BENEMÉRITA UNIVERSIDAD AUTÓNOMA
DE PUEBLA
FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

"OBTENCIÓN DE EXTRACTOS DE TEXTOS CON BASE EN UN CORPUS"

TESIS PROFESIONAL

QUE PARA OBTENER EL TITULO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:
HILARIO SALAZAR MARTÍNEZ

ASESOR:
MC. DAVID EDUARDO PINTO AVENDAÑO

Mayo 2004

Dedicatoria

Dedico mi Tesis a mis Padres: Juan Salazar Martínez y Guadalupe Meza Martínez, porque me enseñaron a ver la luz en tiempos de oscuridad, a mis hermanos que con su apoyo económico y moral hice posible la realización de mi trabajo, a mis amigos de la Maestría en Ciencias de la Computación, y en especial a mis maestros y asesores: M.C David Eduardo Pinto Avendaño y Dr. Héctor Jiménez Salazar, por su paciencia para enseñarme y compartir parte de sus conocimientos.

Agradecimientos

Quiero dar las gracias al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo brindado a través del proyecto I39156A, al Dr. Héctor Jiménez Salazar por su valiosa ayuda, y a quienes me brindaron su apoyo para la evaluación de los cuestionarios y como consecuencia la evaluación de mi Tesis.

Índice general

Índice de figuras

Lista de tablas

Introducción

Por generación automática de resumen de un texto se entiende el proceso por el cual se identifica la información sustancial proveniente de una fuente (o varias) para conocer de qué trata un documento sin necesidad de leerlo completamente y producir una versión abreviada destinada a un usuario en particular (o grupo de usuarios) y a una tarea (o tareas) específica.

Existen muchos enfoques y desarrollos para resolver el problema de la generación del resumen automático en algunos enfoques: Se han generado diversas herramientas, algunas de las cuales hacen uso de recursos lingüisticos grandes. Otros más emplean recursos semánticos y otros métodos estadísticos. En general, la mayoría de los métodos se basan en recursos copiosos.

Los volumenes de información cada vez son mayores y con el surgimiento de Internet el manejo de éstos cobró mayor importancia, por lo que se hace necesario la búsqueda de nuevas formas que ayuden a comprender eficazmente el contenido de un documento, sin tener que leer completamente el mismo; además que la información recuperada tenga que ver con el tema de nuestro interés, ya que mucha información devuelta a una consulta tiene poco que ver con el tema de nuestro interés.

En el marco de lo anteriormente expuesto, se presenta este trabajo, el cual consiste en la generación de extractos de textos y esta compuesto en dos partes, primera: usando solo una función de similitud y segunda: utiliza una función de similitud e información mutua, ambas propuestas hacen uso un corpus [10, 5].

Los resultados hasta el momento son alentadores, y por lo cual hay mucho trabajo por delante.

Este trabajo se divide en dos partes y varios capítulos:

El capítulo 1 muestra algunos métodos para la generación de resumen automático, así como la evaluación de dichos algoritmos.

Parte I Generación de Extracto Utilizando solo una Función de Similitud.

El capítulo 2 describe paso a paso nuestra propuesta para la generación de extractos.

El capítulo 3 describe las pruebas realizadas, así como el procedimiento de evaluación de la parte I.

El capítulo 4 menciona las conclusiones obtenidas, así como el trabajo futuro a realizar.

Parte II Generación de Extractos Utilizando Información Mutua y Función de Similitud.

El capítulo 5 describe paso a paso el proceso de extracción de extracto de documentos utilizando información mutua.

El capítulo 6 describe las pruebas realizadas, así como el procedimiento de evaluación de la parte II.

Además tiene varios apéndices donde se ejemplifica mejor el proceso de extracción.

Capítulo 1

Marco Teórico

1.1. Métodos para la Generación del Resumen Automático

Con el continuo crecimiento de Internet y el volúmen de textos disponibles, es cada vez más importante proveer un mecanismo para permitir una rápida y eficaz recuperación de información contenida en el web, difundir esa información, filtrar la información pertinente entre toda aquella contenida en los documentos almacenados, entre otras tareas.

La generación de resumen automático constituye uno de los grandes temas del tratamiento automático del lenguaje natural. Las técnicas estadísticas, basadas en la frecuencia de ocurrencia y coocurrencia de términos han sido ampliamente utilizadas. Se han proporcionado diversos enfoques de resumen y extracto de documentos [1, 2, 3, 6, 9, 11, 12, 13, 17, 18, 21] para resolver este problema .

El resumen de un texto es una versión abreviada del contenido de ese texto, mientras que el extracto es el conjunto de oraciones del texto con las que se puede construir el resumen. Así, puede decirse, que en particular, que en el extracto de un texto se entiende aquellas oraciones y palabras que son extraidas y que por lo tanto están contenidas en el texto; por el contrario, el resumen incluye nuevas palabras para darle coherencia y congruencia al mismo.

En este trabajo se realizan experimentos usando el concepto de extracto de un

documento, se muestran algunos métodos que ejemplifiquen el procedimiento de extracción.

1.1.1. Construcción Automática de Corpus Grandes para Resúmenes Automáticos

Para la obtención de resumen automático existe dificultades entre ella la carencia de un adecuado corpus: hoy existen únicamente unas pequeñas colecciónes de algunos textos que han sido unidos y anotados manualmente por su importancia textual. Se tiene el costoso y tedioso proceso de anotación. Es muy tedioso que se anote manualmente la importancia textual en corpus de textos grandes. Para evitar este problema, Daniel Marcu [2] desarrolló un algoritmo que construye un corpus automáticamente .
El algoritmo toma un conjunto de tuplas como entrada (Resumen, Texto) y genera el extracto correspondiente, es decir, el conjunto de cláusulas (oraciones) en el texto utilizado para escribir el correspondiente resumen. Para el desarrollo del algoritmo se realizó un experimento que fue evaluado por jueces. El experimento también sugiere estrategias de extracción para mejorar el desarrollo del sistema de resumen automático.
La ventaja de esta propuesta es que crea recopilaciones grandes, específicamente genera un corpus (Resumen, Extracto, Texto) por tuplas dobles:

1. Primero, los pares (Extracto, Texto) se pueden utilizar para entrenar y evaluar la generación del sistema de resumen. Es decir, la construcción de la extracción responsable de identificar la parte más importante del texto.

2. En segundo lugar, los pares (Resumen, Extracto) pueden ser usado para enseñar y evaluar a los generadores de sistemas de resumen, es decir, la interpretación y generación responsable para el mapeo de la selección de unidades textuales dentro del texto coherente.

Un Algoritmo de Aproximación para Determinar el Extracto

La suposición fundamental de aproximación de la construción (resumen, extracto, texto) en tuplas de (resumen, texto) tuplas que estan en el correspondiente extracto, por el subconjunto de cláusulas en el texto cuya similaridad semántica con el resumen es máxima. En caso general, se tiene un texto *(T)* de n cláusulas, ellas son $C_n^1 + C_n^2 + \ldots + C_n^n = 2^n - 1$ extractos de longitud diferente de cero.
La idea dominante del enfoque que determina el extracto E_M es la siguiente: En vez de contestar a la pregunta "se incluye esta cláusula en el extracto?", se contesta una pregunta complementaria: "si quita esta cláusula del texto, se puede aún escribir el extracto A?". Contestando a la pregunta complementaria: hay una manera clara de determinar el número de las cláusulas que deben ser incluidas en el extracto. Para entender la razón, asumase que se asigna inicialmente como E_M todas las cláusulas del texto *(T)*. También asuma que se utiliza un coseno simple, métrico - normalizado, basada en similitud. Si se representa E_M y el extracto *(A)* a través de secuencias $(t, w(t))$ de pares, donde, dado un símbolo t, $w(t)$ es su peso, es posible calcular las semejanzas entre E_M y el extracto *(A)*, del extracto usando la fórmula (1.1), donde el $w(t)E_M$ de $w(t)$A representa los pesos de los simbolos en el extracto A y el E_M del extracto respectivamente.

$$sim(E_M, A) = \frac{\sum_{t \in E_M \cup A} w(t)_{E_M} w(t)_A}{\sqrt{\sum_{t \in E_M} w(t)^2{}_{E_M} \sum_{t \in A} w(t)^2{}_A}} \tag{1.1}$$

Algoritmo

ENTRADA: A tuplas (Resumen, Texto) **SALIDA:** A tuplas (Resumen, Extracto, Texto)

1. Partir el resumen y el texto en cláusulas
2. Efectuar "stemming' y eliminar las palabras cerradas en ambos conjuntos de cláusulas //Construye el centro del extracto

3. E_M = Cláusulas de (texto);
4. **While**$((E = E_M \backslash C_i \mid C_i \in E_M \wedge (\forall C_j \in E_M)(i \neq j \longrightarrow sim(E_M \backslash C_i, Resumen) \geq sim(E_M \backslash C_j, Resumen))) \wedge E > E_M)$
5. $E_M = E$;
6. **end While**
7. Eliminar en E_M las cláusulas que tienen un estado retórico de satélite débil;
8. Eliminar en E_M las cláusulas cortas que tiene un estado retórico de satélite fuerte;
9. Eliminar en E_M los subtítulos;
10. Eliminar en E_M las cláusulas que no son similares a cada cláusula en el resumen;
11. Agregar a E_M las cláusulas en el texto que son más similares a cada clausula en el resumen;
12. Agregar a E_M las cláusulas únicas en el texto que contienen como mínimo dos palabras en el resumen que no son usadas en alguna otra cláusula;
13. Elimina en E_M todas las cláusulas cortas redundantes;
14. **return** $Extracto = E_M$

Los pesos $w(t)E_M$ *y* $w(t)A$ son cancelados por la frecuencia del término en el extracto y el resumen respectivamente. Si se elimina en E_M la cláusula C_i de la que está totalmente sin relación en el extracto *(A)*, se obtiene un nuevo extracto $(E_M \backslash C_i)$ el cual es similar a A y es mejor que el de E_M ya que $sim(E_M\{C_i\}, A)$ y la $sim(E_M, A)$ tienen igual numerador pero el denominador del $sim(E_M \backslash C_i, A)$ es más pequeño que el denominador de $sim(E_M, A)$ y ésta es la suma de menos términos. Al aplica un enfoque ambicioso y se elimina en varias ocasiones las cláusulas de E_M de modo que en cada paso el extracto que resulta tenga semejanza máxima con el resumen, se llega eventual a un estado donde no puede eliminar cláusula de la cancelación sin disminuir la semejanza

de E_M con el resumen. Si se considera que E_M del texto que caracteríza esta etapa es el extracto buscando. Este enfoque codicioso para determinar una componente de las tuplas del resumen (Resumen, Extracto, Texto) a partir de los componentes del extracto y del texto, representa la base del algoritmo.

En el primer paso, se utiliza acotamiento breve de cláusulas y un algoritmo para la identificación de marcado del discurso (CB-DM-1) para determinar las unidades textuales elementales y las frases que desempeñan el papel del discurso del resumen y el texto. El algoritmo CB-CD-DM-I codifica explícitamente conocimiento del papel que juegan con esas 450 frases claves, es decir, frases tales como *sin embargo, además de y aunque,* señala límites de la cláusula y las relaciones retóricas que celebran entre las oraciones adyacentes del texto.

El algoritmo voraz determina un E_M del extracto de semejanza máxima con el resumen suprimiendo cláusulas en el texto original, de modo que la semejanza del extracto restante con el extracto sea máxima en cada paso. Obviamente, tal enfoque no garantiza que **el algoritmo descrito anteriormente** se acerque hacia el extracto de semejanza máxima. Pero sin embargo, se ha notado que incluso un algoritmo tan simple produce un buen extracto, observando la semejanza entre el E_M de los extractos de interés, es decir, los extractos de los cuales las cláusulas repetidas son suprimidas, y el extracto correspondiente para los diez textos del experimento.

Evaluación del algoritmo

Para evaluar el funcionamiento del algoritmo de extracción. Primero se utilizan jueces humanos, para determinar los umbrales del funcionamiento humano en la tarea de identificar los resúmenes que fueron utilizados y escribir los resúmenes para cada juez, determinaron que el sistema de unidades de cada texto que consideraron una mayoría de los otros 13 jueces para ser similar con las unidades en el extracto correspondiente. Por lo tanto se determinó para cada texto "el extracto del patrón oro" en el cual una mayoría de los 13 jueces convino y después se comparó el extracto producido por el juez catorce, con

la evocación tradicional empleanda, comparando la precisión que reflejan los porcentajes de las unidades textuales seleccionadas correctamente por un juez con respecto al patrón oro y a los porcentajes del número total de unidades seleccionadas respectivamente.

Además, los autores calcularón la evocación y los resultados de precisión que caracterizaron los extractos construidos automáticamente, por el algoritmo.

Para permitir una comparación mejor del enfoque manual y automático, además de la evocacion y de la precisión.

1.1.2. ContextO: Una Plataforma para la Extracción de Información y el Resumen Automático de Textos

Este trabajo hace una reseña en cuanto diversos modelos que utilizan conocimiento lingüistico. El modelo de Marcu (Mar 97) se inspira en la teoría de estructura retórica (Man 88) y en el análisis de conectores para construir áboles retóricos que jerarquizan la importancia de los diferentes segmentos textuales. Lehman (Leh 95) selecciona fragmentos del texto sobre la base de puntajes calculados para cada frase en función de términos preestablecidos. Berri (Ber 96) atribuye etiquetas semánticas a ciertas frases para decidir su inclusión o no en un resumen de texto. Mansson (Mas 98) reconoce parcialmente estructuras temáticas en un texto, mientras que Ellouze (Ell 98) explota diferentes tipos de objetos textuales para producir esquemas de resumen.
La evaluación realizada sobre ciertos sistemas (Min 97), (Jin 98) así como los trabajos llevados a cabo en colaboración con profesionales de la producción de resúmenes (End 93) o en comparación con los resúmenes producidos por estos profesionales (Sag 98) ha mostrado la dificultad para la realización de resúmenes estándar, es decir construidos sin tomar en cuenta las necesidades de los usuarios.
Este trabajo propone la definición de un modelo conceptual general de representación para conocimientos lingüisticos [9] y el desarrollo de tareas especializadas que cooperen entre si. Este método identifica los conocimientos lingüisticos ubicandolos en sus contextos y organizándolos en tareas especializadas. Presenta por un lado la ventaja de permitir que el trabajo del lingüista se realice de manera independiente de su implementación informática, y, por otro lado, la de articular efectivamente una misma arquitectura informática los dos tipos de trabajo.
Este enfoque no está orientado a la producción de resúmenes estándar, sino más bien a la construcción de un sistema automático de filtrado de información en textos, con ayuda de criterios semánticos que dan a un usuario la posibilidad de definir su perfil de filtrado en función de sus objetivos. Este sistema emplea conocimientos lingüisticos generales, es decir independientes del dominio particular. Estos conocimientos, que forman una

base de conocimientos ligüisticos, se expresan en formas de marcadores discursivos explícitos (morfemás, palabras, expresiones) que caracterizan una intención pragmática del autor del texto, intención que el sistema debe ser capaz de identificar y de interpretar semánticamente en función del contexto, como lo puede hacer un lector o un profesional del resumen que no sea experto en el dominio tratado en el texto. El sistema de filtrado se basa por un lado en el método de exploración contextual el cual constituye su justificación teórica, y por otro lado en la plataforma informática ContextO que implementa y explota la base de conocimientos lingüisticos necesarios para el filtrado semántico.

El Papel de los Conocimientos Lingüisticos

En este enfoque, el trabajo previo del lingüista consiste en estudiar sitemáticamente un corpus de textos buscando regularidades léxicas y discursivas cuyo empleo es representativo de la categoría sistemática considerada.
Los lingüistas identifican entonces en el corpus las marcas gramaticales pertinentes para la resolución de un problema, y luego conciben y escriben las reglas de exploración del contexto para esas marcas identificadas en los textos.

El Método de Exploración Contextual (MEC)

Este método fue desarrollado por el equipo LaLIC (UMR 8557 du CNRS, EHESS, Université Paris-Sorbonne) que dirige el Profr. Jean-Pierre Desclés. El objetivo de este método consiste en proveer el marco necesario para identificar información semántica específica contenida en los textos. El MEC parte de la hipótesis que establece que todo texto posee unidades lingüisticas que permiten resolver indeterminaciones semánticas en algunos casos, y tomar ciertas decisiones para construir sentidos, en otros.
El método se implementa informáticamente bajo la forma de bases de conocimientos lingüisticos. Este sistema emplea conocimiento exclusivamente lingüistico y presentes en el texto. El método requiere de una descripción fina

y detallada de ciertas unidades lingüisticas llamadas **indicadores** y de otras llamadas **índices**, complementarias de las primeras.
Los **indicadores** son expresiones lingüisticas que disparan la ejecución de ciertas reglas de exploración contextual encargadas de determinar el *valor semántico* del indicador para ciertas tareas específicas, por ejemplo reconocer una conclusión o filtrar una definición del texto. De manera que los indicadores que están asociados a ciertas tareas, son específicos de cada tarea.
Por otra parte, cada indicador tiene asociado un conjunto de reglas de exploración contextual, heurística unas y lingüisticas otras. Las aplicación de una regla, invocada por un indicador, explora el contexto de ese indicador buscando índices lingüisticos con el objetivo de resolver la tarea, esto es, determinar el valor semántico del indicador.
Todos estos elementos - indicadores, índices y reglas - componen la base de conocimiento lingüistico que el método emplea para realizar las actividades requeridas.

Arquitectura General

Sus principales componentes son: el motor de exploración contextual, los agentes especializados y la interfaz con el usuario.

- **El Motor de Exploración** dispara, para una o varias tareas especializadas, el proceso de reconocimiento de indicadores e índices presentes en un segmento textual.

- **El Analizador de Texto** tiene por objetivo construir una primera representación que refleje la estructura del texto. Se basa en un texto señalado por un segmentador construido a partir de un estudio sistemático de las marcas de puntuación. La segmentación se basa en estas marcas y en un estudio de los contextos a la derecha y a la izquierda de las mismas. Aplica además reglas heurísticas para reconocer las secciones, con sus títulos, los párrafos y las oraciones.

- **Los Agentes Especializados** tienen por objetivo explorar las "decoraciones

semánticas"del texto en función de las necesidades definidas por el usuario. Hay entonces un agente que construye un resumen compuesto de oraciones del texto de entrada que corresponde a un tipo de perfil y un agente que construye diferentes extractos del texto de entrada en función de perfiles seleccionados por el usuario.
Estos agentes especializados permiten desarrollar tratamientos específicos para un usuario, explotando el modelo genérico de tratamiento de conocimientos lingüisticos.

- **Interfaz con el Usuario** define la funcionalidad general de la plataforma. Se trabaja sobre la idea de proyectos. Para cada proyecto el usuario define, básicamente, los documentos a procesar, los distintos tipos de tratamientos deseados, los perfiles de cada tratamiento y las salidas deseadas. Permite guardar los proyectos y visualizar de manera jerárquica, acorde a la jerarquía textual definida, o de manera plana. Permite navegar con facilidad entre el texto original y el texto procesado, sincronizándolos dinámicamente.

1.1.3. Generación Automática de Resúmenes Personalizados

Bajo el concepto de resumen automático se agrupan diferentes tipos de resumen que pueden clasificarse atendiendo a su propósito, enfoque y alcance [12]. Atendiendo al **alcance** el resumen puede limitarse a un único documento o a un conjunto de ellos que traten sobre el mismo tema.

1. Según su **propósito** esto es, atendiendo al uso o tarea al que están destinados, a los resúmenes se clasifican como:

   - **Indicativos**, si el objetivo es anticipar al lector el contenido del texto y ayudarle a decidir sobre la relevancia del documento,
   - **Informativos**, se pretenden sustituir al texto completo incorporando toda la información nueva o trascendente, y
   - **Críticos**, se incorporan opiniones o comentarios que no aparecen en el texto original.

2. Finalmente, atendiendo al **enfoque**, se puede distinguir entre resúmenes

    - **Genéricos**, si recogen los temas principales del documento y van destinados a un grupo amplio de personas, y
    - **Adaptados al Usuario**, si el resumen se confecciona de acuerdo a los intereses (es decir conocimientos previos, ámbitos de interés o necesidades de información) del lector o grupo de lectores al que va dirigido.

Este método se centra en una fase de análisis en el que se identifican los segmentos del texto (frases o párrafos, normalmente) que contiene la información más significativa. Durante esta fase se aplica un conjunto de heurísticas a cada una de las unidades de extracción. El grado de significancia de cada una de ellas puede obtenerse mediante combinación lineal de los pesos resultantes de la aplicación de dichas heurísticas. Éstas pueden ser *posicionales*, se tiene en cuenta la posición que ocupa cada segmento dentro del documento, *lingüisticas*, se buscan ciertos patrones de expresiones indicadas, o *estadísticas*, se incluyen frecuencias de apariciones de ciertas palabras.

El resumen resulta de concatenar dichos segmentos de texto en el orden en que aparecen en los textos originales, los posibles problemas de insconsistencia en el resumen resultante constituyen el principal inconveniente de esta aproximación. Una forma de resolver este problema es utilizar el párrafo en lugar de la frase como unidad de extracción.

Generación de Resúmenes Personalizados

Este sistema utiliza tres heurísticas de selección de frases. Las dos primeras se utilizan para para la obtención de resúmenes generales mientras que la tercera se refiere a la personalización de los mismo. Para generar los resúmenes se utiliza la combinación ponderada de las distintas heurísticas.
Las tres heurísticas tienen un objetivo común y es dar puntuaciones a cada una de las frases del texto objeto del resumen, para más tarde elegir las más relevantes. A continuación se describe más detalladamente cada una de ellas.

- **Heurísticas Posicionales:** Esta heurística consiste básicamente en dar mayor puntuación a las cinco primeras frases de un texto.
  En dominios periodísticos, el título y las primeras frases de un texto dan una idea aproximada al lector del contenido de texto que van a leer a continuación. Por esta razón se asigna los siguientes pesos a las N primeras frases del documento que se esta resumiendo. Un valor típico de N se maneja como 5 y los pesos asignados podrían ser:
  Peso frase 1: 1.0
  Peso frase 2: 0.99
  Peso frase 3: 0.98
  Peso frase 4: 0.95
  Peso frase 5: 0.90

- **Heurísticas de Palabras Clave:** Cada texto tiene un número de palabras claves, que son bastante representativas de su contenido. Esta heurística consiste en extraer las M palabras más significativas de cada texto y comprobar a continuación, cuantas de esas palabras clave se encuentran en cada frase. De esta forma se asigna mayor peso a las frases que contengan mayor número de palabras clave.
  Para obtener las M palabras más relevantes de cada noticia se indexan todas las noticias obteniendo así el peso de cada palabra en cada documento utilizando el método $tf.idf$. Se seleccionan así las ocho palabras que más peso tengan para cada documento.
  Para obtener el peso de la frase en el documento se divide el número de

palabras clave del documento que aparece en la frase por el número total de palabras de la frase.

- **Heurísticas de Personalización:** El objetivo de esta heurística consiste en potenciar aquellas frases que tengan mayor relevancia para un modelo de usuario dado, con el fin de personalizar el resumen. En lugar de obtener una idea general del texto resumido, se orienta la elección de frases de tal forma que se elijan aquellas que tengan mayor similitud con las preferencias del usuario.

  El cálculo de los pesos para las frases se realiza de la siguiente manera: del modelo de información con respecto a los pesos que el usuario ha asignado a sus categorías y a sus términos personales. También se extraen del modelo los términos que representan cada categoría así como los términos que el usuario haya definido. Con toda esta información se calcula la similitud existente entre el modelo y la frase, asignado un peso a la frase de acuerdo con la siguiente función de similitud:

$$PesoFrase = \frac{pCat \cdot sim(frase, Categoria) + pTer \cdot sim(Frase, Terms)}{pCat + pTerm} \tag{1.2}$$

  Donde:

$$sim(frase, Categoria) = \frac{\sum_{i=1}^{i=n} sim(frase, tc_i) \cdot pc_i}{\sum_{i=1}^{i=n} pc_i} \tag{1.3}$$

$$sim(frase, Terms) = \frac{\sum_{i=1}^{i=n} sim(frase, t_i) \cdot pt_i}{\sum_{i=1}^{i=n} pt_i} \tag{1.4}$$

  Siendo, $pCat$ el peso general de las categorías, $pTerms$ el peso general de los términos, tc_i los términos que identifican a las categorías i, pc_i el peso asignado a la categoría i, t_i, el término i y el pt_i asignado al término i.

- **Combinación de las tres heurísticas anteriores:** se utiliza para obtener un solo peso para cada frase utilizando la siguiente ecuación:

$$PesoTotalFrase = \frac{(\alpha \cdot pesoH) + (\beta \cdot pesoH2) + (\gamma \cdot pesoH3)}{\alpha + \beta + \gamma} \tag{1.5}$$

Los parámetros α, β y γ sirven para dar más importancia a una heurística sobre otra.

Los cálculos de similitud que realiza el sistema se basan en el modelo de espacio vectorial (Salton 1989), utilizado para la representación de textos, vectores de pesos de términos. Para su obtención se eliminan las palabras más frecuentes usando una lista de parada estándar y los restantes se reducen a una forma canónica usando el extractor de raíces de Porter (Porter 130-137) adaptado al español.

Evaluación

La métrica de evaluación más popular en el marco de recuperación de información es la Evocación/Precisión. Esta métrica permite comparar númerica y gráficamente (por medio de una gráfica precisión/evocación) distintos enfoques de recuperación. Aquí se calculó la precisión media y la gráfica evocación/precisión de los siguientes enfoques:

- La utilización de las noticias completas en la recuperación.
- La utilización de los Resúmenes generales en la recuperación, con los parámetros $\alpha = 0.5$ y $\beta = 0.5$ (es decir se da igual importancia a las dos heurísticas generales).
- La utilización de los Resúmenes personalizados en la recuperación con los parámetros $\alpha = 0.5,\ \beta = 0.5\ y\ \gamma = 1, 2, \ldots, 5$ (es decir, la heurística de personalización es igual de importante que las dos heurísticas generales, o el doble, etc.).

En general, cuanto más próxima se encuentra la precisión media o la gráfica obtenida para un método de extracción de resúmenes a la obtenida usando todo el texto de las noticias, la calidad es mayor porque se retiene más información.

1.2. Conceptos Relacionados

1.2.1. Relaciones de Sentido

Relación de sentido [15] de una palabra x, es el conjunto de todas aquellas palabras relacionadas con x de alguna manera. En este caso, por relación se entiende al hecho de que éstas pertenecen al mismo contexto.
Se concibe una relación de sentido como la aproximación de sus primeros términos de asociación. Estos términos han sido usados con resultados prometedores. Algunas aproximaciones han sido llevadas a cabo para aproximar una representación semántica del texto. Aunque algunas de ellas podrían resultar costosas, debido principalmente al uso de grandes recursos linguisticos.

En nuestro caso, usamos las sentencias más representativas de un documento para representar cada documentos y además usamos relaciones de sentido para expandir consultas, con la finalidad de mejorar la precisión y evocación cuando las palabras que pertenecen a la consulta no aparecen en ningún documento a ser buscado.

Nuestra propuesta esta basada en la teoría de conceptos formales, en la cual cada concepto se representa por dos partes, el 'intent'(es decir el conjunto de propiedades del concepto), y el 'extent'(el conjunto de instancias que tiene cada propiedad). Más formalmente, Sea G el conjunto de instancias, M un conjunto de propiedades y ψ una correspondencia de G a M. El conjunto $\{m \in M \mid \psi(g) = m\}$ está denotado por $\Psi(g), g, \in G$. Un concepto en $\langle G, M, \psi \rangle$ es un par (A, B), donde $A \subset G, B \subset M$, y que mantiene que:

$$\bigcap_{a \in A} \Psi(a) = B$$

Existen algunas aplicaciones de conceptos formales para la representación textual con respecto a los Sistemas de Recuperación de Información y la aproximación para algunas relaciones lexicas basadas en el orden definido entre conceptos.

1.2.2. El Sentido de un Sintagma

El sentido de una frase se conforma agregando las relaciones que dan sentido a por lo menos dos términos del sintagma ($\Psi(t_1) \cap \Psi(t_2) \neq 0$.

Sea la frase $x_1 x_2 \ldots x_n y$, su sentido (burdo, sin composición estricta) está dado por:
$\xi(x_1, \ldots x_n, y) = \xi(x_1, \ldots, x_n) \cup \bigcup_{i \leq n} \Upsilon(x_i, y)$,

siempre que
$\Psi(y) \cap \xi(x_1, \ldots x_n) \neq 0$, donde $\xi(x_1, x_2) \doteq \Upsilon(x_1, x_2) \doteq \Psi(x_1) \cap \Psi(x_2)$

1.2.3. Información Mutua

La información mutua (IM) [16] es básicamente un concepto teórico de información y que finalmente se traduce a una medida. La IM es entonces un cociente de asociación, para medir las normas de asociación de palabras. La medida propuesta es más objetiva y menos costosa que el método subjetivo empleado en Palermo y Jenkins [19]. El cociente de la asociación se puede escalar hasta proporcionar estimaciones robustas de las normas de asociación de palabras para porciones más grandes del lenguaje. Usando la medida del cociente de la asociación, podría obtenerse que las cinco palabras más asociadas dentro de un texto podrían ser: dentistas, enfermeras, tratamiento, y hospitales.

De acuerdo a Fano [4], si dos puntos (palabras), x y y, tienen probabilidades $P(x)$ y $P(y)$, entonces su información mutua, $I(x, y)$, se define como:

$$I(x, y) \equiv log_2 \frac{P(x, y)}{P(x)P(y)} \tag{1.6}$$

Informalmente, la información mutua compara la probabilidad de observar x y y de manera conjunta (la probabilidad de unión) con las probabilidades de observar x y y independientemente. Si hay una asociación genuina entre x y y, entonces la probabilidad común $P(x, y)$ será mucho más grande que $P(x)P(y)$, y por lo tanto $I(x, y) \gg 0$. Si no existe una relación de interés

entre x y y, entonces $P(x,y) \approx P(x)P(y)$, y entonces, $I(x,y) \approx 0$. Si x y y están en distribución complementaria, entonces $P(x,y)$ será mucho menor que $P(x)P(y)$, forzando que $I(x,y) \ll 0$.

En nuestro caso, las probabilidades $P(x)$ de la palabra y $P(y)$ son estimadas contando la frecuencia de ocurrencia de x y de y en un texto. Las probabilidades comunes, $P(x,y)$, son estimadas contando el número de veces que x es seguido por y en una ventanas de w palabras.
Posteriormente se aplica la función normalizando entre el tamaño del corpus (N).

El parámetro del tamaño de las ventanas permite que se observe a través de diversas escalas. Un tamaño más pequeño de las ventanas identificará expresiones fijas y otras relaciones que se mantengan a través de tramos cortos. Ventanas con una tamaño más grande, resaltarán conceptos semánticos y otras relaciones que se matienen a través de grandes escalas.

Se ha observado que el cociente de la asociación llega a ser inestable cuando los valores de conteo de frecuencia son muy pequeños, por lo que regularmente se usa un umbral de 5 para la frecuencia de ocurrencia conjunta, i.e., $f(x,y) < 5$. Esta aproximación es totalmente arbitraria, sin embargo, en la práctica se toma como un valor útil.
En nuestro trabajo, hemos tomado en cuenta el hecho de que los documentos analizados son heterogéneos en cuanto al tamaño, por lo que se ha decidido normalizar la función para el cálculo de información mutua. También se ha detectado ciertas posibilidades de que la frecuencia de ocurrencia conjunta de términos sea cero. En este último caso, la fórmula planteado con anterioridad arrojaría un valor de infinito. La fórmula, finalmente ha quedado expresada como sigue:

$$IM(x,y) = log_2 \left(\frac{N \cdot fr(x,y)}{fr(x) \cdot fr(y)} + 1 \right) \tag{1.7}$$

Donde $fr(x)$ y $fr(y)$ son la frecuencia de ocurrencia del término x y y, respectivamente, y $fr(x, y)$ es la frecuencia de ocurrencia conjunta de los términos x, y.

Parte I

Generación de Extracto Utilizando un Corpus y una Función de Similitud

Capítulo 2

Generación del Extracto de un Texto

La idea en que se basa el sistema implementado es que una oración O_i de un texto T es más representativa que otra O_j si la similitud entre O_i y T es mayor que la similitud entre O_j y T. Con esta idea se propuso llevar a cabo las operaciones que a continuación serán descritas.

Este trabajo consiste en la obtención de extractos de documentos para lo cual se hace uso de un corpus C de 96 documentos de diversos dominios (justicia, cultura, política, sociedad, gobierno, ciencia, tecnología, religión y economía).

Se considera un texto T del cual se desea obtener su extracto. Con el fin de comparar las similitudes entre cada oración de T y el texto mismo, se debe hacer una representación de T usando C. Al corpus C y al texto T se les aplica truncamiento y se obtiene C', T', para dar paso a la obtención del vocabulario V' de T'. Ya obtenido V' se procede a representar V'' de acuerdo con el corpus C'. Enseguida, de acuerdo con V'' se obtiene a grandes razgos la representación de T. La representación de T:T''. En este último, T'', se procede a ordenar las oraciones para obtener el extracto de T utilizando una función de similitud. En resumen, el sistema toma como entrada un texto T del cual se extraen las oraciones más representativa usando un corpus, lo cual se ilustra en la siguiente figura 2

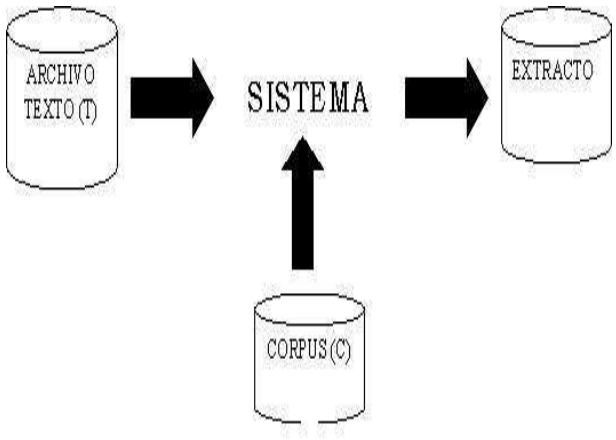

Figura 2.1: Forma gráfica del sistema

A continuación se describen cada uno de los pasos mencionados anteriormente:

1. Preproceso de T para obtener T'
2. Obtención de V a partir de T'
3. Representación de V usando $C' : V'$
4. Representación de las oraciones T' usando V'
5. Cálculo de la similitud entre cada oración de T'' y su complemento en T''
6. Ordenamiento de las oraciones de T según la similitud encontrada
7. Generación de extracto usando un umbral de las 5 oraciones más representativas

Donde:

C'	Corpus preprocesado
T	Es el texto plano
T'	Texto lematizado y truncado
V	Vocabulario de T'
V'	Vocabulario representado
T''	Texto representado utilizando V'

El corpus utilizado en las pruebas realizadas consta de 96 documentos tomados de 9 dominios. La tabla 2.1 muestra los porcentajes de cada dominio a los cuales contribuyen los 96 documentos pueden observarse que "Política" y "Sociedad" tienen los porcentajes mayores, y el dominio de Tecnología es el de menor porcentaje.

Tabla 2.1: Porcentaje de los dominios

Dominio	No. de textos por dominio	Porcentaje
Justicia	10	10.40
Cultura	7	7.3
Política	25	26
Sociedad	23	24
Gobierno	9	9.4
Ciencia	11	11.5
Tecnología	1	1
Religión	2	2.1
Economía	8	8.3
Total	96	100

A continuación se describe el procedimiento para la obtención de los extractos de los textos:

2.1. Preprocesamiento

El preprocesamiento se efectúa sobre el texto T y el Corpus C. El objetivo es tener de cada texto identificadas las oraciones que lo componen con sus palabras truncadas. Se realizan las siguientes operaciones:

1. Conversión a minusculas
2. Identificación de signos (tokens)
3. Eliminación de palabras cerradas y puntuación (excepto puntos que separan oraciones)
4. Stemming
5. Segmentación de oraciones

Este proceso se representa en la figura siguiente:

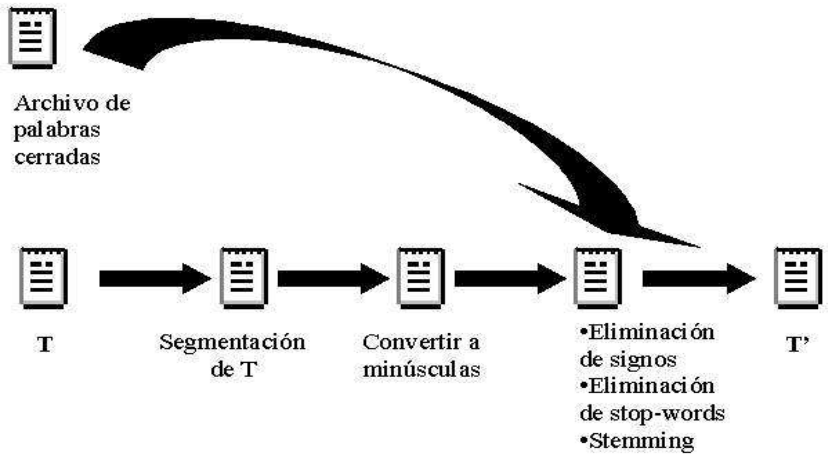

Figura 2.2: Gráfica detallada del proceso de preprocesamiento

El siguiente ejemplo muestra el efecto del procedimiento anterior

Sea T el texto:

> "El Banco Mundial desembolsará este año 1.5 mil millones de dólares en préstamos para México. La mayoría de estos financiamientos se enfocarán a mejorar las condiciones sociales de los pobres, reforzar la estabilidad macroeconómica afianza las reformas en el ejercicio del poder público."

Al identificar tokens en T queda:

> El Banco Mundial desembolsará este año 1.5 mil millones de dólares en préstamos para México . La mayoría de estos financiamientos se enfocarán a mejorar las condiciones sociales de los pobres, reforzar la estabilidad macroeconómica y afianza las reformas en el ejercicio del poder público .

Con lo cual se obtiene el texto sin palabras cerradas:

> Banco Mundial desembolsará año 1.5 mil millones dólares préstamos México. mayoría financiamientos enfocarán mejorar condiciones sociales pobres, reforzar estabilidad macroeconómica afianza reformas ejercicio poder público .

Al truncar las palabras comunes se tiene:

> Banco Mundial desembols año 1.5 mil millon dólar prést México . mayorí financi enfoc mejor condicion social pobr reforz estabil macroeconóm afianz reform ejercici pod públic .

Se forman oraciones en T :

> 1) Banco Mundial desembols año 1.5 mil millon dólar prést México
>
> 2) mayorí financi enfoc mejor condicion social pobr reforz estabil macroeconóm afianz reform ejercici pod públic

Al texto obtenido se le llama $T^{'}$. Se debe de mencionar que el anterior procedimiento tambien se aplica al corpus C, obteniendo $C^{'}$ (mas secuencia de oraciones con palabras truncadas y sin palabras cerradas)

2.2. Obtención del Vocabulario

En este paso se obtiene el vocabulario V del archivo $T^{'}$: Del ejemplo $T^{'}$ del Apéndice A anterior, algunos tokens son:

V ={1.5,afianz,año,Banco,condicion,desembols,dólar,ejercici,. . . ,reforz,social}

2.3. Representación del Vocabulario

En este paso se realizó el siguiente procedimiento

1. ENTRADA $(V, C^{'})$, SALIDA $(V^{'})$

2. Para cada $x \epsilon V^{'}$

   a) Añadir a la representación de $x, x^{'}$: cada contexto (oración) de $C^{'}$ donde ocurra x, en el cuadro 2.2 se muestra la cantidad de tokens asignadas a algunos tokens de V después de efectuar la representación

Tabla 2.2: Tokens que forman $V^{'}$

Tokens	1.5	afianz	año	Banco	condición	desembols	dólar	reforz	social
No. de tokens	13	2	218	20	44	3	36	7	82

2.4. Representación del Texto Preprocesado

De acuerdo con el paso anterior se realiza la representación del archivo $T^{'}$ de la forma siguiente:

1. ENTRADA $(V^{'}, T^{'})$, SALIDA $(T^{''})$
2. Repite mientras no sea el último renglón(i) del archivo $T^{'}$
   - *a*) Repetir mientras no sea la ultima palabra (j) del renglón (i)
     - 1) Sustituir la palabra $x^{'}_{ji}$ por su vector de representación tomado de $V^{'}$ para formar $T^{''}$

Un fragmento de la representación se ilustra en el Apéndice B.

2.5. Obtención de la Similitud y Ordenamiento según el Puntaje

En este módulo se obtiene la similitud utilizando la fórmula (2.1) que una simplificaciín de la funcióin de Jaccard (5.2):

$$sim(O^{''}_i, T^{''} - O^{''}_i) = \#(O^{''}_i \cap (T^{''} - O^{''}_i)) \tag{2.1}$$

Donde:

O_i	es la oración de T
$O^{''}_i$	es la oración de $T^{''}$
$T'' - O^{''}_i$	es el complemento de O_i en $T^{''}$

y se realizó utilizando el siguiente procedimiento:

1. ENTRADA: $(T^{''})$, SALIDA ($T^{''}$ ordenado de acuerdo a 2.1)
2. Repite mientras no sea la última oración de $T^{''}$
   - *a*) Tomar $O^{''}_i$ de $T^{''}$
   - *b*) Obtener complemento de O_i respecto $T^{''} : O^{''}_i$

c) Calcular la similitud entre O_i y O_i'': S_i

d) Asignarle S_i a O_i

e) i++

3. Ordenar T respecto a S_i

En el Apéndice C se muestra un ejemplo de este paso.

2.6. Obtención del Extracto

En este módulo se obtiene el extracto del documento T del Apéndice C, utilizando un umbral de cinco para obtener las oraciones con mayor puntaje, graficamente todo este proceso se muestra en la siguiente figura:

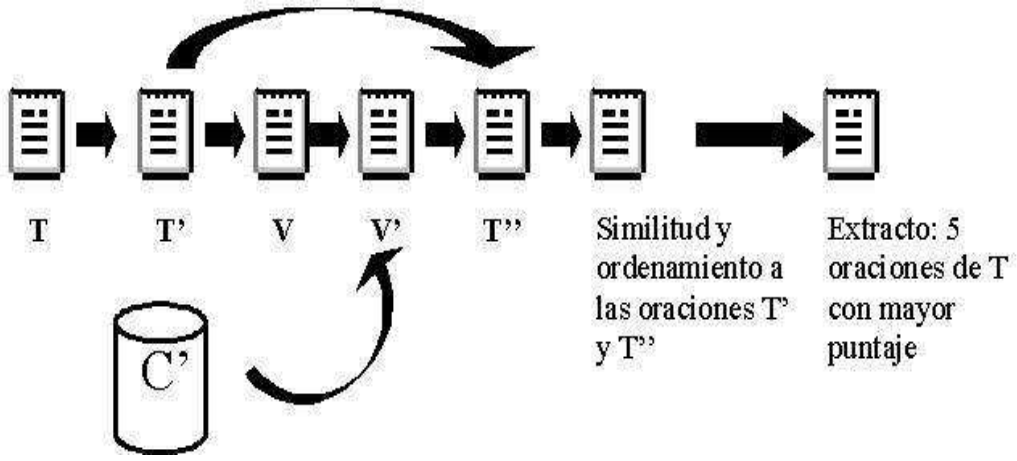

Figura 2.3: Proceso gráfico para la generación del extracto

Para este paso existe un ejemplo en el Apéndice D inciso 2.

En el siguiente capítulo se mostrará las pruebas realizadas y la evaluación de los resultados.

Capítulo 3

Pruebas y Evaluación

Las pruebas consistieron en tomar dos muestras (textos) de cada dominio (cultura, economía, política, sociedad y justícia). De los textos se eliminaron palabras cerradas, se hizo truncamiento y se segmentaron por oraciones, finalmente se obtuvo el extracto de cada texto de acuerdo con el capítulo 2.

Para su evaluación se solicitó la ayuda de cuatro jueces a los cuales se les entregó el paquete de diez documentos (Texto original (T_i) y el extracto obtenido (Ex_i) por el sistema de cada documento) solicitandoles que en T_i marcaran las cinco oraciones más representativas; señalando con A las más representativas sucesivamente hasta con E la menos representativa. Asimismo se solicitó que marcaran el orden de importacia en el que ubicaban las cinco oraciones de Ex_i.

Posteriormente, los cuestionarios constestados por los jueces fueron preprocesados. Se formaron tres clases de oraciones marcadas : **MR** (Muy Representativa), **R** (Representativa) y **SR** (Suficientemente representativa). Las primeras incluyen las oraciones que tiene marca A o B, R las de marca C o D, y SR aquellas con la marca E.

Para cada oración marcada por el sistema : si su marca está en la misma clase que la del juez su puntuación es uno. En cualquier otro caso cero (**a este criterio se le llama duro (CD)**). A el **criterio suave (CS)** es el mismo que el duro y además, se agregan las siguientes reglas, si la marca de una oración esta entre clases cercanas se le asigna $\frac{2}{3}$ (MR y R o R y SR); y si la oración

esta en las clase extremas se le asigna $\frac{1}{3}$ (MR y SR).
Para precisar cómo fue llevada a cabo la evaluación se definen algunas funciones de puntuación.

Donde:
k representa un documento
s refiere al sistema
cl es una clase
j representa un juez
i representa una oración del texto
q representa un dominio
La función de puntuación para el **CD** es:

$$CD_j(k,i) = \begin{cases} & (cl_j(k,i) = MR \wedge cl_s(k,i) = MR) \vee \\ 1 \ si & (cl_j(k,i) = SR \wedge cl_s(k,i) = SR) \vee \\ & (cl_j(k,i) = R \wedge cl_s(k,i) = R). \\ 0 & \text{en otro caso} \end{cases} \tag{3.1}$$

A la función de puntuación del **CD**, se agregan otras reglas, obteniendo asi la función de puntuación para el **CS**, quedando está de la siguiente manera:

$$CS_j(k,i) = \begin{cases} 1 & \text{como en } CD_j(k,i) \\ & \\ & cl_j(k,i) \neq cl_s(k,i) \wedge \\ \frac{1}{3} \ si & (cl_j(k,i) = SR \wedge cl_s(k,i) = MR) \vee \\ & (cl_j(k,i) = MR \wedge cl_s(k,i) = SR) \\ & \\ & cl_j(k,i) \neq cl_s(k,i) \wedge \\ \frac{2}{3} \ si & [(cl_j(k,i),\ cl_s(k,i) \in \{MR,\ R\}) \vee \\ & (cl_j(k,i),\ cl_s(k,i) \in \{R,\ SR\})]. \\ & \\ 0 & \text{en otro caso} \end{cases} \tag{3.2}$$

La exactitud por respuesta para CD o CS, se obtuvo con las siguientes funciones

$$ER_{CD}(k,i) = \frac{1}{4}\sum_{j=1}^{4} CD(k,i) \tag{3.3}$$

$$ER_{CS}(k,i) = \frac{1}{4}\sum_{j=1}^{4} CS(k,i) \tag{3.4}$$

La puntuación promedio **(ED)** del documento con CD se obtuvo con la función

$$ED_{ER_{CD}(k,i)}(k) = \frac{1}{5}\sum_{i=1}^{5} ER_{CD}(k,i) \tag{3.5}$$

Y por ultimo, para el caso del CS se utilizó:

$$ED_{ER_{CS}(k,i)}(k) = \frac{1}{5}\sum_{i=1}^{5} ER_{CS}(k,i) \tag{3.6}$$

Después de haber aplicado las funciones descritas anteriormente, los resultados obtenidos por cada texto se muestran en las tablas: 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9 y 3.10.

Tabla 3.1: Resultados del texto j26020_7E62

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	1	0.25	$2\frac{1}{3}$	$\frac{3}{5}$
2	3	0.75	$3\frac{2}{3}$	1
3	0	0	0	0
4	1	0.25	$1\frac{1}{3}$	$\frac{1}{3}$
5	1	0.25	0	0
ED		0.3		0.37

Tabla 3.2: Resultados del texto j20120_7E82

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	1	0.25	$1\frac{1}{3}$	$\frac{1}{3}$
2	1	0.25	1	$\frac{1}{4}$
3	1	0.25	1	$\frac{1}{4}$
4	1	0.25	1	$\frac{1}{4}$
5	1	0.25	1	$\frac{1}{4}$
ED		0.25		0.27

Tabla 3.3: Resultados del texto j31010_7E22

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	1	0.25	1	$\frac{1}{4}$
2	1	0.25	2	$\frac{1}{2}$
3	0	0	0	0
4	1	0	$\frac{2}{3}$	$\frac{1}{6}$
5	1	0.25	$1\frac{2}{3}$	$\frac{3}{7}$
ED		0.15		0.27

Tabla 3.4: Resultados del texto j16020_7E42

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	1	0.25	1	$\frac{1}{4}$
2	0	0	$\frac{1}{3}$	0
3	2	0.5	2	$\frac{1}{2}$
4	0	0	0	0
5	1	0.25	1	$\frac{1}{4}$
ED		0.2		0.22

Tabla 3.5: Resultados del texto j20120_7E22

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	2	0.5	$2\frac{2}{3}$	$\frac{2}{3}$
2	2	0.5	$2\frac{2}{3}$	$\frac{2}{3}$
3	1	0.25	0	0
4	1	0.25	$2\frac{2}{3}$	$\frac{2}{3}$
5	1	0.25	$1\frac{2}{3}$	$\frac{3}{7}$
ED		0.35		0.48

Tabla 3.6: Resultados del texto j14030_7E62

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	3	0.75	3	$\frac{3}{4}$
2	2	0.5	2	$\frac{1}{2}$
3	2	0.5	2	$\frac{1}{2}$
4	0	0	$1\frac{1}{3}$	$\frac{1}{3}$
5	2	0.5	2	$\frac{1}{2}$
ED		0.45		0.52

Tabla 3.7: Resultados del texto j08110_7E32

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	2	0.5	$2\frac{2}{3}$	$\frac{2}{3}$
2	0	0	$1\frac{1}{3}$	$\frac{1}{3}$
3	1	0.25	$1\frac{2}{3}$	$\frac{3}{7}$
4	0	0	$1\frac{1}{3}$	$\frac{1}{3}$
5	3	0.75	$3\frac{2}{3}$	1
ED		0.3		0.53

Tabla 3.8: Resultados del texto j06020_7E32

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	1	0.25	0	0
2	0	0	2	$\frac{1}{2}$
3	1	0.25	1	$\frac{1}{4}$
4	0	0	0	0
5	4	1	4	1
ED		0.3		0.35

Tabla 3.9: Resultados del texto j06020_7E52

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	0	0	$1\frac{1}{3}$	$\frac{1}{3}$
2	0	0	0	0
3	0	0	0	0
4	1	0.25	1	$\frac{1}{4}$
5	0	0	$\frac{2}{3}$	$\frac{1}{6}$
ED		0.05		0.12

Tabla 3.10: Resultados del texto j30010_7E42

# oración del extracto	CD	ER_{CD}	CS	ER_{CS}
1	4	1	4	1
2	3	0.75	3	0.75
3	1	0.25	$\frac{5}{3}$	0.42
4	2	0.5	2	0.50
5	1	0.25	1	0.25
ED		0.55		0.58

Como se ha visto, el corpus utilizado no es de un dominio en particular, así como los textos usados en las pruebas. La heterogeneidad del dominio produce un sesgo, como es de esperarse en los resultados. Esto se muestra en la tabla 3.11, es decir, cuanto más corpus se tenga de un dominio mejor será el extracto producido. En la tabla 3.11 se muestra la **exactitud del extracto obtenido por dominio (EED)** del texto procesado. Ya que se tenían dos documentos de prueba por dominio se suma su ER_{CD} y se divide entre dos, igualmente para ER_{CS}.

Expresando todo lo anterior en forma de función para CD sería:

$$EED = \frac{1}{\sharp(doc_{q_k \in dom_q})} ED_{ER_{CD}(k,i)}(k) \tag{3.7}$$

Para CS:

$$EED = \frac{1}{\sharp(doc_{q_k \in dom_q})} ED_{ER_{CS}(k,i)}(k) \tag{3.8}$$

Lo cual aparece en la segunda y tercera columna de la tabla 3.11.

Tabla 3.11: Exactitud de respuesta por dominio

Dominio	EED_{CD}	EED_{CS}
Cultura	0.27	0.4
Economía	0.3	0.36
Justicia	0.1	0.21
Política	**0.375**	**0.4**
Sociedad	**0.4**	**0.5**
Promedio	**0.289**	**0.374**

La tabla 3.12 muestra las características de los diez textos usados para las pruebas después de cada paso descrito en el capítulo 2.

Tabla 3.12: Características de los textos usados en el experimento

ID	Tamaño (KB)	Dominio	# de palabras	# de oraciones	Tamaño de V de T'	# de raices T''	# palabras sin truncar
j26020_7E62	4.4	Economía	673	25	234	104515	389
j20120_7E82	4.7	Cultura	614	25	260	10936	412
j31010_7E22	9.3	Justicia	1379	50	460	10568	696
j16020_7E42	8.4	Política	1223	49	370	10350	589
j20120_7E22	3.8	Sociedad	537	19	206	10848	328
j14030_7E62	4.5	Sociedad	646	25	259	10341	400
j08110_7E32	4.5	Cultura	644	25	234	10897	363
j06020_7E32	2.5	Economía	347	16	160	9296	229
j06020_7E52	5.2	Justicia	731	26	131	10971	376
j30010_7E42	6.1	Política	802	21	323	10276	490
Promedio	5.34		760	28	264	1900	428

Capítulo 4

Conclusiones (Parte I)

Se ha presentado una herramienta para obtener las oraciones más representativas de un texto. Estas oraciones pueden constituir lo que es llamado un extracto del texto. El método empleado, a diferencia de otros, usa solamente un corpus sin más preprocesamiento que la eliminación de palabras cerradas y truncamiento de los restantes. Este es un método relativamente eficiente ($O(n^2)$) con n número de oraciones del texto.

Para evaluar la herramienta construida se procedió a obtener el extracto de diez textos tomados al azar dentro de cinco dominios (Cultura, Economía, Justicia, Política y Sociedad) utilizando un corpus de 347 KB. Los textos de pruebas fueron entregados a cuatro jueces para que marcaran las cinco oraciones más representativas del texto. Finalmente, se obtuvo un índice de precisión del sistema comparando las marcas de los jueces y del sistema en las oraciones de cada texto. Globalmente, la mayor precisión fue 0.5 para textos del dominio "Sociedad"usando el **criterio suave**. Ciertamente, este dominio es el segundo más representativo en el corpus. Por ello llama la atención que el segundo dominio más representativo ("Política") haya tenido una precisión inferior. Este fenómeno puede deberse a la composición de cada colección de textos por dominio.

Los resultados obtenidos en esta parte del trabajo son preliminares y han sido presentados en [14], Estos solamente indican que conviene experimentar más con el método utilizado para analizar sus fortalezas. Es necesario comparar el método con otros métodos y así determinar sus ventajas y desventajas. Por

ahora, se debe resaltar que su principal ventaja es que usa pocos recursos lingüisticos y es relativamente eficiente. Asimismo, deben hacerse pruebas aumentando el corpus (como hasta ahora heterogéneo), y además realizar pruebas aplicándolo a textos de un sólo dominio con un corpus de ese mismo dominio.

Por último, conviene experimentar con otras formas de selección de contextos, por ejemplo usar ventanas en lugar de oraciones.

Parte II

Generación de Extractos Utilizando un Corpus, Información Mutua y Función de Similitud

Capítulo 5

Generación del Extracto utilizando Información Mutua

Este parte del trabajo se centra en la idea de que el título de un texto da una mejor aproximación al lector acerca del contenido del mismo, y por lo tanto, éste debería de tener un mayor peso [12] y valiendose de una función de similitud se buscan las oraciones O_i que tengan mayor similitud con el título. Con la idea anterior se propuso realizar el siguiente trabajo.

Considérese un texto T del cual se desea conocer su extracto, para lo cual se hace uso de un corpus C (en este caso del dominio de política). El corpus C y T deben ser preprocesados previamente, para obtener T_1 y C_1; en el caso de T_1 se separa el proceso en dos partes: primero se construye una tabla con todos los encabezados (T_e) de los textos a obtener su extracto y como segunda parte, con el resto de T se forma otro archivo con el cuerpo del texto (T_c). Como paso siguiente se procede a obtener el vocabulario de T_c y T_e que serán V_{Tc} y V_{Te} respectivamente. Los vocabularios son representados de acuerdo con C_1 y se procede a obtener la frecuencia de cada índice con respecto a C_1. A continuación se busca la información mutua de las combinaciones de índices (V_{xy}) de V_{Tc} con respecto a C_a, siguiendo con la eliminación de las palabras que no cumplen con un umbral en la representación del índice ; Así quedan representados $T_c y T_c$ con T_{e1} y T_{c1} según V_{xy}, finalmente, para obtener el extracto T se encuentra la similitud entre el encabezado correspondiente a T_{c2} con cada oración $O_i \in T_{c2}$. El extracto es formado entonces por las cinco oraciones de mayor puntaje que haya obtenido la función de similitud (5.2).

A continuación se describen los pasos que componen el procedimiento anteriormente descrito:

1. Preproceso de T y C para obtener T_1 y C_1
2. T_1 se divide en dos partes: T_c y T_e
3. Obtención de $V_{Tc} : T_c$
4. Obtener la frecuencia (fr_x) de cada índice de V_{tc} con respecto a C_1
5. Obtener la frecuencia ($Vfr_{(k,h)}$) de las combinaciones de índices de V_{tc} con respecto a C_1 y cálculo de la Información Mutúa (IM)
6. Representación de V_{tc} usando IM
7. Representación de las oraciones T_c usando $Vfrc_{(k,h)}$ y obtenemos T_{c2}
8. Cálculo de la similitud entre el encabezado correspondiente (E_{Tc2}) de T_c con $O_i \in T_{c2}$
9. Ordenamiento de las oraciones de T_c según la similitud encontrada
10. Generación de extracto usando un umbral de cinco, para las oraciones más representativas

Donde:

T	Texto plano
C_1	Corpus preprocesado
T_1	Texto preprocesado
T_e	Tabla de encabezados
T_c	Texto sin encabezado
V_{Te}	Vocabulario de T_e y el número de sentencias que la contienen en C_1
V_{Tc}	Vocabulario de T_c y el número de sentencias que la contienen en C_1
$Vfre_{(x,y)}$	Vocabulario por parejas de T_e y el nuúmero de sentencias que la contienen en C_1
$Vfrc_{(k,h)}$	Vocabulario por parejas de T_c y el número de sentencias que la contienen en C_1
V_{repe}	Vocabulario representado de acuerdo a $Vfre_{(x,y)}$
V_{repc}	Vocabulario representado de acuerdo a $Vfrc_{(k,h)}$
T_{e2}	Tabla de encabezados expandida usando V_{repe}
T_{c2}	Cuerpo del texto expandido usando V_{repc}
E_{Tc2}	Encabezado correspondiente a T_{c2}

El corpus utilizado en las pruebas realizadas consta de 375 textos, todos del dominio de **"Política"**, dicho corpus tiene un tamaño de 2700 kb, con 21559 raíces y 29741 oraciones. Cabe mencionar que la muestra de textos tomada (40) para obtener sus extractos, no esta contenida en el corpus.

En la tabla 5.1, se muestran las características de los 40 textos usados para las pruebas al aplicar cada paso descrito en este capítulo.

Tabla 5.1: Característica de los textos usados

ID	Tamaño (K)	# de palabras V_{xy}	# de oraciones	Tamaño de V de T' (KB)	# de raices T''
071203_101co2	2849	11477	12	161	153
Sigue ...					

Tabla 5.1: Característica de los textos usados (Continuación)

ID	Tamaño (K)	# de palabras V_{xy}	# de oraciones	Tamaño de V de T'(KB)	# de raices T''
071203_102co2	8909	63547	35	920	920
071203_103co2	25257	458404	164	6593	959
071203_104co2	5188	24091	17	335	221
071203_105co2	2852	8516	11	199	132
071203_106co2	4609	27262	75	357	235
071203_107co2	4191	27967	22	384	238
071203_108co2	4104	18916	22	260	196
071203_109co2	4041	23872	17	325	220
071203_110co2	7813	64981	67	863	362
071203_111co2	4905	28204	20	404	239
071203_113co2	4944	21529	32	285	209
071203_114co2	17573	286904	132	3992	759
071203_115co2	32008	790654	182	10945	1259
071203_116co2	19548	477754	153	6587	979
071203_117co2	3660	16111	27	215	181
071203_118co2	8299	45452	16	647	303
071203_119co2	5112	40756	26	562	287
071203_121co2	23566	362527	96	5317	853
071203_122co2	1456	5357	11	76	105
071203_123co2	35807	519691	149	7516	1021
071203_124co2	3534	17579	20	250	189
071203_125co2	4388	24754	18	363	224
071203_126co2	28885	492042	187	8283	1089
071203_127co2	5792	19504	40	270	199
071203_128co2	41873	1435666	286	20011	1696
071203_130co2	5684	49142	38	733	315
071203_131co2	7311	54286	28	793	331

Sigue . . .

Tabla 5.1: Característica de los textos usados (Continuación)

ID	Tamaño (K)	# de palabras V_{xy}	# de oraciones	Tamaño de V de T'(KB)	# de raices T''
071203_132co2	12713	204481	110	2832	641
071203_133co2	5340	40187	30	560	285
071203_134co2	5879	44851	30	635	301
071203_135co2	3681	18337	14	266	193
071203_136co2	4668	33154	12	476	259
071203_138co2	6918	58312	23	865	343
071203_139co2	13548	174937	113	2392	593
071203_140co2	3775	15401	14	195	177
071203_142co2	5116	40756	30	2524	287
071203_143co2	11040	180301	100	988	602
071203_144co2	7735	68636	33	1896	372
071203_145co2	10864	129287	50	48	510
Promedio	**16595**	**77640**	**100**	**2964**	**584**
Fin de la tabla					

A continuación se detallan los pasos del procedimiento para la obtención de los extractos.

5.1. Preprocesamiento

El preprocesamiento se efectúa sobre T y C de acuerdo con la sección 2.1. El objetivo es obtener de cada texto las oraciones que lo componen, cada una de ellas con sus palabras truncadas.

El siguiente ejemplo muestra el efecto del procedimiento anterior

Sea T el texto:

> " a este respecto un informe del gobierno distrito federal indica que el 30 de junio de 1989 se celebró un convenio."

Después de identificar tokens en T, se obtiene:

> " a este respecto un informe del gobierno distrito federal indica que el 30 de junio de 1989 se celebró un convenio."

Con lo cual se obtiene además el texto sin palabras cerradas:

> respecto informe gobierno distrito federal indica l 30 junio 1989 celebró convenio."

Al truncar las palabras comunes se tiene:

> respect inform gobierno distrito federal indic 30 juni 1989 celebr conveni

Se segmentan oraciones en T :

> 1) a este respecto un informe del gobierno distrito federal indica que el 30 de junio de 1989 se celebró un convenio
>
> 2) de acuerdo con documentación que analiza la comisión especial de la asamblea legislativa que investiga el caso del paraje san juan aún faltan de ser liberadas dos mil 283 escrituras que significan el 25 por ciento del total de viviendas que se regularizaron con la expropiación de 1989

Al texto obtenido se le llama T_1. El procedimiento anterior también se aplica al corpus C, obteniendo C_1.

Los pasos 2, y 3 son calculados de la misma forma que los pasos 3 y 4 del capítulo 2, respectivamente.

5.2. Obtención de la Información Mutua (IM)

Para obtener la información mutua se busca primero el vocabulario simple y la frecuencia de cada índice $fr(x)$ con respecto a C_1; en seguida se busca

la frecuencia de la combinación por pares de términos $fr(x, y)$. Después de obtener esta información se procede a calcular la información mutua ($IM(x, y)$) utilizando la siguiente fórmula:

$$IM(x, y) = log_2 \left(\frac{N \cdot fr(x, y)}{fr(x) \cdot fr(y)} + 1 \right) \tag{5.1}$$

Donde N es el número de oraciones de C_1.

ejemplo:

V_{tc}={1094/mism, 711/cas,651/años,542/general,486/objet, . . . ,360/ limit}

$Vfr_{(k,h)}$= {42/(mism,nacional),39/(asamblea,general),31/(traves,nacional), . . . , 18/(gobierno,nacional)}

5.3. Representación de V_{Te} y V_{Tc} usando IM

Utilizando la ecuación (5.1) se procede a eliminar las palabras donde su información mutua encontrada es menor que un umbral de tres. Es decir si $(x, y) \in V_{Tc}$ y $IM(x, y) \geq 3$.

El procedimiento es el siguiente:

1. ENTRADA: ($fr(x, y)$ y V_{Tc}), SALIDA(V_{repc})
2. Repite mientras no sea la última palabra de V_{Tc}
   - *a*) Buscar la fr_x y fr_y en V_{Tc}
   - *b*) Obtener N
   - *c*) Calcular la IM de acuerdo a la ecuación (5.1)
3. se construye V_{repc}

Un ejemplo de esta representación se encuentra en el Apéndice G

5.4. Representación de T_e y T_c

En relación a los pasos anteriores se procede a representar T_c utilizando el siguiente procedimiento:

1. ENTRADA (V_{repc}, T_c), SALIDA (T_{c2})
2. Repite mientras no sea la última oración $O_i \in T_c$
   - *a)* Repite mientras no sea la última palabra $x_j^{'}$ de la O_i
     1) Sustituir la palabra $x_{ji}^{'}$ por su vector de representación tomado de V_{repc} para formar T_{c2}

5.5. Obtención de la Similitud y Ordenamiento según el *Puntaje*

Para el caso de esta segunda parte, la función de similitud utilizada es el coeficiente de Jaccard (5.2) y se aplica al encabezado correspondiente al texto E_{Te2} y T_{c2} y dicha función se expresa como sigue:

$$sim(E_{Te2}, T_c) = \#\frac{(E_{Te2} \cap Oi_{c2})}{(E_{Te2} \cup Oi_{c2})} \tag{5.2}$$

Donde:

Oi_{c2} es la $Oi_{c2} \in T_{c2}$

E_{Te2} es el $E_{Tc2} \in T_{e2}$

$(E_{Te2} \cup Oi_{c2}) = \#palabras\ de\ Oi_{c2} + \#palabras\ de\ E_{Te2} - \#(E_{Te2} \cap Oi_{c2})$

y se realizó utilizando el siguiente procedimiento:

1. ENTRADA: (T_c, E_{Te2}, y T_{c2}), SALIDA (T_c ordenado de acuerdo a (5.2))
2. Repite mientras no sea la última oración $O_i \in T_c$
   - *a*) Tomar E_{Te2}
   - *b*) Obtener O_j respecto T_{c2}
   - *c*) Calcular la similitud usando la ecuación 5.2 entre O_j y E_{Te2} : Si
   - *d*) Asignarle S_i a O_i
   - *e*) i++
3. Ordenar T_c respecto a S_i

En el Apéndice E se muestra un ejemplo de este paso.

5.6. Obtención del Extracto

Después de aplicar el algoritmo anterior, se procede a realizar un ordenamiento descendente, de acuerdo con el valor de similitud encontrado. Se hace uso de un umbral con valor igual a cinco para obtener las cinco oraciones más representivas. Estas oraciones conformarán el extracto (Ext) de T_c.

Para entender mejor este paso, se ha anexado un ejemplo en el Apéndice F.

En el siguiente capítulo se mostrarán las pruebas, asi como los criterios de evaluación.

Capítulo 6

Pruebas y Evaluación

El procedimiento de evaluación para la oración (Val_{ora}) de T_c consiste en comparar $O_j \in T_c$ y $O_i \in Ext$. Si $O_j = O_i$ entonces se obtiene el cociente de la diferencia de j e i en valor absoluto más uno; lo cual claramente se muestra en la siguiente función:

$$Val_{ora} = \frac{1}{|j - i| + 1} \tag{6.1}$$

La evaluación del extracto del documento se obtiene sumando todos los valores (Val_{ora}) y dividiéndolos entre el número de oraciones de *Ext*, quedando la función siguiente:

$$Val_{doc} = \frac{\Sigma_{j=1}^{5} Val_{ora}}{5} \tag{6.2}$$

donde: 5 es el número de oraciones de Ext

En las siguientes tablas se muestran los resultados obtenidos:

Tabla 6.1: Evaluación usando la función de similitud (2.1)

ID de T_c	# orac.	Val_{ora} *de Ext*					suma	puntaje de T
		O_1	O_2	O_3	O_4	O_5		
071203pol_101co2	12	0.5	0.5	0.166	0.166	0.5	1.833	0.366
071203pol_102co2	35	1	0.031	0.166	0.5	0.5	2.197	0.439
071203pol_103co2	164	0.1	0.009	0.028	0.009	0.037	0.184	0.036
071203pol_104co2	17	0.125	0.125	0.125	0.25	0.333	0.958	0.191
Sigue ...								

Tabla 6.1: Evaluación usando la función de similitud (2.1) (Continuación)

ID de T_c	# orac.	Val_{ora} *de Ext*					suma	puntaje de T
		O_1	O_2	O_3	O_4	O_5		
071203pol_105co2	11	0.142	0.125	0.5	0.333	0.25	1.351	0.270
071203pol_106co2	75	0.076	0.5	0.5	0.045	0.1	1.222	0.244
071203pol_107co2	22	0.090	0.2	1	0.142	0.333	1.767	0.353
071203pol_108co2	22	0.2	0.066	0.166	0.25	0.5	1.183	0.236
071203pol_109co2	17	0.1	0.125	0.1	0.2	0.111	0.636	0.127
071203pol_110co2	67	0.142	0.2	0.019	0.111	0.03	0.503	0.100
071203pol_111co2	20	0.076	0.066	0.5	0.5	0.5	1.643	0.328
071203pol_113co2	32	0.035	0.166	0.058	0.142	0.125	0.529	0.105
071203pol_114co2	132	1	0.014	0.045	0.017	0.021	1.098	0.219
071203pol_115co2	182	0.006	0.014	0.03	0.007	0.031	0.088	0.017
071203pol_116co2	153	0.013	0.2	0.04	1	0.012	1.265	0.253
071203pol_117co2	27	0.125	0.05	0.111	0.333	0.055	0.675	0.135
071203pol_118co2	16	0.25	0.5	0.076	0.166	0.142	1.136	0.227
071203pol_119co2	26	0.111	0.058	0.076	0.5	0.083	0.830	0.166
071203pol_121co2	96	0.037	0.058	0.111	0.09	0.045	0.343	0.068
071203pol_122co2	11	0.111	0.111	0.166	0.333	0.333	1.055	0.211
071203pol_123co2	149	0.011	0.027	0.011	0.010	0.007	0.068	0.013
071203pol_124co2	20	0.25	0.071	0.090	0.142	0.142	0.698	0.139
071203pol_125co2	18	0.142	0.2	0.083	0.076	0.5	1.003	0.200
071203pol_126co2	187	0.02	0.005	0.007	0.007	0.111	0.15	0.03
071203pol_127co2	40	0.055	0.071	0.038	0.058	0.066	0.290	0.058
071203pol_128co2	286	0.004	0.003	0.166	0.007	0.025	0.206	0.041
071203pol_130co2	38	0.1	0.083	0.333	0.04	0.032	0.588	0.117
071203pol_131co2	28	1	0.076	0.142	0.111	1	2.329	0.465
071203pol_132co2	110	0.02	0.02	0.02	0.013	0.009	0.082	0.016
071203pol_133co2	30	0.038	0.037	0.2	0.5	0.111	0.886	0.177

Sigue ...

Tabla 6.1: Evaluación usando la función de similitud (2.1) (Continuación)

ID de T_c	# orac.	Val_{ora} *de Ext*					suma	puntaje de T
		O_1	O_2	O_3	O_4	O_5		
071203pol_134co2	35	0.035	0.04	0.052	0.09	0.25	0.469	0.093
071203pol_135co2	14	1	0.1	0.333	1	0.25	2.683	0.536
071203pol_136co2	14	0.166	0.076	0.5	0.166	1	1.910	0.382
071203pol_138co2	23	0.142	0.25	0.062	0.166	0.09	0.712	0.142
071203pol_139co2	113	0.062	0.009	0.02	0.041	0.04	0.173	0.034
071203pol_140co2	14	0.090	0.25	0.125	0.25	0.111	0.827	0.165
071203pol_142co2	30	0.090	0.052	0.066	0.5	0.071	0.781	0.156
071203pol_143co2	104	0.014	0.5	0.025	0.013	0.013	0.567	0.113
071203pol_144co2	33	0.083	0.034	0.5	0.125	0.043	0.786	0.157
071203pol_145co2	56	0.029	0.022	0.125	0.052	0.025	0.254	0.05
Promedios	**61.97**	**0.190**	**0.126**	**0.172**	**0.211**	**0.198**	**0.913**	**0.179**
Fin de la tabla								

Tabla 6.2: Evaluación usando la función de similitud (5.2)

ID de T_c	# orac.	Val_{ora} *de Ext*					suma	puntaje de T
		O_1	O_2	O_3	O_4	O_5		
071203pol_101co2	12	0.25	0.125	0.5	0.142	0.25	1.267	0.253
071203pol_102co2	35	0.055	0.030	0.033	0.5	0.047	0.666	0.133
071203pol_103co2	164	0.038	0.008	0.012	0.008	0.009	0.076	0.015
071203pol_104co2	17	0.333	0.071	0.333	0.076	0.1	0.915	0.183
071203pol_105co2	11	0.5	0.25	0.2	0.2	0.2	1.35	0.27
071203pol_106co2	75	0.020	0.037	0.047	0.024	0.025	0.155	0.031
071203pol_107co2	22	0.333	0.125	0.25	0.2	0.090	0.999	0.199
071203pol_108co2	22	0.090	0.111	0.333	0.066	0.333	0.935	0.187
071203pol_109co2	17	1	0.071	0.1	0.166	0.166	1.504	0.300
Sigue . . .								

Tabla 6.2: Evaluación usando la función de similitud (5.2) (Continuación)

ID de T_c	# orac.	Val_{ora} *de Ext*					suma	puntaje de T
		O_1	O_2	O_3	O_4	O_5		
071203pol_110co2	67	0.5	0.016	0.021	0.016	0.022	0.577	0.115
071203pol_111co2	20	0.062	0.142	0.058	0.25	0.142	0.657	0.131
071203pol_113co2	32	0.166	0.090	0.333	0.035	0.066	0.693	0.138
071203pol_114co2	132	0.010	0.015	0.011	0.012	0.02	0.069	0.013
071203pol_115co2	182	0.024	0.018	0.076	0.008	0.5	0.626	0.125
071203pol_116co2	153	0.007	0.01	0.01	0.047	0.012	0.086	0.017
071203pol_117co2	27	0.111	0.041	0.25	0.142	0.055	0.601	0.120
071203pol_118co2	16	0.142	1	0.111	0.083	0.1	1.437	0.287
071203pol_119co2	26	0.166	0.111	0.333	0.111	0.25	0.972	0.194
071203pol_121co2	96	0.041	0.013	0.05	0.037	0.045	0.187	0.037
071203pol_122co2	11	1	1	0.142	0.25	0.166	2.559	0.511
071203pol_123co2	149	0.006	0.014	0.018	0.333	0.017	0.391	0.078
071203pol_124co2	20	0.2	0.1	0.090	0.111	0.5	1.002	0.200
071203pol_125co2	18	0.058	0.25	0.076	0.111	0.2	0.696	0.139
071203pol_126co2	187	0.009	0.04	0.062	0.006	0.125	0.242	0.048
071203pol_127co2	40	0.041	0.125	0.045	0.333	0.035	0.581	0.116
071203pol_128co2	286	0.012	0.007	0.008	0.004	0.008	0.040	0.008
071203pol_130co2	38	0.142	0.052	0.1	0.071	0.2	0.566	0.113
071203pol_131co2	28	0.045	0.066	0.052	0.25	0.071	0.484	0.096
071203pol_132co2	110	0.25	0.09	0.027	0.01	0.034	0.411	0.082
071203pol_133co2	30	1	0.047	0.166	0.125	0.076	1.416	0.283
071203pol_134co2	35	0.076	0.076	0.166	0.333	0.071	0.725	0.145
071203pol_135co2	14	0.1	1	0.111	0.333	0.25	1.794	0.358
071203pol_136co2	14	0.333	1	0.333	0.25	1	2.916	0.583
071203pol_138co2	23	1	0.071	0.090	0.090	0.25	1.503	0.300
071203pol_139co2	113	0.013	0.019	0.041	0.009	0.052	0.136	0.027
Sigue ...								

Tabla 6.2: Evaluación usando la función de similitud (5.2) (Continuación)

ID de T_c	# orac.	Val_{ora} *de Ext*					suma	puntaje de T
		O_1	O_2	O_3	O_4	O_5		
071203pol_140co2	14	0.142	0.083	0.333	0.125	0.166	0.851	0.170
071203pol_142co2	30	0.083	0.5	0.083	0.142	1	1.809	0.361
071203pol_143co2	104	0.026	0.25	0.011	0.010	0.023	0.322	0.064
071203pol_144co2	33	1	0.25	1	0.25	0.25	2.75	0.55
071203pol_145co2	56	0.027	0.052	0.022	0.041	0.021	0.166	0.033
Promedios	**61.97**	**0.235**	**0.184**	**0.151**	**0.133**	**0.173**	**1.308**	**0.175**
Fin de la tabla								

Capítulo 7

Conclusiones (Parte II)

En esta parte del trabajo se han realizado experimentos para determinar el extracto de un texto (sin formato). Las técnicas utilizadas están basadas en el uso de la información mutua, como una forma de controlar la expansión de palabras usando contextos de un corpus, y la obtención de valores de similitud entre una oración y su complemento (para todas las oraciones del texto) y por otro lado, la obtención del valor de similitud entre el título de un texto y las oraciones restantes del mismo.

Estás técnicas derivaron dos experimentos, los cuales se describen a continuación:

En primera instancia se obtuvo el extracto de un texto, usando una función simplificada de Jaccard, para obtener el grado de similitud entre cada oración del texto y su correspondiente complemento. En este caso, la complejidad del problema es $O(n^2)$, para n igual al número de oraciones en el texto. Los resultados obtenidos son alentadores, mas no contundentes, ya que en el segundo experimento, en donde se hace uso de la función de Jaccard, para obtener la similitud del título del texto con las oraciones restantes del mismo (orden $O(n)$), se pueden o bservar algunos resultados que mejoran aquellos del primer experimento de esta parte, sin embargo en el segundo experimento, los cálculos se reducen dramáticamente.

Para los experimentos de esta parte del trabajo, se usó un nuevo corpus que consta de 375 textos, todos del dominio de **"Política"**. El tamaño del corpus

es de 2700 kb, y posee alrededor de 21559 raíces y 29741 oraciones. Cabe mencionar que la muestra de textos tomada (45) no esta contenida en el corpus.

Algunos resultados parciales de este trabajo fueron utilizados presentados en el Primer Congreso Internacional de Inteligencia Artificial, llevado a cabo en la ciudad de Hyderabad, India [15].

La experimentación en la generación de extractos para textos sin formato, ha sido gratificante para el autor y presenta aún más retos por vencer que al momento de iniciar el trabajo de investigación.

Algunas de las perspectivas planteadas para este trabajo se presentan a continuación:

a) convendria que un experto humano en la obtención de extractos evaluara los resultados obtenidos, de tal manera que se pudiese dar un sustento mayor a los valores calculados en el capítulo de pruebas.

b) El corpus usado no fue totalmente homogéneo, por lo que sería interesante poseer un banco de información (corpus) homogeneo y suficientemente extenso para continuar con las pruebas.

c) Sería deseable implantar otros métodos para la obtención de extractos, a fin de precisar la potencialidad de las técnicas desarrolladas.

Bibliografía

[1] Brigitte Endres-Niggemeyer *"SimSum: an empirically founded simulatio of summarizing"*, REGAMON 2000, pp. 659-682, 2000.

[2] Daniel Marcu: "The Automatic construction of large-scale corpora for summarization research", ACM-SINGIR '99, pp. 137-144, 1999.

[3] F.C. Johnson,C.D. W.J. Paice,Black & A.P. Neal *"The application of linguistic processing to automatic abstract generation"*, SIGIR '94

[4] Fano, R. *"Transmission of Information: A Statistical Theory of Communications"*, MIT Press, Cambridge, MA.

[5] García J.F.: "Estructural conceptual y comunicación", *"Dimensión antropológica"*, Año 2, Vol. 3, pp. 75-84, México, 1995.

[6] Gerard Salton, James Allan & Amit Singhal: "Automatic text decomposition and structuring", *"Information Processing and Management"*, V. 32, pp. 127-138, Elsevier, 1996.

[7] Grefentette, Gregory: "Automatic thesaurus generation from raw text using knowledge-poor techiniques", *"Xerox"*, Grenoble Lab.,1995.

[8] Grefentette, Gregory: "Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic relation", *"SINGLE/ACL, Workshop on Acquisition of Lexical Knowledge from Text Columbus"*, OH, 1993.

[9] Gustavo Crispino,Jean-Luc Minel & Javier Couto "Contexto: Una plataforma para la extracción de información y el resumen automático de textos", *"Proceedings of the 2nd. Workshop on Spanish Processing and Language Technologies"*,pp. 153-157, Universidad de Jaén, España, 2001.

[10] G. Ruge: "Combining corpus linguistics and human memory models for automatic term association", *"Text Information Retrieval"*, T. Strzalkowski *Ed.*, Kluwer, 1999. Universidad de Jaén, España, 2001.

[11] Hongyn Jing & Katheleen R. Mckeown *"The Decomposition of human-written summary sentences"*, ACM-SIGIR '99, pp. 129-151, 1999.

[12] I. Acero, M. Alcojor,A. Díaz, J.M. Gómez & M. Maña "Generación Automática de Resúmenes personalizados", *"Procesamiento de Lenguaje Natural* 27 ", pp. 281-298, SEPLN 2001.

[13] Jade Goldstein, Mark Kantrowitz, vibhu Mittal & Jaime Carbonell *"Summarizing Text Documents: Sentence Selection and Evaluation Metrics"*, ACM-SIGIR '99, pp. 121-128, 1999.

[14] Jiménez Salazar, H., Pinto Avendaño D. & Salazar Martínez H. *"Text extraction: a corpus-based approach"*, XXX Aniversario FCC-BUAP, noviembre 2003. ISBN: -968 863 7114, pag 92-94.

[15] Jiménez Salazar, H., Pinto Avendaño D. & Salazar Martínez H. *"Information Retrieval Base on Text Extraction"*, 1^{st} International India Conference on Artificial Intelligence, Session: Conceptual Information Retrieval, hynderabad la India 2003, 17 y 18 diciembre 2003.
ISBN: 0-9727412-0-8

[16] Kenneth Ward Church & Patrick Hanks *"Word Association Norms, Mutual Information"*, Computational Linguistics Volume 16, Number 1, March 1990.

[17] Kevin Knight & Daniel Marcu *"Summarizing beyond sentence extraction: A probabilistic approach to sentence comprensión"*, ELSEVIER SCIENCE 2002, pp. 91-107, 2002.

[18] Maher Joana & Abdel Majid Ben H.: "Automatic text summarization of scientific articles based on classification of extract's population", *"Lecture Notes in Computer Science 2588"*, A. Gelbukh *Ed.*, pp. 623-634, Springer, 2003.

[19] Palermo, D. S. & Jenkins, J. J. *"Word association norms:Grade school through college"*, Minneapolis: University of Minnesota

[20] Varaschin Gasperin, C. & Strube de Lima, V.L.: "Experiment on extracting semantic relations from syntactic relation", *"Lecture Notes in Computer Science "*, Vol. 2588, Springer, 2003.

[21] Wesley T. Chuang & Jihoon Yang *"Extracting Sentence Segments for text Summarization: A machine learning approach"*, ACM-SIGIR 2000, pp. 152-159, 2000. Press., 1964.

Apéndice A

Oraciones de un Texto T

El archivo descrito abajo corresponde al archivo etiquetado como $j06020\_7E32$.

Número de oración	Oración
1	El Banco Mundial desembolsará este año 1.5 mil millones de dólares en préstamos para México
2	La mayoría de estos financiamientos se enfocarán a mejorar las condiciones sociales de los pobres reforzar la estabilidad macroeconómica y afianza las reformas en el ejercicio del poder público
3	Según reporte de la institución además de los préstamos para el sector financiero y la descentralización otro tipo de financiamiento existente o planeado apoyará el creciente acceso a educación y servicios de salud para los pobres
4	Los programas de apoyo a México del Banco Mundial buscarán expandir los programas de desarrollo rural tratarán los problemas estructurales en el sector energía y mejorarán la protección ambiental
5	El documento analiza el comportamiento de los mercados que dice desarrollaron la capacidad de disciplinar a los Estados castigar sus errores y desenmascararlos
6	Esto se hace evidente en rápidas alzas y repentinos hundimientos de flujos de capital privados a los países en vías de desarrollo en la última década

7 Los flujos de capital privado netos eran de casi 43 mil millones de dólares en 1990 y subieron a 304 mil millones en 1997

8 En 1999 por la crisis asiática la devaluación rusa y la crisis de la divisa en Brasil cayeron a 239 mil millones de dólares

9 Así destaca el BM los flujos de capital se han vuelto inestables y selectivos

10 Construir y mantener políticas e instituciones para atraer y sostener la inversión no es un fácil pero debe lograrse incluso en las épocas del Estado más fuerte

11 Subraya que para los mexicanos este mensaje es obvio tienen atravesando su frontera norte a la población mayor de edad más rica del mundo

12 Los baby boomers de la posguerra estadunidense están en la última década de sus vidas económicamente activas

13 A medida que se acercan al retiro están bombeando dinero a los fondos de pensión

14 Para el 2002 se espera que los activos por pensiones globales alcancen 13.7 billones de dólares

15 Los administradores buscarán altos rendimientos en todo el mundo y México está bien posicionado para atraer cuantiosa parte de este capital

16 Sin embargo advierte atraer y mantener este capital depende de varios factores que incluyen un gobierno democrático bien administrado y con firmes políticas macroeconómicas y disciplina fiscal

Apéndice B

Representación de $T^{'}$

Para la oración $O_1 \epsilon\ T^{'}$ del ejemplo del apéndice A:

> Banco Mundial desembolsará año 1.5 mil millones dólares

Un fragmento de su representacion es:

> **Banco** [aplic gir Vicente previ conoc Guillermo Ferranti empres coincid primer económ validez economí internacional correct consultor recomend ambiental integral evalú Consejo Público secretari Desarrollo Latina mexican encuentr reduc deserción increment proces certidumbr Europeo apoy cit Económica particip Lafourcade deleg intención desempeñ Crédito gobi pequeñ visit VENEGAS vid federal altas Instituto perd millon mes áre nuev gobiern permanent reg República dispon sugier recient program independient económicohacendari Unidad organ Económico orient Kliksberg Bernardo logr director Quesada materi regl mil Ferranti división país año econom desarroll misión ciudad libr gerent reconstrucción inform recomend propósit competit Monetario visión prepar pobrezaignoranci año Guillermo Economía Grupo Turismo titul vía universaliz Política nacional pobrez Económica Martha median días reprob 1 2 Washington president entrant noviembr gent econom próxim entrevist camin Organización

abandon atra Desarrollo gubernatur logr ingres estructural impact financier JUAN señal afirm MANUEL 90 círcul eléctr apart atac liberaliz diversif sector irse part tributari principal clar foxist director Derbez fundamental orígen Jiménez agres merc financier anual Sahagún Pinos problem Ortiz inversión Francisco fiscal Interamericano Luis últim América Sahagún Avantel David pervers Banco Hacienda Perry integrant globaliz impuls energí miembr gobern fond Mundial Martínez BID Bancomer empresari export Secretaría administr march expand sistem coordin Europea FMI present protección expus divis tas salient equidad reform romp reun agres inform internacional frent déc Estudios trat Ernesto paralel rural integral gestión fij análisis pobrez Social mejor Díaz oficin subsecretari fin aspir estad Fondo Bancomext form México manifest Presidente analiz pas Europa son encarg crític contribu tres hotel vocer Internacional públic program Foro México representant aprovech oportun sostuv admit Olivier prelimin transparent nivel diseñ 470 gen trat gobiern primari tarif económ desembols enseñ expus Gil comerci emple Investigación oportun dólar Presidencia jef Fox educ ampli busc decision advirt Financiero dedic general proyección distribu Programa promotor margin eléctr vigent oper Unión list]

Mundial [aplic mism realiz Vicente comenz egres conoc anterior Ferranti 2001 empres globalifób coincid primer económ alianz correct consultor tanqu ocurr recomend Comercio ambiental coop model Mundial Desarrollo Consejo encuentr Latina reduc favorec proyect certidumbr apoy cit multitudinari Cancún Naciones particip loc gast Lafourcade intención desempeñ gobi visit anunci federal escenific millon mes apliqu áre nuev gobiern sugier program altern independient región discurs 2100 organ orient ocurr Económico vez callejer escenari materi regl acud excluy mil Ferranti agend país año grad María econom desarroll misión ciudad exclu recomend Monetario visión prepar OCDE viv Guillermo coeficient

año neoliberal loc riquez pobrez Unidas Seattle días cient 1 2 Washington president expert afirm entrant noviembr infantil salud 1988 próxim IPCC lun entrevist Organización Banco Bertrab asiát conferent 5 ingres logr estructural afirm von señal Global 90 part apart atac sector jef orig tributari aprob present principal clar respect foxist director grup segund agres Roo población anual octubr tierr Figueres Ortiz heroic problem inversión fiscal últim Económicos América opositor David ambient inmens Perry integrant impon energí OMC gobern trabaj empresari administr promov expand activ sistem FMI present protección UNEP caball salient Guerra equidad reform seman Foro nazis agres nivel inform polít entrev exist trat posibil rural paralel gestión fij pobrez Sureste protest presupuest mejor peor oficin reunión José Económico Fondo form México manifest cre pas Congreso Quintana encarg mortandad social tres Internacional program 40 Foro México representant sostuv Olivier prelimin transparent Mexicano OMM mayorí gobiern 1990 tiemp quep económ prens desembols mart Herman habl cas polac regional lanz dólar jef Fox inferior mund meteorológ 1999 educ Estado busc funcionari decision Suiza dedic distribu Programa general Centro margin Unión Ley vigent oper divers]

...

dólar [República proced dedic aplic Vicente comenz Bush Bolivia primer conoc 550 etap Malloch supuest record val 2001 sociedad estim empres 800 arruin comun lectur exig Brown conserv económ 5 continu asegur aproxim EU 2006 obtien cicl solucion 25 pobr años inclus continent pretendí lir mar Desarrollo mexican reduc UEDO ayud percib capital cit dat destac apoy 364 muj public cambi Guanajuato Training rival códig Naciones Disneylandia deleg reportaj 613 años intención 400 anual direct enfermedad utiliz

9 alocu relacion adult prést VIHsida Pérez ofrec pequeñ aument 2000 vid Fox and ola tres millon mes teorí esenci part millon piraterí subdeleg ex 10 República áre 408 nuev disminu program asciend crud cártel sensibl genét cápit Dominicana Unidas organ plaz Military raz const consider Ferranti Uruguay continu 16 mil mitad esfuerz crisis país 17 Wall Chile 1948 alrededor inyect Defensa result libr fronter empleador recibí design herman York necesit Unidos viv diferent intelectual arroj sól año 918 explot conflict Ecuador nacional propiedad gran 53 titul merc estudi 239 raíz 500 famili cient president 2 chilen luch próxim gent Education ataqu 3 Banco fallec Perú Union superior cianur temát 4 consign Pérez asiát 5 permitirí ingres Amado descens Brasil 7 entrad Norberto señal destin negoci laboratori argentin compr 9 recurs net cifr parqu devalu part mundial orig Díaz recurs priv gananci satisfac ahorr detall vent quint manten junt obrer maquil Honduras 1991 población coop Corporation merc prest prueb Ministerio recaud garantizaí niñ problem financ imagínens cuest 560 106 acuest cient mil Africa devolv militar recib Italia últim migración entreg mayor traduj control cardiac requier trabaj Larrieta gobern reducción antinarcót Mundial solicit mensual muj sexual George tardanz identific comercial billon export democrát Chihuahua traicion primari requ presupuest 569 administr ayud infraestructur febrer present incorpor peculi deleg plant Ecuador buen divis Fuentes cost Carrasco compañí fluj sanción renunci inform encontr clímax internacional nivel diari frent déc Salvador Trinidad Ernesto investig pais posibil crisis revel tem asegur capacit mejor póliz Indicó PGR confront sueld mercaderí acces mont José estad gan 20 opción través Zedillo punt Unidos Street rus form sol Gómez 206 lad pas son Alemania solicitant cercan día requer Carrillo Juárez crític Colombia nazi crec millón 40 Estados program Tamaulipas flagranci Forli Estados uva México acap nec 225 seri ener 700 cánc sostuv Venezuela convirt 43 pen arrest proporcion tip sub nivel Argentina genét grand 60 gobiern tonel 1990 pid señor

construcción razon millón per económ desembols ciel pes Pacific vitalici deten histór factur cas dólar moment 900 repercutirí 491 478 1997 80 Journal IMET siet w Fox 493 sum international conmin predispuest mund 1999 terci educ Suárez anunc plaz globalifobi 497 crític distribu Programa cayeron evident supuest 304 1988 vigent barril 533 oper pensión MEXICO preci]

Apéndice C

Oraciones de T Ordenadas por Similitud

4844 La mayoría de estos financiamientos se enfocarán a mejorar las condiciones sociales de los pobres reforzar la estabilidad macroeconómica y afianza las reformas en el ejercicio del poder público

4613 El documento analiza el comportamiento de los mercados que dice desarrollaron la capacidad de disciplinar a los Estados castigar sus errores y desenmascararlos

4528 Sin embargo advierte atraer y mantener este capital depende de varios factores que incluyen un gobierno democrático bien administrado y con firmes políticas macroeconómicas y disciplina fiscal

4351 Los baby boomers de la posguerra estadunidense están en la última década de sus vidas económicamente activas

4269 El Banco Mundial desembolsará este año 1.5 mil millones de dólares en préstamos para México

4164 En 1999 por la crisis asiática la devaluación rusa y la crisis de la divisa en Brasil cayeron a 239 mil millones de dólares

3818 Según reporte de la institución además de los préstamos para el sector financiero y la descentralización otro tipo de financiamiento existente o planeado apoyará el creciente acceso a educación y servicios de salud para los pobres

3730 Subraya que para los mexicanos este mensaje es obvio: tienen atravesando su frontera norte a la población mayor de edad más rica del mundo

3675 Los programas de apoyo a México del Banco Mundial buscarán expandir los programas de desarrollo rural tratarán los problemas estructurales en el sector energía y mejorarán la protección ambiental

3388 Así destaca el BM los flujos de capital se han vuelto inestables y selectivos

3380 Los flujos de capital privado netos eran de casi 43 mil millones de dólares en 1990 y subieron a 304 mil millones en 1997

2567 A medida que se acercan al retiro están bombeando dinero a los fondos de pensión

2516 Para el 2002 se espera que los activos por pensiones globales alcancen 13.7 billones de dólares

2248 Esto se hace evidente en rápidas alzas y repentinos hundimientos de flujos de capital privados a los países en vías de desarrollo en la última década

1689 Los administradores buscarán altos rendimientos en todo el mundo y México está bien posicionado para atraer cuantiosa parte de este capital

1107 Construir y mantener políticas e instituciones para atraer y sostener la inversión no es un fácil pero debe lograrse incluso en las épocas del Estado más fuerte

Apéndice D

Cuestionario

A continuación se muestra el cuestionario utilizado para cada uno de los diez textos de prueba. En este apéndice se hace referencia al texto $j06020\_7E32$

Estimado colaborador se solicita su ayuda para el siguiente procedimiento:

1. **Por favor lea el siguiente texto y marque las 5 oraciones que más representan al texto (señale con A la más represntativa hasta con E la menos representativa)**

Número de oración	Oración
1	El Banco Mundial desembolsará este año 1.5 mil millones de dólares en préstamos para México
2	La mayoría de estos financiamientos se enfocarán a mejorar las condiciones sociales de los pobres reforzar la estabilidad macroeconómica y afianza las reformas en el ejercicio del poder público
3	Según reporte de la institución además de los préstamos para el sector financiero y la descentralización otro tipo de financiamiento existente o planeado apoyará el creciente acceso a educación y servicios de salud para los pobres
4	Los programas de apoyo a México del Banco Mundial buscarán expandir los programas de desarrollo rural tratarán los problemas estructurales en el sector energía y mejorarán la protección ambiental
5	El documento analiza el comportamiento de los mercados que dice desarrollaron la capacidad de disciplinar a los Estados castigar sus errores y desenmascararlos

6 Esto se hace evidente en rápidas alzas y repentinos hundimientos de flujos de capital privados a los países en vías de desarrollo en la última década

7 Los flujos de capital privado netos eran de casi 43 mil millones de dólares en 1990 y subieron a 304 mil millones en 1997

8 En 1999 por la crisis asiática la devaluación rusa y la crisis de la divisa en Brasil cayeron a 239 mil millones de dólares

9 Así destaca el BM los flujos de capital se han vuelto inestables y selectivos

10 Construir y mantener políticas e instituciones para atraer y sostener la inversión no es un fácil pero debe lograrse incluso en las épocas del Estado más fuerte

11 Subraya que para los mexicanos este mensaje es obvio: tienen atravesando su frontera norte a la población mayor de edad más rica del mundo

12 Los baby boomers de la posguerra estadunidense están en la última década de sus vidas económicamente activas

13 A medida que se acercan al retiro están bombeando dinero a los fondos de pensión

14 Para el 2002 se espera que los activos por pensiones globales alcancen 13.7 billones de dólares

15 Los administradores buscarán altos rendimientos en todo el mundo y México está bien posicionado para atraer cuantiosa parte de este capital

16 Sin embargo advierte atraer y mantener este capital depende de varios factores que incluyen un gobierno democrático bien administrado y con firmes políticas macroeconómicas y disciplina fiscal

2. **LLenar el siguiente cuestionario**
   ¿Para usted cual sería la representatividad que tiene cada una de las oraciones siguientes sobre el texto del inciso 1.

Ordenamiento de Oración	Oración
◯	La mayoría de estos financiamientos se enfocarán a mejorar las condiciones sociales de los pobres reforzar la estabilidad macroeconómica y afianza las reformas en el ejercicio del poder público
◯	El documento analiza el comportamiento de los mercados que dice desarrollaron la capacidad de disciplinar a los Estados castigar sus errores y desenmascararlos
◯	Sin embargo advierte atraer y mantener este capital depende de varios factores que incluyen un gobierno democrático bien administrado y con firmes políticas macroeconómicas y disciplina fiscal
◯	Los baby boomers de la posguerra estadunidense están en la última década de sus vidas económicamente activas
◯	El Banco Mundial desembolsará este año 1.5 mil millones de dólares en préstamos para México

Apéndice E

Ejemplo de Similitud (Parte II)

Este ejemplo hace referencia al texto 071203*pol*_101*c*02
El archivo origina es el siguiente:

Por escriturar dos mil 238 predios en el Paraje San Juan
Armando Calderón
3 Nov 03

No. oración	Oración
1	A este respecto un informe del Gobierno Distrito Federal indica que el 30 de junio de 1989 se celebró un convenio
2	De acuerdo con documentación que analiza la comisión especial de la Asamblea Legislativa que investiga el caso del Paraje San Juan aún faltan de ser liberadas dos mil 283 escrituras que significan el 25 por ciento del total de viviendas que se regularizaron con la expropiación de 1989
3	A este respecto un informe del Gobierno Distrito Federal indica que el 30 de junio de 1989 se celebró un convenio entre la delegación Iztapalapa y la Dirección General de Registro Territorial con la Asociación de Residentes del Paraje San Juan el cual tuvo por objeto la regularización de los lotes de terreno ahí ubicados
4	De ahí nació la elaboración de ocho mil 770 escrituras de las cuales aún no se han liberado las dos mil 283 ya señaladas
5	Para todas ellas se estableció la modalidad de Contrato de Donación con Carga lo que implicaba que estaban limitadas hasta el pago de un gravamen por concepto de indemnización que sería utilizado para sufragar las reclamaciones de pago a quienes acreditaran ser sus propietarios

No. oración	Oración
6	De acuerdo al decreto transcurrido el plazo de 10 años sin que se presentaran reclamos de este tipo quienes adquirieron los lotes bajo el programa de regularización quedaron liberados de la obligación del pago
7	En cada una de las escrituras de los colonos regularizados expresa el documento se estableció el pago de la carga luego de un avalúo realizado por la Comisión de Avalúos de Bienes Nacionales mismo que se depositaría en un fideicomiso de administración que nunca se constituyó
8	De las ocho mil 770 escrituras emitidas 874 fueron liberaciones de carga mediante pago a la Tesorería y 293 a través de la oficina de Consignaciones del Tribunal Superior de Justicia del DF mediante respectivo billete adquirido en Nacional Financiera
9	Diez años después de la regularización de los predios el 6 de agosto de 1999 se emitió una declaratoria de liberación de escrituras del Paraje San Juan de manera gratuita y con ello a cinco mil 320 escrituras se les canceló la carga ante el Registro Público de la Propiedad y de Comercio del DF
10	El GDF informa a los diputados que fue por dicha declaratoria que Enrique Arcipreste del Ábrego quien hoy reclama una indemnización de mil 810 millones de pesos solicitó el amparo y protección de la justicia federal que fue sobreseído el 7 de marzo de 2001
11	Sólo que inmediatamente promovió otro amparo solicitando la indemnización expediente 508 98 que se radicó en el Juzgado Octavo de Distrito en Materia Administrativa y ahí recayó la sentencia que condena al GDF al pago
12	Ayer el secretario de Gobierno Alejandro Encinas reiteró No vamos a prestarnos a ninguna caricatura promovida por los abogados de Arcipreste porque son unos coyotes que buscan beneficiarse del erario público

A continuación se muestran los puntaje y la oraciónes de acuerdo a tipo de funcion utilizada:

A)Uitilizando la funcion de similitud (2.1)

Puntaje	Oración
52943	de acuerdo con documentación que analiza la comisión especial de la asamblea legislativa que investiga el caso del paraje san juan aún faltan de ser liberadas dos mil 283 escrituras que significan el 25 por ciento del total de viviendas que se regularizaron con la expropiación de 1989
52041	a este respecto un informe del gobierno distrito federal indica que el 30 de junio de 1989 se celebró un convenio entre la delegación iztapalapa y la dirección general de registro territorial con la asociación de residentes del paraje san juan el cual tuvo por objeto la regularización de los lotes de terreno ahí ubicados
51429	de las ocho mil 770 escrituras emitidas 874 fueron liberaciones de carga mediante pago a la tesorería y 293 a través de la oficina de consignaciones del tribunal superior de justicia del df mediante respectivo billete adquirido en nacional financiera
50961	diez años después de la regularización de los predios el 6 de agosto de 1999 se emitió una declaratoria de liberación de escrituras del paraje san juan de manera gratuita y con ello a cinco mil 320 escrituras se les canceló la carga ante el registro público de la propiedad y de comercio del df
50687	de acuerdo al decreto transcurrido el plazo de 10 años sin que se presentaran reclamos de este tipo quienes adquirieron los lotes bajo el programa de regularización quedaron liberados de la obligación del pago
50549	en cada una de las escrituras de los colonos regularizados expresa el documento se estableció el pago de la carga luego de un avalúo realizado por la comisión de avalúos de bienes nacionales mismo que se depositaría en un fideicomiso de administración que nunca se constituyó
49652	el gdf informa a los diputados que fue por dicha declaratoria que enrique arcipreste del ábrego quien hoy reclama una indemnización de mil 810 millones de pesos solicitó el amparo y protección de la justicia federal que fue sobreseído el 7 de marzo de 2001
43962	a este respecto un informe del gobierno distrito federal indica que el 30 de junio de 1989 se celebró un convenio
41763	para todas ellas se estableció la modalidad de contrato de donación con carga lo que implicaba que estaban limitadas hasta el pago de un gravamen por concepto de indemnización que sería utilizado para sufragar las reclamaciones de pago a quienes acreditaran ser sus propietarios
33529	de ahí nació la elaboración de ocho mil 770 escrituras de las cuales aún no se han liberado las dos mil 283 ya señaladas
32182	sólo que inmediatamente promovió otro amparo solicitando la indemnización expediente 508 98 que se radicó en el juzgado octavo de distrito en materia administrativa y ahí recayó la sentencia que condena al gdf al pago

B) Uitilizando la funcion de similitud (5.2)

Puntaje	Oración
0.515175	de ahí nació la elaboración de ocho mil 770 escrituras de las cuales aún no se han liberado las dos mil 283 ya señaladas
0.417902	diez años después de la regularización de los predios el 6 de agosto de 1999 se emitió una declaratoria de liberación de escrituras del paraje san juan de manera gratuita y con ello a cinco mil 320 escrituras se les canceló la carga ante el registro público de la propiedad y de comercio del df
0.378678	de acuerdo con documentación que analiza la comisión especial de la asamblea legislativa que investiga el caso del paraje san juan aún faltan de ser liberadas dos mil 283 escrituras que significan el 25 por ciento del total de viviendas que se regularizaron con la expropiación de 1989
0.373479	el gdf informa a los diputados que fue por dicha declaratoria que enrique arcipreste del ábrego quien hoy reclama una indemnización de mil 810 millones de pesos solicitó el amparo y protección de la justicia federal que fue sobreseído el 7 de marzo de 2001
0.338968	de las ocho mil 770 escrituras emitidas 874 fueron liberaciones de carga mediante pago a la tesorería y 293 a través de la oficina de consignaciones del tribunal superior de justicia del df mediante respectivo billete adquirido en nacional financiera
0.332407	sólo que inmediatamente promovió otro amparo solicitando la indemnización expediente 508 98 que se radicó en el juzgado octavo de distrito en materia administrativa y ahí recayó la sentencia que condena al gdf al pago
0.327317	a este respecto un informe del gobierno distrito federal indica que el 30 de junio de 1989 se celebró un convenio
0.32465	a este respecto un informe del gobierno distrito federal indica que el 30 de junio de 1989 se celebró un convenio entre la delegación iztapalapa y la dirección general de registro territorial con la asociación de residentes del paraje san juan el cual tuvo por objeto la regularización de los lotes de terreno ahí ubicados
0.292217	para todas ellas se estableció la modalidad de contrato de donación con carga lo que implicaba que estaban limitadas hasta el pago de un gravamen por concepto de indemnización que sería utilizado para sufragar las reclamaciones de pago a quienes acreditaran ser sus propietarios
0.289683	de acuerdo al decreto transcurrido el plazo de 10 años sin que se presentaran reclamos de este tipo quienes adquirieron los lotes bajo el programa de regularización quedaron liberados de la obligación del pago
0.275632	en cada una de las escrituras de los colonos regularizados expresa el documento se estableció el pago de la carga luego de un avalúo realizado por la comisión de avalúos de bienes nacionales mismo que se depositaría en un fideicomiso de administración que nunca se constituyó

Apéndice F

Ejemplo de Extracto (Parte II)

A continuación se muestra las cinco oración obtenidas que componen el extracto del texto.
Extrtacto obtenido del inciso A) Apéndice E

Puntaje	Oración
52943	de acuerdo con documentación que analiza la comisión especial de la asamblea legislativa que investiga el caso del paraje san juan aún faltan de ser liberadas dos mil 283 escrituras que significan el 25 por ciento del total de viviendas que se regularizaron con la expropiación de 1989
52041	a este respecto un informe del gobierno distrito federal indica que el 30 de junio de 1989 se celebró un convenio entre la delegación iztapalapa y la dirección general de registro territorial con la asociación de residentes del paraje san juan el cual tuvo por objeto la regularización de los lotes de terreno ahí ubicados
51429	de las ocho mil 770 escrituras emitidas 874 fueron liberaciones de carga mediante pago a la tesorería y 293 a través de la oficina de consignaciones del tribunal superior de justicia del df mediante respectivo billete adquirido en nacional financiera
50961	diez años después de la regularización de los predios el 6 de agosto de 1999 se emitió una declaratoria de liberación de escrituras del paraje san juan de manera gratuita y con ello a cinco mil 320 escrituras se les canceló la carga ante el registro público de la propiedad y de comercio del df
50687	de acuerdo al decreto transcurrido el plazo de 10 años sin que se presentaran reclamos de este tipo quienes adquirieron los lotes bajo el programa de regularización quedaron liberados de la obligación del pago

Extracto obtenido del inciso B) Apéndice E

Puntaje	Oración
0.515175	de ahí nació la elaboración de ocho mil 770 escrituras de las cuales aún no se han liberado las dos mil 283 ya señaladas
0.417902	diez años después de la regularización de los predios el 6 de agosto de 1999 se emitió una declaratoria de liberación de escrituras del paraje san juan de manera gratuita y con ello a cinco mil 320 escrituras se les canceló la carga ante el registro público de la propiedad y de comercio del df
0.378678	de acuerdo con documentación que analiza la comisión especial de la asamblea legislativa que investiga el caso del paraje san juan aún faltan de ser liberadas dos mil 283 escrituras que significan el 25 por ciento del total de viviendas que se regularizaron con la expropiación de 1989
0.373479	el gdf informa a los diputados que fue por dicha declaratoria que enrique arcipreste del ábrego quien hoy reclama una indemnización de mil 810 millones de pesos solicitó el amparo y protección de la justicia federal que fue sobreseído el 7 de marzo de 2001
0.338968	de las ocho mil 770 escrituras emitidas 874 fueron liberaciones de carga mediante pago a la tesorería y 293 a través de la oficina de consignaciones del tribunal superior de justicia del df mediante respectivo billete adquirido en nacional financiera

Apéndice G

Representación de una Palabra (Parte II)

La representacion de la palabra **nacional** sin aplicar IM es:
nacional apoyand recient inversion asum eficaz poder drogadict ana evacu presidencial reconcili presupuesto hawai trop sufrimient encarg 38p7 diputados tendenci alter obstacul and apoyari republicanos chilen torn are movimient guard altas 1998[a]

Un fragmento del vocabulario $Vfrc_{(k,h)}$ con su IM es:

$Vfrc_{(k,h)}$={. . . ,0.89866/(nacional,recient), 0.56465/(nacional,asum), 0.363512/
(nacional,chilen), 5.190287/(nacional,apoyari), . . . ,4.685215/(nacional,altas),. . . }

de acuerdo a $Vfrc_{(k,h)}$ se eliminan las palabras h donde $IM \leq 3$.

Ejemplo: $(nacional, recient)$ se elimina *recient* del vector de representación y así sucesivamente.

Al terminar de aplicar IM con el fragmento de vocabulario anterior, la representación de la palabra queda asi:

nacional apoyand inversion eficaz poder drogadict ana evacu presidencial reconcili presupuesto hawai trop sufrimient encarg 38p7 diputados tendenci alter obstacul and apoyari republicanos torn are movimient guard altas 1998[a]