con el Colegio de Posgraduados (Colpos), tiene el propósito de hallar el método para manipular la producción de azúcares y los tiempos de floración, los cuales tienen aplicaciones importantes para el manejo de las plantas en el campo y, en consecuencia, también para la producción de bebidas alcohólicas.

La meta a largo plazo es hacer más eficiente la explotación del agave en el campo y además descubir nuevas aplicaciones, como el desarrollo de suplementos alimenticios o de la medicina tradicional a partir de esta planta.

La ganadora al Premio Uhuari Mujer y Ciencia 2009, del Instituto de la Mujer en Irapuato, detalló que los estudios se hacen por medio de un sistema heterólogo in vitro, el cual consiste en purificar las enzimas y confirmar que tienen las actividades que la investigación propone.

Los investigadores realizan ese trabajo en tres etapas: la primera es la extracción del ARN, es decir, colectar muestras del Agave tequilana Weber variedad azul.

En la segunda se obtiene información de los genes expresados y en tejidos específicos de la planta que permiten ver sus pautas de expresión y regulación.

Finalmente se realiza un análisis bioinformático por medio del cual se obtienen datos más precisos de los genes que ayudan a conocer las enzimas presentes y cómo, cuándo y dónde se producen las inulinas (azúcares) que son aprovechadas para la producción del tequila.

Fragmentos extraídos por *ResúmeME*:

- Científicos del Centro de Investigación y de Estudios Avanzados (Cinvestav) Unidad Irapuato, del Instituto Politécnico Nacional (IPN), secuenciaron más de 35 mil genes del agave, de los cuales han identificado aquellos que pueden estar involucrados en el proceso de maduración y síntesis de inulinas (frútanos), con la finalidad de que en un futuro se pueda acortar el tiempo de producción en el sector.
  puntaje : 0.805100128388

- La ganadora al Premio Uhuari Mujer y Ciencia 2009, del Instituto de la Mujer en Irapuato, detalló que los estudios se hacen por medio de un sistema heterólogo in vitro, el cual consiste en purificar las enzimas y confirmar que tienen las actividades que la investigación propone.
  puntaje: 0.772622947185

- Finalmente se realiza un análisis bioinformático por medio del cual se obtienen datos más precisos de los genes que ayudan a conocer las enzimas presentes y cómo, cuándo y dónde se producen las inulinas (azúcares) que son aprovechadas para la producción del tequila.
  puntaje: 0.815007252184

Texto 3. (inglés)

Resultados obtenidos en la evaluación. Precisión: 0.93333 Exhaustividad: 0.73333

Is Technology Killing Leisure Time?

New surveys suggest that ubiquitous technological tools are killing off leisure time, especially for younger workers and students -- that would be you -- who are working longer hours, taking fewer and shorter vacations (when they do go away, they take their cells, Palms and laptops along) and say they are more stressed than any other segment of the population. Opportunistic employers aren't helping, actually encouraging employees to do personal chores on the Net -- from their desks. Wasn't technology supposed to free us from workplace shackles?

Americans for centuries have believed that new labor saving devices will free us from the burdens of the workplace and give us more time to ponder philosophy, goof off, explore the arts, and hang around with friends and family.

So here we are at the start of the 21st Century, enjoying one of the greatest technological boom times in human history, and nothing could be further from the truth.

The very tools that were supposed to liberate us have bound us to our work (and schools) in ways that were inconceivable just a few years ago. But technology almost never does what we expect.

Almost all of us -- especially the people reading this -- have less leisure time than ever. We work harder, take fewer vacations for shorter periods of time, report more stress than almost any other demographic group and find the boundaries between work and play increasingly blurred. Computing and communications technologies are destroying the idea of privacy and leisure.

According to a new study reported in the July issue of American Demographics magazine, as the distinctions between home and the workplace fade, more and more of us go online from our offices to buy the things and perform the tasks we used to do when we got home. At first, employers were wary of workers going on the Net. But they've learned to love and encourage it, since it keeps employees chained to their desks for longer hours.

In l999, the researchers report, 19 percent of the total population had Net access at work, compared with just seven percent in l996. Employers, who now expect workers to be available for longer periods, understand that they have to let them to do their chores online. At work, Net surfers go first to news, information and entertainment sites. Then they hit search engines, marketing/corporate sites, sex sites and retailing shopping sites, in that order.

But there's a huge trade off for this convenience. Inforum's l999 Survey from the MEDSTAT group, reports American Demographics, found that adults aged 35 and younger were the most stressed people in the population. Nearly seven in 10 said they were "somewhat" to "extremely" stressed, an astonishing contrast to adults over 65: 31 percent of them said they had almost no stress in their lives at all.

More than a third of adults under the age of 25 say they don't get enough sleep most or all of the time. No wonder. More than half of them report that they didn't have time to take a vacation, according to the Travel

Industry Association of America. When younger people do travel, they don't take much of a break: 42 percent of travelers who go away for just a weekend are aged 18 to 34 -- the largest share of any single demographic group. Of course, maybe they have less disposable income or have young children they can't leave for long. But if you think about people you know in this age group, it's also obvious that they have trouble disconnecting from work, thanks mostly to technology, and they're also afraid to show employers that they're not indispensable. It may also be true that openly or not, more employers expect their workers to be around all the time.

Before the Net, cell phones and Palms, the lines between work and leisure time were markedly clearer. People left their offices at a predictable time, were often completely disconnected from and out-of-touch with their jobs as they traveled to and from work, and were off-duty once they were home. That' s no longer true. Even in a competitive job market, employers expect workers to put in longer hours and to be available almost constantly via fax, cell, e-mail or other communications devices. Bosses, colleagues and family members -- lovers, buddies and spouses too -- expect instant responses to voice-and e-mail messages.

Employers have thus begun to pay the small price of allowing their round-the-clock workers to shop and communicate online, found the AD study.

The American Demographic report validates the suspicion that corporatist employers are taking advantage of new technologies and of workers' anxieties to demand longer hours and increased productivity -- the very things new technologies were supposed to liberate people from.

Although there are no known studies relating to college students and their work hours, it seems they are also bound to their desks and dorms by environments in which faculty, friends and other members of the college community increasingly do their work online. Studies of time spent on instant messaging services would probably show staggering use. And research possibilities online are boundless.

Few of us manage to buck this trend, apart from some neo-Luddites. Half of all Americans now own a cell phone, and more than 46 per cent of pleasure travelers take their phones with them when they go away, reports the Travel Industry Association. More than 18 per cent take their pagers and 6 per cent their laptops, while 10 per cent check e-mail on vacation. Younger Americans are living in a hyperactive information culture.

According to the Bureau of Labor Statistics, 40 per cent of men worked more than 40 hours a week in l998, an increase of 5 percentage points in the last two decades. As for women, 22 per cent worked more than 40 hours aweek, compared with just 14 per cent in 1979.

So it's not surprising that a l998 General Social Survey conducted by the National Opinion Research Center at the University of Chicago found that more than 40 per cent of American workers say they come home from work exhausted, up from 36 per cent in l989. Young married couples report that they work an average 26 per cent more hours each year than they did 30 years ago.

Aside from long hours, the nature of work has changed. Economist and author Richard Sennett (The Corrosion of Character: The Personal Consequences of Work in the New Capitalism) and Joanne B. Ciulla (The Working Life: The Promise and Betrayal of Modern Work), point out changes in the nature of work itself.

"Flexible" work projects, the growing number of part-time workers, and a culture that embraces and even celebrates continuous layoffs, down-sizings and re-engineerings have rendered almost everyone's work life

stressful and unstable. Workers work harder and longer, move more often, change their work tasks more frequently, and are nevertheless constantly subject to dismissal or its threat.

This isn't what technology is supposed to be doing for us. New technologies, from genetic research to the Net, offer all sorts of benefits and opportunities. But when new tools make life more difficult and stressful rather than easier and more meaningful -- and we are, as a society, barely conscious of it -- then something has gone seriously awry, both with our expectations for technology and our understanding of how it works.

Fragmentos extraídos por *ResúmeME*:

- New surveys suggest that ubiquitous technological tools are killing off leisure time, especially for younger workers and students -- that would be you -- who are working longer hours, taking fewer and shorter vacations (when they do go away, they take their cells, Palms and laptops along) and say they are more stressed than any other segment of the population.
  puntaje: 0.735560030389

- But technology almost never does what we expect.
  puntaje: 0.545887650569

- According to a new study reported in the July issue of American Demographics magazine, as the distinctions between home and the workplace fade, more and more of us go online from our offices to buy the things and perform the tasks we used to do when we got home.
  puntaje: 0.549678341434

- But if you think about people you know in this age group, it's also obvious that they have trouble disconnecting from work, thanks mostly to technology, and they're also afraid to show employers that they're not indispensable.
  puntaje: 0.531841683833

- Although there are no known studies relating to college students and their work hours, it seems they are also bound to their desks and dorms by environments in which faculty, friends and other members of the college community increasingly do their work online.
  puntaje: 0.521116405753

- This isn't what technology is supposed to be doing for us.
  puntaje: 0.555079556404

- But when new tools make life more difficult and stressful rather than easier and more meaningful -- and we are, as a society, barely conscious of it -- then something has gone seriously awry, both with our expectations for technology and our understanding of how it works.
  puntaje: 0.674850113097

Texto 4. (inglés)

Resultados obtenidos en la evaluación. Precisión: 0.86666 Exhaustividad: 0.73333

Following Trash and Recyclables on Their Journey

By MIREYA NAVARRO

Where does all the trash go?

Karin Landsberg, 42, a self-described "eco-geek" in Seattle, was so curious that she invited researchers from the Massachusetts Institute of Technology into her home last month to fish 12 items out of her garbage and recycling bins — a can of beans, a compact fluorescent light bulb — and tag them with small electronic tracking devices.

Her trash is now on its journey to the place where it goes to die or be reborn.

The Architectural League of New York went through a similar trash-tagging exercise as part of the same project when it moved its offices from midtown Manhattan to SoHo two weeks ago. Among the discarded items tagged were a coffee cup, a filing cabinet, a book shelf, a broken wine glass and an empty plastic bottle that had held liquid soap.

"All they can tell me up to this point is that some of the stuff has gone through the Lincoln Tunnel," said Gregory Wessner, director of digital programs and exhibitions for the league. "It is on the move. We're really excited to know what happens."

Through the project, overseen by M.I.T.'s Senseable City Laboratory, 3,000 common pieces of garbage, mostly from Seattle, are to be tracked through the waste disposal system over the next three months. The researchers will display the routes in real time online and in exhibitions opening at the Architectural League of New York on Thursday and the Seattle Public Library on Saturday.

One purpose of the project, said Carlo Ratti, director of the lab, is to give people a concrete sense of their impact on the environment in a way that might lead them to change their habits.

"If you see where a plastic bottle ends up, a few miles down the road in a dump, you may want to get tap water or some other container for the water," Mr. Ratti said.

Collecting, transporting, storing and getting rid of garbage is a costly and often daunting task for cities. Lynn Brown, a spokeswoman for Waste Management Inc., a company that runs both landfills and recycling centers nationwide and is helping to underwrite the tracking project with $300,000, said garbage moved through a vast network of sites run by multiple contractors, which makes it challenging to find the most efficient way to handle it.

It also means hundreds of possible journeys for trash.

“From a logistics standpoint, it’s a very complicated situation,” Ms. Brown said. “When you look at how waste is handled in different cities, it’s like snowflakes. It’s all different.”

Other factors are also in play in the travel of recyclables like metal and plastic. Among them are price fluctuations that may make it cheaper for a company to ditch items than to recycle them, contamination that makes a can or paper useless, and human error in sorting or transporting material.

Even when an item is headed where it is supposed to go, “does it fall off the boat, or truck, or whatever?” said Ms. Landsberg, a transportation planner for Washington State. “Is the stuff actually made into something useful in this country? Does it all end up shredded and shipped to China, where who knows what happens to it?”

To answer some of those questions, the M.I.T. team is using battery-powered tags based on cellphone technology.

The researchers say it will take several months to analyze the data generated by the cellular signals. But they have already noticed that while some trash reaches its destination in a couple of days, other items may take four or five weeks to wind their way to landfills or recycling and waste processing plants.

In Seattle, where researchers recruited volunteers for the project through the Seattle Public Library’s Web site, the Seattle Public Utilities newsletter and other local publications, about 500 pieces have been tagged. One item, an aluminum can disposed of at a residence, traveled 2.5 miles to a recycling facility in the city in just under two days.

In New York, where 50 items were tagged at the Architectural League’s offices, a recyclable plastic bottle picked up at Madison Avenue and 51st Street traveled 18.3 miles over four days to Kearny, N.J., and is still en route, said Assaf Biderman, associate director of the M.I.T. lab.

The tracking has its limitations. Even though the tags have a battery life of two to six months and can report back from overseas, they can easily be crushed in transit inside garbage trucks and at processing facilities. Mr. Biderman said a paper cup taken from a Seattle residence sent signals for seven and a half days before it went silent and is assumed to have been destroyed.

But the researchers say most tags are likely to travel far enough to show which items go where and how long it takes them to reach a destination, yielding information about inefficiencies in the waste management system. In coming weeks the project is expected to gain an international component when 50 items are tagged in London, Mr. Biderman said.

Ms. Brown of Waste Management said her company hoped that the experiment could eventually help shorten or avoid overlaps in routes traveled by its 24,000 garbage trucks and to find more central locations for transfer and disposal.

Ultimately, she said, “we’re looking for ways to recycle more and to do it all more efficiently.”

Brett Stav, a senior planning and development specialist for the Seattle Public Utilities, which collects about 2,100 tons of trash and recyclables a day, said that aside from the help with logistics, he saw “tremendous educational value” in the experiment.

“There is this hidden world of trash, and there are ramifications to the choices that people make,” Mr. Stav said. “People just take their trash and put it on the curb and they forget about it and don’t think about all the time and energy and money put into disposing of it.”

The point is well taken by Ms. Landsberg of Seattle, who is so environmentally conscious that she keeps a worm bin to compost her food waste.

“If I found out that it wasn’t going where I think it does, if it is less recycled than I hoped,” she said she “might think about buying less of it or doing without.”

“Maybe it is more about the reduce than the re-use,” she said.

Fragmentos extraídos por *ResúmeME*:

- Karin Landsberg, 42, a self-described “eco-geek” in Seattle, was so curious that she invited researchers from the Massachusetts Institute of Technology into her home last month to fish 12 items out of her garbage and recycling bins — a can of beans, a compact fluorescent light bulb — and tag them with small electronic tracking devices.
  puntaje: 0.755925725143

- We’re really excited to know what happens.” Through the project, overseen by M.I.T.’s Senseable City Laboratory, 3,000 common pieces of garbage, mostly from Seattle, are to be tracked through the waste disposal system over the next three months.
  puntaje: 0.690118458703

- In New York, where 50 items were tagged at the Architectural League’s offices, a recyclable plastic bottle picked up at Madison Avenue and 51st Street traveled 18.3 miles over four days to Kearny, N.J., and is still en route, said Assaf Biderman, associate director of the M.I.T.
  puntaje: 0.627402672804

- Ultimately, she said, “we’re looking for ways to recycle more and to do it all more efficiently.” Brett Stav, a senior planning and development specialist for the Seattle Public Utilities, which collects about 2,100 tons of trash and recyclables a day, said that aside from the help with logistics, he saw “tremendous educational value” in the experiment.
  puntaje: 0.634043619189

- “If I found out that it wasn’t going where I think it does, if it is less recycled than I hoped,” she said she “might think about buying less of it or doing without.” “Maybe it is more about the reduce than the re-use,” she said.
  puntaje: 0.803489932036

Texto 5. (inglés)

Resultados obtenidos en la evaluación. Precisión: 0.466667 Exhaustividad: 0.533333

Émile Zola

Émile Zola (1840-1902), French novelist, essayist, and critic, the chief advocate and practitioner in France of a movement known as naturalism. Naturalist writers aimed at an objective depiction of life and regarded human behavior as determined by hereditary instincts and emotions and the social and economic environment, rather than by free human choice.

Born in Paris, Émile Édouard Charles Antoine Zola spent his formative years in Aix-en-Provence in the south of France. Although Zola's father died when Émile was seven, Émile and his mother remained in Aix until poverty forced them to move to Paris in 1858. There the young Zola eked out a living working as a clerk for the publishing house Hachette and writing literary and political articles for newspapers. His knowledge and understanding of poverty, evident in his later novels, was due in part to personal experience.

Zola's published work began with a collection of stories, Contes à Ninon (1864; translated as Stories for Ninon, 1888), and a full-length novel, La confession de Claude (1865; Claude's Confession, 1882). Neither received much attention, but in 1867 Zola achieved notoriety with Thérèse Raquin (translated 1962), a lurid tale of lust and murder.

Inspired in part by La comédie humaine (1842-1848; The Human Comedy, 1895-1900), a vast cycle of novels by French writer Honoré de Balzac, Zola then conceived of a series of 20 novels, Les Rougon-Macquart, which would relate the history of a single family during the reign of French Emperor Napoleon III (1852-1870). In these novels he sought to imitate the scientific method through detailed, objective observation of his characters under controlled conditions. He also sought to incorporate ideas on the ways in which heredity and the environment shaped human character and determined human behavior—ideas that he had encountered in his reading of French critic and philosopher Hippolyte Taine, British scientist Charles Darwin, and French scientist Prosper Lucas. Zola considered heredity modified by environment to have the force of fate.

Zola accomplished his great task, beginning in 1871 with La fortune des Rougon (The Fortune of the Rougons, 1886) and ending in 1893 with Le docteur Pascal (Doctor Pascal, 1957). After publishing the seventh of these novels he read Introduction à l'étude de la médecine expérimentale (1865; An Introduction to the Study of Experimental Medicine, 1927) by French physiologist Claude Bernard and tried to adapt this scientific method of observation and experimentation in the remainder of his work. In 1880 Zola published the essay "Le roman expérimental" ("The Experimental Novel," 1893), in which he developed these ideas and articulated his concept of naturalism and the naturalistic novel. He further explored these ideas in "Les romanciers naturalistes" (The Naturalist Novelists, 1881).

Zola visited the locations in which the action of his books took place, observed closely, and took copious notes. In his novels he introduced characters inspired by his research, studied their hereditary backgrounds (often familiar to readers of earlier novels in the cycle), and observed how their lives played out in their

world. Although Zola's science sometimes seems amateur, it lent coherence to the enormous cycle of novels. Some think it fortunate that Zola's epic imagination often eclipses his scientific aspirations.

Although Les Rougon-Macquart includes many excellent novels, two works in this series are recognized as among the best French novels of the 19th century: L'assommoir (1877; translated 1879) and Germinal (1885; translated 1885). The protagonist of L'assommoir, Gervaise Macquart, a launderer in a cheap quarter of Paris, is abandoned by Étienne Lantier, the father of her two illegitimate children. She seems to have a change of luck when she meets and marries the roofer Coupeau and acquires her own laundry facility. She prospers until Coupeau falls from a roof, takes to drink, and carries her with him in his decline into moral depravity. The French word assommoir means a club or sledgehammer used to fell something by a blow; in slang it means a low-life tavern. In L'assommoir it refers to Colombe's cheap saloon that houses a distilling apparatus, which Zola transforms into a fantastic beast, an evil monster spouting steam and chomping furiously as if to devour the world. This distillery becomes the emblem for alcoholism in the novel. Both Coupeau and Lantier are genetically predisposed to the disease, and their environment and circumstances leave them little chance of escape. Despite the baseness of these characters and their world, Macquart is genuinely appealing, and the reader is moved to compassion by her fate.

Germinal, which takes place among a community of exploited miners, examines such issues as unionization and the economic and political doctrine known as socialism. The son of Gervaise Macquart, Étienne Lantier, wanders into this community, which is on the verge of a strike. He finds work in the mine pits (which are depicted as a voracious monster devouring human bodies) and is befriended by Toussaint Maheu, with whose daughter, Catherine, he falls in love. Lantier, predisposed by heredity to homicidal violence and alcoholism but fundamentally a moral and good man, soon becomes a leader of the miners. His emerging socialist sympathies are reinforced by the surrounding misery, his reading of socialist literature, and his acquaintance with Souvarine, an exiled Russian revolutionary. Zola contrasts the degradation and suffering of the Maheu family with the complacency and prosperity of the Grégoire family, who own stock in the mines. Near the end of Germinal, the head of the Maheu family strangles the Grégoires' only child, and Catherine Maheu dies in the mines. During the course of the novel the miners stage an unsuccessful strike and return to the pits, humiliated and desperate, and Lantier returns to the road that had brought him to the mining community. Despite the pessimism of this ending, it is spring and the world is germinating, offering the hope of "a revolutionary April, a flight of a decrepit, sick society into the springtime."

Zola is also famous as the author of "J'accuse" ("I accuse"), an open letter to the president of France (published in the newspaper L'aurore in 1898), in which he denounced French army officials for lying in their effort to convict Captain Alfred Dreyfus, a Jew, of treason. Dreyfus was later found innocent. Zola wrote a fictional account of the case in his novel Vérité (1903; Truth, 1903).

Zola died accidentally of carbon monoxide poisoning in 1902. The streets of Paris were lined with mourners as his casket passed through the city. He had come to be known as a champion of the innocent, an upholder of justice, and a defender of the downtrodden. As novelist Anatole France declared in his eulogy, Zola had become "the conscience of mankind."

Fragmentos extraídos por *ResúmeME*:

- Naturalist writers aimed at an objective depiction of life and regarded human behavior as determined by hereditary instincts and emotions and the social and economic environment, rather than by free human choice.
  puntaje: 0.419258166267

- Zola's published work began with a collection of stories, Contes à Ninon (1864; translated as Stories for Ninon, 1888), and a full-length novel, La confession de Claude (1865; Claude's Confession, 1882).
  puntaje: 0.484649764614

- Inspired in part by La comédie humaine (1842-1848; The Human Comedy, 1895-1900), a vast cycle of novels by French writer Honoré de Balzac, Zola then conceived of a series of 20 novels, Les Rougon-Macquart, which would relate the history of a single family during the reign of French Emperor Napoleon III (1852-1870).
  puntaje: 0.472612934338

- He also sought to incorporate ideas on the ways in which heredity and the environment shaped human character and determined human behavior—ideas that he had encountered in his reading of French critic and philosopher Hippolyte Taine, British scientist Charles Darwin, and French scientist Prosper Lucas.
  puntaje: 0.46650850677

- Despite the pessimism of this ending, it is spring and the world is germinating, offering the hope of "a revolutionary April, a flight of a decrepit, sick society into the springtime."

  Zola is also famous as the author of "J'accuse" ("I accuse"), an open letter to the president of France (published in the newspaper L'aurore in 1898), in which he denounced French army officials for lying in their effort to convict Captain Alfred Dreyfus, a Jew, of treason.
  puntaje: 0.607274616704

- Dreyfus was later found innocent.
  puntaje: 0.503451729044

Texto 1. (alemán)

Resultados obtenidos en la evaluación. Precisión: 0.83333 Exhaustividad: 0.83333

MilitärAnschlag

Italienische Soldaten und Zivilisten sterben bei Anschlag auf Nato-Militärfahrzeug
Selbstmordanschlag in Kabul: Mindestens 16 Tote
Blutiger Angriff auf ein Nato-Militärfahrzeug im Zentrum Kabuls: Bei einem Selbstmord-Anschlag kamen mindestens sechs italienische Soldaten und zehn Zivilisten ums Leben. Mehr als 50 Menschen wurden verletzt. Italiens Regierungschef Berlusconi will die Truppen nun "schnellstmöglich" abziehen.

Kabul - Bei einem Selbstmordattentat in der afghanischen Hauptstadt Kabul sind am Donnerstag sechs italienische Soldaten sowie mindestens zehn Zivilisten getötet worden. Mindestens 55 weitere Menschen wurden nach Angaben eines Sprecher des Innenministeriums verletzt.
Ein Selbstmordattentäter rammte einen italienischen Militärkonvoi und zündete den Sprengsatz. Dutzende Fahrzeuge brannten. Nach Informationen von SPIEGEL ONLINE versuchten mehrere Mitglieder der Isaf-Truppen, die Opfer aus den brennenden Fahrzeugen zu befreien. Augenzeugen berichten von vielen Toten, die nach der Attacke auf der Straße lagen. Bei dem Anschlag wurden 21 Geschäfte sowie mehrere Autos zerstört.

Die Taliban bekannten sich zu der Tat. Ziel seien die ausländischen Truppen in Afghanistan gewesen, sagte ein Sprecher der radikalen Islamisten.
Der Anschlag ereignete sich auf einer Straße, die die US-Botschaft mit dem Flughafen der afghanischen Hauptstadt verbindet und oft von ausländischen Militärkonvois genutzt wird. Unweit des Ortes befinden sich zahlreiche Vertretungen. Kurz vor der Explosion hatte der afghanische Staatschef Hamid Karzai eine Pressekonferenz im ebenfalls nahe gelegenen Präsidentenpalast abgehalten.

Es war der vierte größere Anschlag in Kabul innerhalb von fünf Wochen. Zu dem Anschlag am 8. September in der Nähe des Eingangs zum Militärflughafen bekannten sich ebenfalls die radikalislamischen Taliban.
Italiens Ministerpräsident Silvio Berlusconi reagierte umgehend auf den Anschlag: "Wir sind alle überzeugt, dass wir Afghanistan schnellstmöglich verlassen müssen", sagte er am Rande des EU-Sondergipfels in Brüssel. Dies könne Italien aber nicht alleine entscheiden, sondern müsse es mit seinen Verbündeten absprechen - "sonst werden wir das Vertrauen der anderen Länder verraten."

"Das ist ein schmerzhafter Tag für Italien", sagte Berlusconi zu dem Attentat. Rom habe bereits vor dem Attentat "eine starke Reduzierung" seiner Truppen vorgesehen gehabt. "Wir werden auf diesem Weg weitergehen." Die in Italien mitregierende rechtspopulistische Lega Nord hatte zuvor einen Abzug bis Weihnachten verlangt. Berlusconi sagte dazu, bisherige Abzugspläne könnten beschleunigt werden, "sobald der Aufbau der afghanischen Sicherheitskräfte es erlaubt".

Fragmentos extraídos por *ResúmeME*:

- Selbstmordanschlag in Kabul: Mindestens 16 Tote

  Blutiger Angriff auf ein Nato-Militärfahrzeug im Zentrum Kabuls: Bei einem Selbstmord-Anschlag kamen mindestens sechs italienische Soldaten und zehn Zivilisten ums Leben. puntaje:0.640448583112

- Italiens Regierungschef Berlusconi will die Truppen nun "schnellstmöglich" abziehen.
  puntaje: 0.620273319033

- Ein Selbstmordattentäter rammte einen italienischen Militärkonvoi und zündete den Sprengsatz.
  puntaje0.601257994045

- Kurz vor der Explosion hatte der afghanische Staatschef Hamid Karzai eine Pressekonferenz im ebenfalls nahe gelegenen Präsidentenpalast abgehalten.
  puntaje:0.603790422001

- Italiens Ministerpräsident Silvio Berlusconi reagierte umgehend auf den Anschlag: "Wir sind alle überzeugt, dass wir Afghanistan schnellstmöglich verlassen müssen", sagte er am Rande des EU-Sondergipfels in Brüssel.
  puntaje:0.855435351427

Texto 1. (chuj)

Resultados obtenidos en la evaluación. Precisión: 1.0 Exhaustividad: 1.0

EL ÉXODO

A ton kel t'a chi' jun, kom a in tik a eb' ix innulej t'a Waxakana, t'a Jalumk'u, cha'wan eb' ay t'a Jalumk'u tik, jun ix ay t'a Waxakana, komo ay eb' ix t'a jun, ja' eb' ix, a in tik leman yoch wo'och, leman inchenhi, wilani yem te witz t'a jun pak'an t'a jun b'e ¿b'aj tax laj k'och jun tzan anima' chi'?,

--ilna' yemta' eb' winh ula' chi', tekan ula' eb' winh tzemta' chi', xinchi' icha tik t'a jun tzan eb' ix unin,

--tekan am laj eb' winh, ixschi' eb' ix,

--ilek'ek to naik b'aj olk'och eb' winh, ixinchi' t'a eb'.

Lan yilan eb' tzkot eb', ixk'och eb', k'och jun tzan anima' chi', axo winal jun, ay ix innulej, xit'iti' tzk'och eb' ix yet' spop yet' sk'u, to to niwan spop eb' ix yet'nak, niwan sk'ael eb' ix yet' nak,

--ay dios, tob'an a eb' ix konulej tzjawi, b'aj til tzk'och jun tzan eb' ix konulej tik, xinchi' t'a jun tzan eb' ix une'.

Ixk'och eb' ix b'ian:

--¿ayamach?

--ayin, xinchi', ochanh, xinchi' t'a eb' ix, ay dios, ¿b'aj tzachkoti? ¿b'aj tzexk'ochi?, ixinchi' t'a eb' ix,

--a tik naik tom tzachb'at ket'ok mato tzachkani, to lan kolajwieli, lan onhsmilancham eb' winh ejercito ¿tom wach' kaj tik naik?, lan kochami, ixchamkan cha'wan eb' winh ket' b'eyum t'a Jalumk'u, ay eb' winh ejercito, ixyik'kanelta' eb' winh t'a yol spat, ixsmilcham eb' winak cha'wan eb' winh ket' b'eyum, smilkancham eb' winh, yuj chi' jun ixonhkoti, to masanil anima' olsmilcham eb' winak, ina' San Pransisko ixcham anima', ixsmolb'ej jantak anima' eb' winh t'a yol templo, ixsjulanoch oxe' bombe winh, oxe', chanhe', oye', ixcham nhanhal yuj bomba, yuj to ay to eb' jun pitzan to, ixkani, ixyak' rosiar gasolina eb' winh spatik templo, ixyak'kanoch k'ak'al eb' winh, ixtz'aem jun a jun patil templo yib'an eb', tob' ixcham eb' masanil, tob' ichta' olyutok eb' anima' eb' masanil, aeb' olb'oanok, yuj chi' jun ixonhxiwi', lan lajcham anima', chab' aldea ixsmilcham eb'winh chi' naik, ixsmolb'ej anima' chi' eb' winh, yuj chi' jun ixonhxiwkoti, ixonhkot b'ian, yujnhej chi' ixonhelta', yuj chi' tzek'ach kik'a, ta to anab'en achcham jun, kananh, xalonh tik, tekan olonhkaxpaxb'atok, ta to agana achchami, kananh, olonhkaxpaxb'atok, olonhb'at t'a Mexico chi', olkotzab'ejb'ati, a t'a olkak' defender kob'a, tik lak'an kolonia tik, a ton kolonia tik olonhajok, tob' olyak' jak koposado eb' ¿b'aj wal wach' kok'och jun?, xschi' eb' ix innulej chi' t'ay in.

Ixink'e juknaj b'ian, ixin to pzx lan incheni, lan yo'och wo'och, ixwik'kankot jun in chokal wixim chi', ixwek'kanem t'a jun yol inb'el jun tzin chab' wo'ochi, yo ixinb'at tzakan yet' eb' ix b'ian, man xa ink'u, man xa tas ixwik'a, ixwak'tej kani, ixinoch tzakan yet' eb' ix, ixonhb'at t'a jun kolonia, a to t'a ixonhk'och elelal, chi' ton a t'a ayonhek' elelal chi' jun. Ixemkan jun k'u chi', axo t'a yewial ma'ay ay tekan olonhk'eek' chi',

--tekan tonhej, yal eb' winak,

--tekan olmeltzaj eb' winak, ixkochi', konhk'exek.

Muy ke a in tik toxo ixb'o ink'o'ol yet' jun ix wune' Soila tik, yab'ix-in yuj ix, b'o'el chi', kob'at b'ian, inmeltzaji a t'a,

--olb'at wak' a jun wune' tik, a t'a olwak' ajok, xinchi'.

Ixink'ex t'ay inchonhab' chi', lan to in k'exk'och jun, lan to wochk'och tek'ek-ok t'a yol sti' inpat, ixwilani, ixjaw k'en kopiro, ixjaw jun k'en ich tik, ixjaw jun k'en ich tik, masanil b'aj tzemk'och k'e'en, tinini' tzkan k'e'en, ijan ixtoineli, axo kosat b'at smak eb' winh ejersito chi' yet' k'en kopiro, tzyak'kanemta' chu'an k'en yarme winh t'ay onh, jantak mal olachxiwok, tzonxiwi, anima' tzonhxiwi, jantak tzonhtesxiwtej eb' winak, icha ton to olonhsjul eb' winak, a in tik pax wib' yuj xiwelal, jantak anima' laj te xiw eb', tzijtum unin ixcham yuj xiwelal, tzijtum anima' ixtexiwi, yuj chi' elnak, konhkoti, ichokta' maj onhxiwtej eb' winh, mato elonhta', mato ma'ay ayonh am laj t'a kopat chi' naik, yuj chi' eb' winh xiw t'ay onh, yuj chi' konh na konhkan t'a tik, yuj chi' kot na konh.

Ichta' yaj swayan, ichta' ajnak, ichta' swayanil, kelta' yuj eb' winh ejercito chi', yuj k'en k'opiro, el na konhkoti, jantak anima' chi' smilnak eb' winh yol nab'anej, malaj smul eb' malaj, masanil unin to to ajnak, masanil unin jun ujal, chab' ujal, oxe' ujal, chanhe' ujal unin, masanil eb' tzsmilcham eb' winh, nenis tak, kotak unin, tzsmilcham eb' winak, chuk tzsk'ulej eb' winak, ichachon tzsmilancham nok' eb' winak, ichachon tzonhyutej eb' winak, yuj chi' a onh tik tzonhkusi, siempre tzonhkusi yuj jantak kopamilia, ket' b'eyum chamnakkani, yuj chi' jun maxonh meltzaj laj, pero yiknhej tik t'a tik to malaj lum kolu'um b'aj tzonhaji, b'aj tzonhwa'l, yuj chi' tzkona' meltzaj kob'a, ta ma'ay wach' kajek' t'a tik, ichta' elnakonhkankoti, yuj chi' ayonhek' t'a skaxepal Mexico.

Fragmentos extraídos por *ResúmeME*:

- ¿b'aj tzexk'ochi?, ixinchi' t'a eb' ix,

  --a tik naik tom tzachb'at ket'ok mato tzachkani, to lan kolajwieli, lan onhsmilancham eb' winh ejercito ¿tom wach' kaj tik naik?, lan kochami, ixchamkan cha'wan eb' winh ket' b'eyum t'a Jalumk'u, ay eb' winh ejercito, ixyik'kanelta' eb' winh t'a yol spat, ixsmilcham eb' winak cha'wan eb' winh ket' b'eyum, smilkancham eb' winh, yuj chi' jun ixonhkoti, to masanil anima' olsmilcham eb' winak, ina' San Pransisko ixcham anima', ixsmolb'ej jantak anima' eb' winh t'a yol templo, ixsjulanoch oxe' bombe winh, oxe', chanhe', oye', ixcham nhanhal yuj bomba, yuj to ay to eb' jun pitzan to, ixkani, ixyak' rosiar gasolina eb' winh spatik templo, ixyak'kanoch k'ak'al eb' winh, ixtz'aem jun a jun patil templo yib'an eb', tob' ixcham eb' masanil, tob' ichta' olyutok eb' anima' eb' masanil, aeb' olb'oanok, yuj chi' jun ixonhxiwi', lan lajcham anima', chab' aldea ixsmilcham eb'winh chi' naik, ixsmolb'ej anima' chi' eb' winh, yuj chi' jun ixonhxiwkoti, ixonhkot b'ian, yujnhej chi' ixonhelta', yuj chi' tzek'ach kik'a, ta to anab'en achcham jun, kananh, xalonh tik, tekan olonhkaxpaxb'atok, ta to agana achchami, kananh, olonhkaxpaxb'atok, olonhb'at t'a Mexico chi', olkotzab'ejb'ati, a t'a olkak' defender kob'a, tik lak'an kolonia tik, a ton kolonia tik olonhajok, tob' olyak' jak koposado eb' ¿b'aj wal wach' kok'och jun?, xschi' eb' ix innulej chi' t'ay in.

puntaje: 0.931288972867

Texto 1. (tarahumara)

Resultados obtenidos en la evaluación. Precisión: 1.0 Exhaustividad: 0.66666

BACHÁWARA RA'ÍCHARI

Mapuarí nejé pé usáni á osá naó ne bamíire a'rí míi "Wagéachi" ne betére muribépi San Luis Majimáchi. Siné rawé kené onó míi Kirílichi simíre pé okwáa namúti rarísia mapu regá: Okuá kilo a'káwari mapu gajé akáwibo, a'rí okuá kilo harina, okua kilo oná oréame mapu sinéame Rarámuri we ra'ire gumiyá Kobísi lókaga, Á bera raráre okuá kilo plátano, okuá kilo na'rási a'rí ajaré limóni.Yé plátano, na'rási a'rí limóni bire oserí periódico aniríachi achagá yáriru.

Mapuarí ramé go'yá mochíwe wera plátano a'rí na'rási sí, kéne Onó anére á míi kéne Eyé oserí e'negá: "Yé bawéra mapu bawerága uchúwi jéna ra'íchari jú, a míi chabóchi ko wé ga'rá ra'ícha wera bawéra e'negá. Nejé á né anére á ajaré chabochi mapu nejé machí a'rí abói á ga'rá machí cho oserichi ra'ícha mapu nichí a'íchema biré utáa oserí mapu nejé togé. Arí sinéame níchí nére mapu ikí nichí anére wera bachabéera. Á bera wé sema regá rá ícha wera oserí, a'rí wera chabochi á bera ga'rá epené osayá a'rí ra'ícha wera oseríchi". Ke né machí chú regá, nori nejé binói pachána né nátare: "Nejé ko pée á né ochérosa nejé simée aminá mekabé nóchaga, a'rí ne anéma wera chabochi mapu nichí binírima osayá a'rí oseríchi ra'ícha."

We wi'rí simírore -¿chi yéna wi'rí ré? Tabiré ne machí. Siné rawé anére kéne Onó á míi kéne Eyé: "Ne mayé mapu má ga'rá jú mapu tabóo Patricio á míi Sogíchi mapu binimée osayá a'rí oseríchi ra'ícha, mapu má á warubée jú..." Arí uchécho simírore wé wiribé ru, ke tabiré umérore nichi toyá á míi Sogíchi mapu ké wési itegé mapu tibuma chibá a'rí suwába namuti. A'rí weká rawé osípo ku rowína á míi Sogíchi. Kéne Onó binéri biré rawé osimí nawayá míi Kirílichi a'rí Kéne Eyé a'rí nejé sí, a'rí kó biré rawé miná nasipa sí nima re. Naa Kirílichí a'rí míi Sogíchi nasipa rawe jú rú, a'rí beikiá rawé risábasa re míi Sogiichi, a'rí osá naó rawé á ga'rá gurípo re. Ara regá wé simáre rawéwari.

Siné bamíbari San Luis omáwachi, okuá siríame á míi Sogíchi simíbare ba'ré bayéma mapu Misa animée míi San Luis Gonzága omáwachi a'rí mapu pagóma kúuchi. A bera norínare biré ba'ré má wa'rú remarí Daniel García de Alba rewéame. Pée achigá ra'íchame nori wé garega ra'íchame a'rí wé chá regá e'néname nori pée á achigá.

San Luis omáwachi (osá makói miná biré awé Júniochi mechá, a'rí biré mili miná kimakói siento aminá naósa makói okúa bamíbari). Kimakói be'á re má ta síré mí mapu wirí re'obá kené Onó a'rí nejé sí, arí má tá imíbe míi ré'oba jubá ara wirí komerách yé komerách á bera niwe biré wa'rúportali. Ara bera biré wa'rú reé moba asáre wera ba'ré wé gániriga ra'ichaga ajaré chabochi úuga arí ajaré rarámuri á ajaré kuruwi chó.

Má reporása wera ba'ré má ne wiríbare wera ba'ré owiná matóchi jubá e'néga mapu ikí osíi atigé wera ba"ré. Kene Onó ko nichí owiná wiríbare cho e'negó. Wera ba'ré biré kurí gite re'éa atigé. Wera kurí biré guwáara ami biré chabochi yása mapu chapimée arí binói biréera guwáara gite osá buresa ekárari, a'rí wera guwáara nasípa seméroga a'ri banisúka kú suráa nokayé wera burerúame.

Ara regá osísa besá, má nijire wera kurí a míi chabóchi mapu ara regá oráma.

Sinéame besá chokéame orása tabiré wési umérore surána wera burerúame. Nori wé iwéga burerúga ripíi wera kuri. Chabóchi ko tabiré umérore, A'rí má nijíre wera kuri á míi rarámuri.

Nibiré wesi cho umérore surána wera burerúame. A'rí má yáriru kene onó wera kurí á bera osíre besá nori tabiré cho umérore surana. A'rí wera ba'ré má nichí e'nére wé chá regá kiríi achigá á nichí majárire.

A'rí má nichí yáre wera kurí guwáara. Má ne osíre mapu rega orayé binói, mapu gite nejé wé ga'rá né e'negá wirigé, mapu regá osíi asáre binói. A'rí nasípa seméroga wera guwáara mapu regá orayé binói a'rí ne banisúre wé ga'rá suráre wera burirúame. We achigá nichí e'nére wera ba're a'rí uchecho osá ne osíre abe ga'rá ne umérore surána wera burerúame ajaréera pé nichí e'negá mochiwe e'wéri buséga. Namuti nichí anére wera ba'ré, nori tabiré ne inámure mapu gite tabiré ne machiyé castía ra'ícha. Arí wera ba'ré kene onó úuga ra'íchare nejé sí wirigé ara nori tabiré ne inámure chú aníriru.

A'rí misa suwinísamá ta simíbare kú bitichí wé sapúuga Nawagá kene onó
anére á míi kene eyé: "Sogichi simí towí ba'ré úuga. Ga'rá biwíame napácha uchési a'ri gasibáchari si, (mapu regá oráriwa mi re'rége bakochi). A'rí binói níwara goyáchi". (Biré chéerame goyáchi mapu kene eyé newáare chabée rakú gite, nori abe jíi nichí chéwi rayénari). "Má bá", anire kene onó. A'rí wé ne sewére á wé ne naráre cho, wé sewériru ru areweyá ru eyé. Nori ara regá ne simíre. Wé seweka ne anére « Ariosi ba » a míi kene eyé. Kene onó ko á simáma ruyé á míi Sogíchi nichi e'nema. Arí kene onó nichí weká wenomí nichí yáre: beikiá peso aminá naó nomí wera mapu 0.720 plata aniríame mapu ne akáame rarimée á míi Sogíchi. A'rí wera gawé mérame mapu ba'ré úuga eyéne, wera bera úuga nichí asériru.

Ma ta síre míi Sogíchi, ara nichí narére wera ba'ré Martínez Aguirre S.J., ermano "Rosákame" Luciano Blanco, ermano Enrique Ureña a'rí ermano Leopoldo de León « Mawiyá ». Sinéame jesuita aníire. A'rí sí akináaka suwába ermano jesuita biníriame nichí binírire osayá, castiya ra'ícha cho á'rí oseríchi raícha a'rí uché weká namuti. A'rí má ke itéere éruka nichí binírima usánisa bamíbari. Arí má nichí jurárirú míi mapu goná nocha wera etemari.

Wera ermano Mawiyá ma nichí yáre wera jiéro mapu gite ne nocháma a'rí
á ga'rá nichí binírire wera jiéro gite nocha. Arí wera ermano Pulido nichí binírire uché namúti nocha mapu regá zapato newayá a'rí á uché ajaré namúti orayá. A'rí wera okúanika ermano Mawiyá a'rí Pulido sí nichí aneyé: "Ga'rá machibóo pichíka nocha a'rí ga'rá cho machibóo wichí biwayá nocha biré chiní samíame gite". Ara regá bera simárore weká bamíbare. Siné rokó ne norínare anáaka Chiwawa sí biré ermano marista Arkadio García rewéame úuga. A'rí ne ra'íchare suwába kene rawéwari, we ga'níriga gipúre mapu iki ne ra'íchega enagé a'rí anire binoi: "Osá kiri biré oserí suwába mujé níwara rawéwari". "Cha rináti ju". Ne anere. A'rí aníre binói: "Nejé nimí guwíroma. Nejé nimí anéma chu regá osibóo". Binoi míi Walajara simíre kú norínama rugá mapu nichí guwíroma anigá. Nori Onorúame kú repá bayére mapu binoi muribé asimée.

A'rí wé akiná ma sire okuá ermano marista biré Manueli Hernández Gaona rewéame, biréera Migeli García García rewéame. Abói nichí tánire mapu nejé osíima ajaré ra'íchari rarámuri a'rí ajaré ra'íchari castilla mapu aminá rarámuri newáma. Arí á gá níriga ne osáre. A'rí chiriwéga á míi abói sinéame, má ne níire mapu ké ne jú: Rarámuri ra'íchari osáame, castilla ra'íchari aminá rarámuri ra'íchari newáame.

Fragmentos extraídos por *ResúmeME*:

- Siné rawé kené onó míi Kirílichi simíre pé okwáa namúti rarísia mapu regá: Okuá kilo a'káwari mapu gajé akáwibo, a'rí okuá kilo harina, okua kilo oná oréame mapu sinéame Rarámuri we ra'ire gumiyá Kobísi lókaga, Á bera raráre okuá kilo plátano, okuá kilo na'rási a'rí ajaré limóni.Yé plátano, na'rási a'rí limóni bire oserí periódico aniríachi achagá yáriru.
  puntaje:0.831390314898

- Ke né machí chú regá, nori nejé binói pachána né nátare: "Nejé ko pée á né ochérosa nejé simée aminá mekabé nóchaga, a'rí ne anéma wera chabochi mapu nichí binírima osayá a'rí oseríchi ra'ícha.
  puntaje:0.530778518098

- Siné rawé anére kéne Onó á míi kéne Eyé: "Ne mayé mapu má ga'rá jú mapu tabóo Patricio á míi Sogíchi mapu binimée osayá a'rí oseríchi ra'ícha, mapu má á warubée jú..." Arí uchécho simírore wé wiribé ru, ke tabiré umérore nichi toyá á míi Sogíchi mapu ké wési itegé mapu tibuma chibá a'rí suwába namuti.
  puntaje:0.591248204553

- Sinéame jesuita aníire.
  puntaje:0.50245824711

- A'rí ne ra'íchare suwába kene rawéwari, we ga'níriga gipúre mapu iki ne ra'íchega enagé a'rí anire binoi: "Osá kiri biré oserí suwába mujé níwara rawéwari".
  puntaje:0.515994339286

- Nejé nimí anéma chu regá osibóo".
  puntaje:0.528035498023

- Abói nichí tánire mapu nejé osíima ajaré ra'íchari rarámuri a'rí ajaré ra'íchari castilla mapu aminá rarámuri newáma.
  puntaje: 0.531326759142

- A'rí chiriwéga á míi abói sinéame, má ne níire mapu ké ne jú: Rarámuri ra'íchari osáame, castilla ra'íchari aminá rarámuri ra'íchari newáame.
  puntaje: 0.601755881299

APÉNDICE C. RESULTADOS DETALLADOS DE LA EVALUACIÓN PARA LOS RESÚMENES GENERADOS

- Textos en Español

Jueces	**Texto 1**		**Texto 2**		**Texto 3**		**Texto 4**	
	Precisión	***Exhaus.***	***Precisión***	***Exhaus.***	***Precisión***	***Exhaus.***	***Precisión***	***Exhaus.***
Juez 1	1	0.66666	0.66666	1	0.33333	0.33333	0.33333	0.66666
Juez 2	1	0.66666	1	0.33333	0.66666	0.66666	0.33333	0.33333
Juez 3	0	0	0.33333	0.33333	0	0	0	0
Juez 4	1	0.33333	1	0.33333	0.33333	0.33333	0	0
Juez 5	0.66666	0.66666	1	0.66666	0.33333	0.66666	0.66666	0.66666
Juez 6	0.33333	0.66666	1	1	0.33333	0.33333	0.66666	1
Juez 7	0.66666	0.33333	0.66666	0.66666	0.66666	0.33333	0.33333	0.33333
Juez 8	1	0.66666	1	0.66666	0.66666	0.66666	0.66666	0.33333
Juez 9	0	0	0.66666	0.66666	0.66666	0.66666	0.33333	0.33333
Juez 10	0.33333	0.66666	1	0.66666	0.33333	0.33333	0.66666	0.66666
Juez 11	0.33333	0.33333	0.66666	0.66666	0.66666	0.66666	0.66666	0.66666
Juez 12	0.33333	0.66666	1	0.66666	0.66666	0.33333	0.66666	0.33333
Juez 13	1	0.66666	1	0.66666	0.33333	0.33333	0.33333	0.33333
Promedio	**0.58974**	**0.48717**	**0.84615**	**0.64102**	**0.46153**	**0.43589**	**0.43589**	**0.43589**

Jueces	Texto 5		Texto 6		Texto 7		Texto 8	
	Precisión	***Exhaus.***	***Precisión***	***Exhaus.***	***Precisión***	***Exhaus.***	***Precisión***	***Exhaus.***
Juez 1	0.33333	0.33333	0.33333	0.66666	1	1	1	0.66666
Juez 2	0.33333	0.33333	0.33333	0.33333	0.66666	0.33333	0.66666	0.33333
Juez 3	0.33333	0.33333	0.33333	0.33333	0	0	0.33333	0.33333
Juez 4	0.33333	0.33333	1	0.66666	0	0	1	0.66666
Juez 5	1	0.66666	1	1	0.33333	0.33333	1	0.66666
Juez 6	0.33333	0.33333	0.66666	1	0.33333	0.33333	0.33333	0.33333
Juez 7	0.66666	0.66666	0.66666	0.66666	0.33333	0.33333	0.66666	0.66666
Juez 8	0.66666	0.66666	1	0.66666	0.33333	0.33333	1	0.33333
Juez 9	1	1	0.33333	0.33333	0.33333	0.33333	1	0.66666
Juez 10	0.66666	0.66666	1	1	0	0	0.33333	0.33333
Juez 11	0.66666	0.66666	0.33333	0.33333	0	0	1	0.33333
Juez 12	0.33333	0.66666	0.66666	0.66666	0	0	1	1
Juez 13	0.33333	0.33333	0.66666	0.66666	0.33333	0.33333	1	0.66666
Promedio	**0.53846**	**0.53846**	**0.64102**	**0.64102**	**0.28205**	**0.2564**	**0.79487**	**0.53846**

Jueces	Texto 9		Texto 10		Texto 11	
	Precisión	***Exhaus.***	***Precisión***	***Exhaus.***	***Precisión***	***Exhaus.***
Juez 1	0.33333	0.66666	0.66666	0.66666	1	0.66666
Juez 2	1	0.33333	0.66666	0.33333	0.33333	0.33333
Juez 3	0.33333	0.33333	0.33333	0.33333	0.33333	0.33333
Juez 4	1	0.66666	1	0.66666	1	0.66666
Juez 5	0.66666	0.66666	0.33333	0.66666	1	0.66666
Juez 6	0.66666	0.66666	0.66666	0.33333	0.66666	0.66666
Juez 7	1	1	0.66666	0.66666	0.66666	0.66666
Juez 8	1	0.66666	1	0.66666	0.66666	0.66666
Juez 9	0.66666	0.66666	0.33333	0.33333	1	0.66666
Juez 10	1	0.66666	0.33333	0.33333	0.33333	0.33333
Juez 11	1	0.66666	0.66666	0.66666	0.66666	0.66666
Juez 12	1	0.66666	0.66666	0.66666	0.66666	0.66666
Juez 13	0.66666	0.33333	0.33333	0.33333	0.66666	0.66666
Promedio	**0.79487**	**0.61538**	**0.58974**	**0.51282**	**0.69230**	**0.58974**

Precisión total del experimento: **0.606061**

Exhaustividad total del experimento: **0.517483**

F-score: **0.55828**

- Textos en inglés

Jueces	Texto 1		Texto 2		Texto 3		Texto 4	
	Precisión	***Exhaus.***	***Precisión***	***Exhaus.***	***Precisión***	***Exhaus.***	***Precisión***	***Exhaus.***
Juez 1	1	0.66666	1	1	1	0.66666	1	1
Juez 2	0.33333	0.33333	0.66666	0.66666	1	0.66666	1	0.66666
Juez 3	1	0.33333	0.33333	0.33333	1	0.66666	0.66666	0.66666
Juez 4	0.66666	0.66666	1	1	1	1	1	0.66666
Juez 5	0.33333	0.33333	1	1	0.66666	0.66666	0.66666	0.66666
Promedio	**0.66666**	**0.46666**	**0.8**	**0.8**	**0.93333**	**0.73333**	**0.86666**	**0.73333**

Jueces	Texto 5	
	Precisión	***Exhaus.***
Juez 1	0.33333	0.33333
Juez 2	0.66666	0.66666
Juez 3	0.66666	0.66666
Juez 4	0.33333	0.33333
Juez 5	0.33333	0.66666
Promedio	**0.46666**	**0.53333**

Precisión total del experimento: **0.746667**

Exhaustividad total del experimento: **0.653333**

F-score: **0.696889**

- Textos en alemán

Jueces	**Texto 1**	
	Precisión	***Exhaus.***
Juez 1	1	0.66666
Juez 2	0.66666	1
Promedio	**0.83333**	**0.83333**

Precisión total del experimento: **0.833333**

Exhaustividad total del experimento: **0.833333**

F-score: **0.833333**

- Textos en chuj

Jueces	**Texto 1**	
	Precisión	***Exhaus.***
Juez 1	1.0	1.0
Promedio	**1.0**	**1.0**

Precisión total del experimento: **1.0**

Exhaustividad total del experimento: **1.0**

F-score: **1.0**

- Textos en tarahumara

Jueces	**Texto 1**	
	Precisión	***Exhaus.***
Juez 1	1.0	0.66666
Promedio	**1.0**	**0.66666**

Precisión total del experimento: **1.0**

Exhaustividad total del experimento: **0.666667**

F-score: **0.7999**

REFERENCIAS

[1] Spark Jones K. (1998), "Automatic Summarizing: Factors and Directions", En: Mani I. y Maybury M. (1999), *Advances in Automatic Text Summarization*, Cambridge: MIT Press.

[2] Mani I. y Maybury M. (1999), *Automatic Text Summarization*, Cambridge: MIT Press.

[3] Luhn, H. P. (1958*),* "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development.

[4] Edmundson, H. P. (1969). "New Methods in Automatic Extraction", Journal of the Association for Computing Machinery 16, 264-285.

[5] Brandow, R.; Mitze, K.; Rau, L. (1994). "Automatic condensation of electronic publications by sentence selection". Information Processing and Management 31. 675-685.

[6] Lin. C.; Hovy E. (1997), "Identifying Topics by Position", Applied Natural Language Processing Conference. Washington. 283-290.

[7] Pauce, C. D. (1990), "Constructing literature abstracts by computer: Techniques and prospects", Information Processing and Management 26. 171-186.

[8] Teufel, S.; Moens, M. (1999), "Discourse-level argumentation in scientific articles: human and automatic annotation", ACL Workshop: Towards Standards and Tools for Discourse Tagging. Maryland, USA. 84-93.

[9] Jurafsky Daniel, Martin James H. (2009), *Speech and language processing,* (2ª edición) Pearson Prentice Hall.

[10] Manning Christopher D., Schütze Hinrich (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

[11] Méndez Cruz, Carlos Francisco y Alfonso Medina Urrea (2005), "Extractive Summarization Based on Word Information and Sentence Position", *Lecture Notes in Computer Science*, 3406, pp. 653-656. En: GELBUKH, Alexander, ed., *Computational Linguistics and Intelligent Text Processing*, Springer, Berlín.

[12] Marcu Daniel (1999), "The automatic construction of large-scale corpora for summarization research"*,* The 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley.

[13] Weiss Sholom, Indurkhya Nitin, Zhang Tong, Fred Damerau (2004), *Text Mining: Predictive Methods for Analyzing Unstructured Information,* Springer 1ª edición.

[14] Ross, Sheldon M. (2007), *Introducción a la Estadística*, trad. Valdés Sánchez Teófilo, Ed. Reverte. pp. 96-101

[15] Brown J., Glazier E. (1978), *Telecomunicaciones*, Barcelona. Ed, Marcombo,

[16] Kiss Tibor and Strunk Jan (2006), "Unsupervised Multilingual Sentence Boundary Detection", Computational Linguistics volume 32, pp. 485-525.

[17] Bieler Heike and Dipper Stefanie (2008), "Measures for Term and Sentence Relevances: an Evaluation for German", Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).

[18] *www.python.org*

[19] Bird Steven, Klein Ewan, Loper Edward (2009), *Natural Language Processing with Python*, O'Reilly Media.

[20] Medina Urrea Alfonso (2003), *Investigación cuantitativa de afijos y clíticos del español de México Glutinometría en el Corpus del Español Mexicano Contemporáneo*; Tesis de doctorado, Centro de Estudios Lingüísticos y Literarios, El Colegio de México, ciudad de México.

[21] Mani Inderjeet (2001), "Summarization Evaluation: An Overview", Proceedings of the NTCIR Workshop 2nd meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization.

[22] Sparck Jones, K., Galliers, J. (1996), "Evaluating Natural Language Processing Systems: An Analysis and Review", Lecture Notes in Artificial Intelligence 1083, Springer-Verlag.

[23] Geraldene Walker,Joseph Janes,Carol Tenopir (1999), *Online retrieval: a dialogue of theory and practice*, Second edition ,Libraries Unlimited. pp 264-266

[24] C.J. van Rijsbergen (1975), *Information Retrieval*, Buttersworth, USA.

[25] Brandow, R., K. Mitze, L. Rau (1995), "Automatic condensation of electronic publications by sentence selection", Information processing and management: an international journal, volume 31, issue 5, Pergamon Press.

[26] Barzilay, R., Elhahad, M. (1997), “Using lexical chains for text summarization”, Actas del ACL/EACL Workshop on Intelligent Scalable Text Summarization. Madrid: ACL. 10-17.

[27] Boguraev, B.; Kennedy, C. (1997), “Salience-based content characterization of text documents.”, Actas del ACL/EACL Workshop on Intelligent Scalable Text Summarization. Madrid: ACL. 2-9.

[28] Vivaldi, Jorge; da Cunha, Iria; Torres-Moreno, Juan-Manuel; Velázquez, Patricia (en prensa). “Automatic Summarization Using Terminological and Semantic Resources”, En actas del *7th International Conference on Language Resources and Evaluation* (LREC 2010). Valletta, Malta.

[29] VanDijk, T.A. (1979), “Recalling and summarizing complex discourse”, en W. Burghardt & K. Hölker, eds., *Text Processing* (Berlin New York: de Gruyter).

[30] Correira, A. (1980), “Computing Story Trees”, en *American Journal of Computational Linguistics*, 6, pp. 251-273.

[31] Marcu, D. (2000), *The Theory and Practice of Discourse Parsing Summarization*, Massachusetts: Institute of Technology.

[32] da Cunha, Iria; Wanner, Leo (2005), “Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria”, En Saggion, H.; Minel J. (eds.) *Crossing Barriers in Text Summarization Research* (RANLP-2005). Borovets (Bulgaria): INCOMA Ltd. 46-51. ISBN 954-90906-8-X.

[33] da Cunha, Iria ; Torres-Moreno, Juan Manuel.; Velázquez, Patricia; Vivaldi, Jorge (2009). “Un algoritmo lingüístico-estadístico para resumen automático de textos especializados”, *Linguamática* 2. 67-79. ISSN 1647-0818.

[34] da Cunha, Iria; Fernández, Silvia; Velázquez, Patricia; Vivaldi, Jorge; SanJuan, Eric; Torres-Moreno, Juan Manuel (2007). “A new hybrid summarizer based on Vector Space Model, Statistical Physics and Linguistics”. En Gelbukh, A.; Kuri Morales, A. F. (eds.) MICAI 2007: *Advances in Artifical Intelligence. Lecture Notes in Computer Science*. Berlín: Springer. 872-882. ISSN 0302-9743.

[35] Morris, A. H.,Kasper G., & Adams, D. (1992). “The effects and limitations of automated text condensing on reading comprehension performance”, Information Systems Research, 3(1), 17-35.

[36] Lin, C. Y. (2004), “Rouge: A Package for Automatic Evaluation of Summaries”, En *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*. 25-26.

[37] DUC (The Document Understanding Conference). http://duc.nist.gov

[38] Text Summarization Challenge. http://www.lr.pi.titech.ac.jp/tsc/index-en.html

[39] Hovy, E.; Lin, C. Y.; Zhou, L. (2005), “Evaluating DUC 2005 using basic elements”, En *Proceedings of the Document Understanding Conferences (DUC)*. Vancouver. 1-6 .

[40] Nenkova, A.; Passonneau, R. (2004). “Evaluating content selection in summarization: The pyramid method”. En *Proceedings of the HLT-NAACL Conference*. Boston.145-152.

[41] Torres-Moreno, Juan-Manuel; Saggion, Horacio; da Cunha, Iria; Velázquez-Morales, Patricia; SanJuan, Eric (en prensa). “Évaluation automatique de résumés avec et sans référence”. En actas de la *17e Conférence sur le Traitement Automatique des Langues Naturelles*. Université de Montréal et École Polytechnique de Montréal: Montreal (Canadá).

[42] Annie Louis, Ani Nenkova (2009), “Automatically evaluating content selection in summarization without human models”, *Empirical Methods in Natural Language Processing*, Singapore. 306-314.

[43] Mani, I.; House, D.; Klein, G.; Hirschman, L.; Obrst, L.; Firmin, T.; Chrzanowski, M.; Sundheim, B. (1998), *The Tipster Summac Text Summarization Evaluation: Final report. Technical report*. DARPA.