1. RESUMEN AUTOMÁTICO

1.1. Resumen

El objetivo de esta web es explicar en qué consiste el **resumen automático de documentos** y cómo es posible combinar técnicas de **recuperación y organización de la información** para conseguir resultados favorables al respecto.

1.2. Introducción

En la actualidad, gracias a la amplia difusión de **información** en formato digital que se ha producido desde la aparición de la World Wide Web y sobre todo a través de Internet, ha surgido en los últimos años la necesidad de impulsar el desarrollo de tecnologías relacionadas con la **recuperación y extracción de información**.

En relación a dichas tecnologías, se ha dado un pequeño salto a la optimización del procesamiento de **información**, abriendo nuevas líneas de investigación centradas en el **resumen automático de documentos**.

1.3. Objetivos del resumen automático

Los objetivos que persigue el **resumen automático de documentos** son:

- Aprovechar las ventajas que ofrecen las tecnologías, para sustituir lo costoso que puede llegar a suponer para un humano realizar **resumenes de documentos** textuales muy extensos.
- Presentar de forma abreviada y precisa el contenido principal de un **documento**.
- Evitar que los usuarios malgasten tiempo innecesario en localizar una **información** concreta que estén buscando.

2. TIPOS DE RESUMEN AUTOMÁTICO

2.1. Tipos de resumen automático

La generación de **resumenes automáticos** depende de la cantidad de **información** a resumir y del tipo de **información** manejada. Así, se pueden distinguir 3 tipos de **resumen automático**:

- **Mono-documento**: es el más habitual y el que se va a explicar en mayor profundidad.
- **Multi-documento**.
- **Información no textual**.

2.2. Resumen automático mono-documento

Los **resumenes mono-documento** se realizan sobre un único **documento de texto** cuyo formato puede ser muy variado. Así, dependiendo de dicho formato, las técnicas utilizadas para generar un **resumen** pueden ser muy diversas. En esta sección, se explicarán cuáles son las técnicas básicas que implementan las herramientas hasta ahora desarrolladas.

Las 3 técnicas básicas para generar **resumenes automáticos** son:

- Técnicas superficiales.
- Técnicas basadas en entidades.
- Técnicas basadas en la estructura discursiva.

2.2.1. Técnicas superficiales

Estas técnicas consideran el **texto completo del documento** como un "todo" formado únicamente por cadenas de caracteres separadas en bloques grandes que constituyen los párrafos. Los procedimientos más habituales para distinguir las partes más importantes de ese "todo" y poder generar **resumenes automáticos** aceptables son:

2.2.1.1. Palabras frecuentes

Esta técnica consiste en seleccionar las frases que contengan un mayor número de las palabras que aparecen más veces en el **texto** completo.

2.2.1.2. Posición del texto

Esta técnica se basa en escoger como más relevantes las frases o palabras que aparezcan en lugares concretos del **documento**, que se consideran de mayor relevancia. Según la ubicación a la que se le dé más importancia se usan los siguientes métodos:

- Lead Method: si se considera que lo importante aparece al principio o al final.
- Según el género del **documento**.

  Algunos tipos de documento según su género podrían ser:

  Noticias de prensa: lo más importante aparece en el titular y primer párrafo.
  Artículos de investigación: lo esencial aparece en el abstract y las conclusiones.

2.2.1.3. Términos o frases indicativos

Esta técnica le da importancia a títulos, subtítulos, enlaces, **cadenas de texto** introductorias a algo relevante ("es importante", "en conclusión", "principalmente", "en **resumen**"…), que permiten valorar positivamente las frases donde aparecen. Sin embargo, **cadenas de texto** "penalizadoras" ("imposible", "difícilmente" …), harán que se valoren negativamente las oraciones en las que aparecen.

2.2.2. Técnicas basadas en entidades

Estas técnicas se basan en la utilización de técnicas de análisis morfosintáctico del **texto**. De esta forma, es posible determinar la categoría léxica de cada término (sustantivos, verbos, adjetivos, artículos, pronombres, preposiciones, …).

Una vez realizado dicho análisis, se construye un grafo que represente las relaciones o conexiones entre los términos del **texto** y poder elaborar los **resumenes**. Dichas relaciones pueden ser de tipo semántico (una naranja, una manzana y un plátano, son frutas), o de tipo temático (un alumno, un catedrático y un campus, son términos relacionados con el entorno de una universidad).

2.2.3. Técnicas basadas en la estructura discursiva

Son sistemas bastante más complejos basados en la estructura del discurso, y en donde hay que prestar especial atención a la cohesión y coherencia de los **resumenes** elaborados.

2.3. Resumen automático multi-documento

En este caso, el **resumen** se realiza sobre los contenidos de un conjunto de **documentos**.

En esta sección se abordarán los 2 temas siguientes:

- Requisitos que debe cumplir el **resumen automático multi-documento**.
- Tipos de **resumen automático multi-documento**.

2.3.1. Requisitos

El **resumen automático multi-documento** debe cumplir una serie de requisitos que se enumeran a continuación:

- Clustering: habilidad para agrupar **documentos** parecidos y buscar **información** relacionada.
- Cobertura: habilidad para localizar y extraer los puntos más importantes de varios **documentos**.
- Anti-redundancia: habilidad para minimizar redundancias entre los pasajes del **resumen**.
- Cohesión del **resumen**: habilidad para combinar pasajes textuales de forma que le resulte útil al lector. Las principales técnicas en este sentido son: ordenación de los pasajes del más importante al menos importante, de forma que si el lector deja de leer el **resumen** haya podido obtener los contenidos de mayor relevancia, ordenación de los pasajes por fecha, ordenación por temas...
- Coherencia: los **resumenes** generados deben ser totalmente entendibles por el usuario.
- Inconsistencias de las fuentes: como algunos **documentos** pueden contener errores con frecuencia, el **resumen** debería ser capaz de reconocer e informar de dichas inconsistencias.
- Actualizaciones: cuando se genera un nuevo **resumen automático**, debería tener en cuenta los **resumenes** previos generados.

2.3.2. Tipos

Los tipos de **resumen automático multi-documento** que existen son:

- **Resumen** a partir de las secciones comunes de los **documentos**: localiza las partes importantes que la colección de **documentos** tienen en común y las utiliza para efectuar el **resumen**.

- **Resumen** a partir de las secciones comunes y de las secciones únicas de los **documentos**: exactamente igual que el anterior, pero teniendo en cuenta también para el **resumen** las partes más importantes únicas de cada **documento**.
- **Resumen del documento** central: crea un **resumen sencillo del documento** central del grupo.
- **Resumen del documento** central y del resto: idéntico al anterior, pero incluyendo también pasajes y palabras claves del resto de **documentos** para obtener una cobertura más completa del conjunto de todos ellos.
- **Resumen del último documento** y del resto: genera el **resumen** a partir del **documento** más reciente de la colección, incluyendo pasajes y palabras claves del resto de **documentos**.
- **Resumen** a partir de las secciones comunes y de las secciones únicas de los **documentos** teniendo en cuenta la fecha: realiza un **resumen** de forma similar al segundo tipo pero dándole más importancia a los pasajes de los **documentos** más recientes.

2.4. Resumen automático no textual

El **resumen automático de información** no textual abarca todo tipo de **documentos** multimedia. No existen aún demasiadas soluciones para abordar este problema, y es que el tratamiento de **información no textual** plantea problemas adicionales a los que hasta ahora se han planteado. Algunos de estos problemas son:

- Los contenidos audiovisuales no aportan ningún tipo de **información textual** de la que extraer términos relevantes.
- Del problema anterior se deriva el que no sea posible realizar segmentaciones de contenido en oraciones o frases.

Al no estar esta página orientada al **resumen automático multimedia**, sino al **resumen automático de documentos de texto**, no se va a aportar más **información** en este campo, pero cabe destacar que el punto fuerte de investigación en esta línea está orientado a la segmentación temática.

3. LÍNEAS DE INVESTIGACIÓN

3.1. Líneas de investigación

Existen dos grandes líneas de investigación en el campo del **resumen automático de documentos**. Éstas son:

- Top-down: análoga a la **extracción de información**.
- Bottom-up: similar a la **recuperación de información**.

3.1.1. Línea Top-down

Se basa en la idea de que el usuario necesita localizar una determinada **información muy concreta de un texto**. Por eso, son necesarias y aplicables las técnicas de **extracción de información**.

El tratamiento de los **documentos** se realiza considerando a éstos como una colección de oraciones de las que se hace una selección de acuerdo a unos criterios específicos dados por el usuario y que guían el proceso de **resumen automático**. Dicha selección se realiza extrayendo los términos o frases relevantes para los criterios de búsqueda establecidos. Finalmente, de esa selección de frases se construye un **documento** único basado en el original, obteniendo así un **resumen automático** totalmente válido.

3.1.2. Línea Bottom-up

Se basa en la idea de que el usuario necesita realizar un **resumen automático de un documento**, en el que aparezca la **información** más importante contenida en el **texto original**.

Las técnicas que se están aplicando en este campo de investigación están estrechamente relacionadas con la **recuperación de información**. Así, la metodología empleada se basa en la **recuperación** de los términos más frecuentes del **texto** y construir posteriormente grafos semánticos que permitan construir frases totalmente nuevas y redactar un **resumen automático** totalmente genuino, en contraposición con el caso anterior, que era elaborado copiando las frases relevantes de forma idéntica a como estaban en el **documento original**.

Sin embargo, dichas técnicas son bastante más sofisticadas que para el caso anterior y los **resumenes automáticos** obtenidos hasta el momento en esta línea no han sido demasiado satisfactorios.

4. SOFTWARE

4.1. Herramientas de resumen automático

Existen ya herramientas o programas que realizan **resumen automático de textos**. Algunas de ellas son:

- Summarizer: software de **resumen automático** que reconoce varios formatos de **documento** (.doc, .pdf, .html …). Su **resumen** está formado únicamente por frases completas del **texto original**.
- Extractor: software de **resumen automático de textos**, correo electrónico y páginas web. El **resumen** que genera consta de listas de palabras clave y frases importantes.
- TextAnalyst: software muy completo de análisis de contenido textual. La calidad del **resumen** que elabora se debe a un buen equilibrio entre técnicas lingüísticas y redes neuronales.
- SweSum: software de **resumen automático de texto** que permite seleccionar el porcentaje de **resumen**. Además, puede sacar listados con palabras clave y estadísticas.
- Microsoft Word: también incluye un sistema básico de **autorresumen**.