ÍNDICE

Capítulo 1

INTRODUCCIÓN A LA MINERÍA DE TEXTO

Esquema

1. Introducción
2. Elementos básicos
3. Estructuras de texto
4. Paquetes R de análisis de texto
5. Parámetros estadísticos

1.1. INTRODUCCIÓN

La abundancia de información de toda índole ha llevado, en la actualidad, a la denominada sociedad del conocimiento. Pero no toda información es conocimiento, ya que éste implica una información generalmente extraida de texto después de procesar, resumir o relacionar el mismo. La información se almacena primordialmente en documentos de texto como libros, revistas, periódicos, diarios, artículos, correos, páginas web, cartas, etc. Por otra parte, se ha generado la necesidad de encontrar maneras de clasificar documentos y organizar su información de manera innovadora y relevante en lugar de las tradicionales listas ordenadas.

Como respuesta a los problemas derivados de la multiplicidad de información surge la minería de texto (MT) o text mining en su expresión inglesa. La minería de texto emerge con el propósito de extraer, analizar y procesar textos procedentes de grandes conjuntos de datos, así como también facilitar su presentación para la comprensión de un nuevo conocimiento. Ofrece a las organizaciones la posibilidad de explorar textos no estructurados para establecer patrones y conseguir información relevante.

En resumen, la minería de texto busca extrae información de datos no estructurados y encontrar patrones que son nuevos y desconocidos anteriormente. Su finalidad, por tanto, es el descubrimiento de grupos interesantes, tendencias, asociaciones y derivaciones en los patrones encontrados y su visualización para la deducción de nuevas conclusiones.

La minería de texto se puede considerar como un subconjunto de la minería de datos, al ser los textos datos en si mismo. La minería de datos utiliza técnicas de "machine learning" donde la estadística y la informática tienen especial relevancia. Además, la minería de texto utiliza técnicas basadas en el procesamiento de textos como la lingüística computacional y la recuperación de información.

1.2. OPERACIONES EN MINERÍA DE TEXTO

Básicamente los procesos que se realizan con los textos en la MT, según Montes y Gómez (2005), se pueden sintetizar así:

- Una *fase de preprocesamiento*, donde los textos son transformados en algún tipo de representación semiestructurada que permita su análisis automático, y

- una *fase de descubrimiento*, donde las representaciones intermedias son analizadas y algunos patrones interesantes, como por ejemplo: agrupamientos, asociaciones, desviaciones y/o tendencias pueden ser descubiertos.

Liddy (1998) sugiere que estos procesos se componen de tres etapas:

- *Preparación del texto*: selección, limpiado y preprocesamiento del texto. En esta etapa tienen lugar procesos como la identificación y el etiquetado de las partes de la oración.
- *Procesamiento del texto*: uso de algoritmos para procesar los datos, comprimiendo y transformándolos para identificar partes importantes de información.
- *Análisis del texto*: evaluación del rendimiento para valorar si la información fue proporcionada de forma correcta y relevante.

1.3. USOS DE LA MINERÍA DE TEXTO

Existen diferentes herramientas estadísticas e informáticas de apoyo a la MT con la finalidad de ofrecer nuevo conocimiento. Estas herramientas permiten realizar una serie de funciones:

• *La extracción de características (entidades)*.- Proceso de identificar referencias de nombres de personas, instituciones, eventos, autoridades existentes y sus relaciones.

• *La generación de agrupamiento también conocido como clustering*.- Para agrupar documentos similares sin conocimiento previo de las agrupaciones, lo que se denomina en "machine learning" aprendizaje no supervisado. Esto significa que la agrupación será definida por el programa informático y no por una lista de clases predefinidas. Permite evaluar la relevancia de los documentos de cada grupo. Además, identifica relaciones desconocidas y duplicados potenciales. Conjuntamente, optimiza la organización de los resultados (Eíto Brun, 2004).

• *La categorización automática*.- Determina el tema o temas que trata una colección de documentos, lo que se denomina en "machine learning" aprendizaje supervisado. Este a diferencia del clustering, decide la clase a la que un documento pertenece dentro de una lista de clases predefinida. Como ejemplos se tiene la detección de spam en emails, etiquetar automáticamente flujos de artículos, etc. La clasificación empieza con un conjunto de entrenamiento de los documentos que son previamente clasificados; se crea un modelo de clasificación que basado en el conjunto de entrenamiento es capaz de asignar la clase correcta de un nuevo documento denominado de validación (Hotho, Nürnberger, & Paaß, 2005).

• *Identificar ideas principales (topic)*.- Reconoce y extrae los principales temas o ideas tratados por la colección de documentos. A diferencia de la categorización de documentos, este procedimiento permite extraer los términos que son representativos del texto sin asignarlos a una clase. Una idea se identifica buscando la ocurrencia de

términos y combinaciones de términos en los documentos. Al identificar cada idea se creará redes conceptuales a través de los documentos que traten de la misma idea.

• *Elaboración automática de resúmenes*.- Los resúmenes automáticos de texto se pueden realizar mediante extracción de frases relevantes y mediante abstracción, es decir, recogiendo las ideas relevantes con frases contenidas en el texto o ajenas al mismo. El procedimiento más utilizado es el primero, aún cuando está en continuo desarrollo el segundo. La extracción se sustenta en la frecuencia estadística de los términos encontrados, así como de la posición que ocupan estas frases en el texto. Facilita el análisis de grandes colecciones de documentos.

• *Análisis de opiniones*.- Permite el tratamiento de datos de texto recogidos en opiniones de establecimientos o productos, cartas, opiniones en medios de comunicación, etc. Básicamente el contenido de texto sigue un proceso de contraste con léxico evaluado en una escala dicotómica (bien, mal) o multicategórica (muy bien-muy mal) obteniendo resultados a nivel de frase o a nivel de documento, documentos, etc.

• *Visualización de documentos*.- Permite mostrar los textos en un formato que facilita la interpretación y navegación de colecciones de texto o elementos relevantes de los mismos (por ejemplo: nube de palabras).

1.4. ELEMENTOS BÁSICOS

Existen unos elementos básicos en los textos que permiten realizar el tratamiento informático-estadístico de los mismos. De forma detallada son:

- *Caracteres*.- Unidades elementales numéricas, textuales o de símbolos que configuran cualquier documento; siendo, por tanto, los pilares para la formación de palabras, frases, documentos o corpus.

- *Palabras*.- La palabra como, unidad textual primaria, no aporta en sí misma información semántica al documento, pero permite mediante el análisis de su frecuencia destacar los términos más utilizados en un determinado documento. Además, añade información morfológica según el papel que desarrolla y sobre la riqueza léxica del documento o corpus (relación entre el número de palabras distintas y el total de palabras del corpus). Es relevante destacar que las palabras no son independientes del contexto y por tanto tendrán una carga semántica según el mismo. Una representación visual que se utilizará con las palabras es la llamada nube de palabras o bolsa de palabras que será objeto de un estudio detallado más adelante.

- *Frases*.- Una frase es una secuencia de palabras, con un cierto nexo sintáctico, en un determinado texto.

- *Párrafo*.- De una forma operativa un párrafo es un conjunto de frases que terminan en un punto y aparte. Pero conceptualmente un párrafo es una o más frases que expresan una idea y por tanto existe coherencia dentro del mismo.
- *Documentos*.- Un documento está formado por un conjunto de párrafos cuyo objetivo sean expresar una o varias ideas. Puede ser, por tanto, uno de los elementos constituyentes de un corpus.
- *Corpus*.- Etimológicamente la palabra corpus proviene del latín y significa cuerpo. Por tanto, un corpus puede significar simplemente un grupo de textos.Sin embargo, un corpus difiere de un simple conjunto de textos porque es creado para ser sometido análisis lingüístico (C. F. Meyer, 2004) y en ese sentido está estructurado y planificado con la idea de ser una muestra representativa de datos para un estudio o proceso cuyo objetivo final es obtener información relevante. El tamaño del corpus viene derivado de la representatividad exigida, por lo cual, no hay valores patrones.

Pautas para la construcción de un corpus terminológico se puede ver en: Vargas-Sierra (2006). Se destaca que la adecuación textual del corpus al proyecto terminológico implica tener en cuenta aspectos como:

1) que el dominio, tema o tipo de textos que contenga el corpus estén bien definido y bien delimitado;
2) que los textos sean suficientemente representativos y que el conjunto de los mismos resulte en una muestra adecuadamente equilibrada para, con todo ello, fundamentar las conclusiones que se deriven;
3) que la organización y el contenido del corpus favorezcan su explotación;
4) que los textos sean adecuados en tamaño y formato, de modo que no surjan problemas de compatibilidad con las diferentes herramientas informáticas que utilicemos en su tratamiento y explotación (Sánchez, 1995).

1.5. ESTRUCTURAS DE TEXTO

1.5.1. N-gramas

Un n-grama es una subcadena de caracteres obtenida de una palabra dada de mayor tamaño. Por ejemplo: de la palabra **programa** podemos obtener los bigramas: pr, ro, og, gr, ra, am, ma; los trigramas: pro, rog, ogr, gra, ram, ama; o cuadrigramas: prog, rogr, ogrm, gram, rama.

La idea de n-gramas se puede generalizar a la unión de palabras en una frase. Así por ejemplo, sea la siguiente frase: **el castellano es un idioma muy utilizado** se pueden formar los trigramas: “el_castellano_es”,”el_castellano_un”, “el_es_un”,”el_es_idioma”, “castellano_es_un”,”castellano_es_idioma”, “castellano_un_idioma”,”castellano_un_muy”, “es_un_idioma”,”es_un_muy”, “es_idioma_muy”,”es_idioma_utilizado”, “un_idioma_muy”,”un_idioma_utilizado”, “un_muy_utilizado”,”idioma_muy_utilizado”.

1.5.2. Concepto

La palabra concepto viene del latín conceptus y representa la expresión, generalmente por un término, de un pensamiento mediante palabras. Es una idea abstracta que como resultado de la interacción con el entorno produce unos resultados que finalmente se expresan con palabras.

Por ejemplo, el concepto de *electricidad* se puede expresar así: ”Forma de energía que produce efectos luminosos, mecánicos, caloríficos, químicos, etc., y que se debe a la separación o movimiento de los electrones que forman los átomos”.

1.5.3. Tesauro

Palabra que proviene del griego thēsaurós que significa almacén, tesorería. Representa una secuencia de palabras, empleadas para representar conceptos.

En líneas generales, un tesauro comprende lo siguiente (wikipedia):

- Una lista de términos preferentes, ordenados en forma alfabética, temática y jerárquica.
- Una lista de sinónimos de esos términos preferentes, llamados descriptores, con la leyenda “úsese (término preferente)” o una indicación similar.
- Una jerarquía o relaciones entre los términos. Esto se expresa con la identificación de “términos más generales” y ”términos más específicos”.
- Las definiciones de los términos, en caso de ambigüedad, para facilitar la selección de los mismos por parte del usuario.
- Y un conjunto de reglas para usar el tesauro.

Ejemplos de tesauros son entre muchos: AGROVOC de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) sobre temas de agricultura y Thesaurus of Psychological Index Terms (APA) de términos de Psicología.

1.5.4. Jerarquía de conceptos o taxonomía de términos

Se entiende por taxonomía aquella clasificación compuesta por uno o más niveles, siendo la forma más simple la de un sólo nivel, como por ejemplo un vocabulario en el que cada nivel representa un término o concepto (Wason, 2006).

Las taxonomías se suelen representar con una jerarquía de términos donde los niveles superiores se utilizan para representar los conceptos más generales.

Estas construcciones también se suelen presentar mediante un árbol o grafo dirigido acíclico, es decir, existe un nodo padre del que cuelgan uno o varios hijos.

Por ejemplo, se puede expresar la siguiente taxonomía: medios didácticos (nodo padre) y radio, televisión, video, etc. (nodos hijos).

1.5.5. Redes semánticas (IS-A)

Las redes semánticas son un caso particular de red jerárquica que viene representada por un grafo dirigido, donde las características del nodo padre son heredadas por los nodos hijos. Asi por ejemplo: un *gato* es un *felino*, un *felino* es un *mamifero*, un *mamifero* es un *animal*.

Como vemos las redes semánticas permiten la generalización de conceptos. En el ejemplo se parte del concepto gato y se llega a la generalización de animal.

Las redes IS-A tienen dos tipos de nodos:

- *Token*: nodos de nivel inferior que heredan las características de los nodos de capas superiores y tienen características propias.
- *Type*: nodos de nivel superior que representan a las clases de individuos.

1.6. PAQUETES R DE ANÁLISIS DE TEXTO

El programa R(https://cran.rediris.es), software utilizado en estadística y análisis de datos, tiene una serie de paquetes muy utilizados en minería textual. En el CRAN de R (https://CRAN.R-project.org/view=NaturalLanguageProcessing) puede encontrar información pormenorizada del conjunto de los mismos. No obstante, se presenta a continuación los de mayor relevancia:

Marcos:

- **tm** proporciona un marco integral para la minería de texto con R. En un artículo de la revista *Journal of Statistical Software* (Feinerer, 2013) se presenta un resumen detallado de los métodos y técnicas utilizados para el recuento de palabras, la agrupación y preprocesamiento de texto, etc.