Análisis de Datos Masivos para el Negocio

Tema 8. Técnicas de procesamiento del lenguaje natural

Índice

Esquema

Técnicas de Procesamiento del Lenguaje Natural

- La mayoría de la información almacenada por la humanidad está en forma de texto almacenado.
- Los emails, blogs, libros, páginas web, etc., se basan en textos expresados en el denominado **lenguaje natural**.
- Lenguaje natural se refiere al lenguaje que utilizan los humanos para comunicarse de forma general o natural.
- Los textos generados por personas contienen mucho **valor empresarial** ya que incluyen información sobre lo que les gusta o disgusta, qué sabe o qué les gustaría saber, sus intereses, deseos, etc.

TÉCNICAS BÁSICAS DE PROCESAMIENTO DEL LENGUAJE NATURAL

Técnicas que procesan textos no estructurados para obtener información estructurada de ellos.

Técnicas básicas

- **Tokenización.** Dividir en tokens la secuencia de caracteres de un texto.
- **Normalización del texto.** Homogeneización de términos. Lemización y stemming.
- **Etiquetado gramatical**. Asignación de una categoría gramatical a cada palabra o token.
- **Reconocimiento de entidades.** Identificación de secuencias de caracteres que se refieren a nombres de personas, lugares, organizaciones, expresiones de fecha y hora, cantidades, valores monetarios, porcentajes, etc.
- **Creación de índice invertidos**. Agilización de búsquedas en documentos.
- **Extracción de la información.** Obtener información estructurada relevante a partir de texto no estructurado. Se basa en las técnicas anteriores.

MINERÍA DE TEXTOS

Proceso que aplica técnicas de aprendizaje automático a textos no estructurados.

- El objetivo es extraer información no trivial de un conjunto de textos no estructurados.
- Se necesita una fase de preprocesado que extraiga información estructurada del texto no estructurado.
- El preprocesado se realiza mediante técnicas básicas de procesamiento del lenguaje natural

Aplicaciones

- **Clasificación de documentos**. Dado un nuevo documento, asignarle una categoría de entre un grupo prefijado. Descubre la temática tratada en el texto.
- **Agrupación de documentos**. Dado un conjunto de documentos de texto, los *agrupa* en subconjuntos en función de lo relacionados que estén entre sí.
- **Resúmenes automáticos**. Proporciona un nuevo texto que incluye únicamente la información relevante del original.
- **Análisis de sentimientos**. Minería de opinión. Estudia los sentimientos y opiniones que sobre productos, temas, personas, etc han sido incluidas en textos.

8.1. Introducción y objetivos

El procesamiento del lenguaje natural (NLP) es un campo de la ciencia de la computación o la inteligencia artificial cuyo centro de interés es el tratamiento de los lenguajes humanos. El NLP se aplica actualmente a simuladores de conversación, reconocimiento de voz, traducción automática, chequeo ortográfico y también a la minería de texto. Esta última aplicación es en la que se va a centrar este tema. La minería de texto y análisis de texto es el procesado de un texto en lenguaje natural con el objetivo de extraer información útil. Por tanto, el fin último de la minería de textos es obtener información a partir de un conjunto de documentos de texto que están expresados en lenguaje natural o lenguaje humano. Entre las aplicaciones que podemos encontrar para la minería de textos se encuentran, la extracción de información estructurada a partir de texto no estructurado, la clasificación automática de documentos por temática tratada, la agrupación automática de documentos por contenidos similares, el resumen automático de documentos, el análisis de sentimientos basados en texto, etc.

Objetivos que se pretenden conseguir en este tema:

- Entender qué es el procesamiento del leguaje natural y la minería de texto.
- Conocer las técnicas básicas de procesamiento de lenguaje natural.
- Conocer las principales aplicaciones de la minería de texto.
- Conocer la estructura de un sistema basado en minería de texto.

8.2. Información basada en texto

La mayoría de la información almacenada por la humanidad está en forma de texto almacenado. Los emails, blogs, libros, páginas web, etc., se basan en textos expresados en el denominado **lenguaje natural.** Lenguaje natural se refiere al lenguaje que utilizan los humanos para comunicarse de forma general o natural. El mundo empresarial está convencido de que los textos que las personas generan contienen mucho valor ya que incluyen información sobre lo que les gusta o disgusta, qué sabe o qué les gustaría saber, sus intereses, deseos, etc. Esta información puede ser muy relevante para las empresas o investigadores. Sin embargo, el procesamiento de toda esa ingente cantidad de información no la pueden realizar las personas, es necesario la intervención de computadores que realicen ese trabajo.

La información textual puede estar expresada en lenguaje natural o no. Ejemplos de textos no expresados en lenguaje natural pueden ser los ficheros de registro o historial generados por las aplicaciones de los computadores, ficheros de intercambio de información entre aplicaciones, etc. Las técnicas comentadas en este tema también se pueden aplicar a este tipo de información textual, pero el tema se centrará en la información textual expresada en lenguaje natural, el mismo que utilizan los humanos para comunicarse.

Sin embargo, el procesamiento del lenguaje natural es difícil debido a su complejidad inherente. A continuación, se nombran algunas de las características del lenguaje natural que dificultan su procesamiento:

- **Ambigüedad.** Una expresión en lenguaje natural puede tener más de una interpretación. La resolución de la ambigüedad puede necesitar una amplia comprensión del contexto donde está ubicado el texto. Por supuesto, para un computador esta comprensión no es sencilla. Existen ambigüedades a nivel de palabra (palabras con más de un significado). A nivel referencial, por ejemplo,

pronombres que se refieren a una entidad antes citada. A nivel estructural, por ejemplo, no claridad en los complementos, lo que necesita de la interpretación semántica para resolver la ambigüedad. O incluso a nivel de oración, al utilizar recursos como la ironía o metáforas.

- **Sinónimos.** Muchas palabras pueden tener significados similares o parecidos.
- **Imperfecciones en el texto.** Los textos debido a errores humanos, fallos en las transmisiones o en la digitalización pueden incluir errores que para un humano pueden ser fácilmente identificables, pero que para el computador puede ser complicado.

Técnicas básicas de análisis de texto

Existen un conjunto de técnicas básicas que permiten analizar la información textual. Este tipo de técnicas son aplicables a cualquier información textual y conforman un primer paso a la hora de generar información estructurada del texto no estructurado. A continuación, se muestran algunas de estas técnicas de análisis básicas.

Tamaño del texto

Dado un texto, la primera información que se puede obtener es su tamaño en caracteres, palabras, párrafos, etc.

Frecuencia de palabras

Un primer análisis de un texto es realizar calcular la frecuencia con la que aparece cada palabra en el texto. Puede ayudar a localizar las palabras más relevantes del texto o la probabilidad de aparición de una palabra.

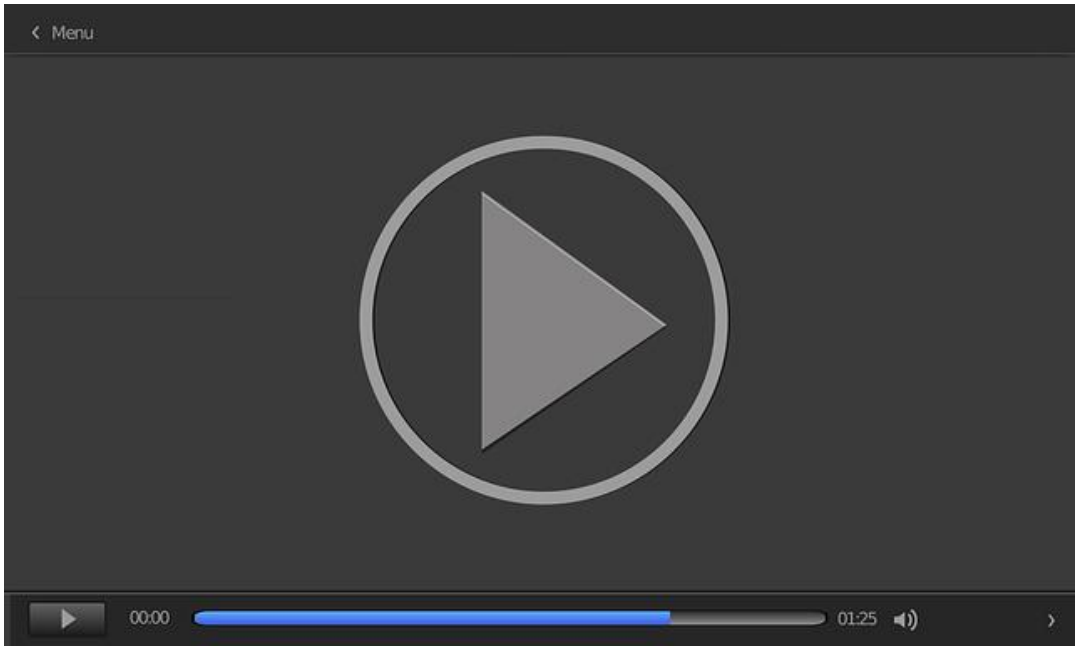
N-gramas

Cálculo de la frecuencia de aparición de una secuencia de palabras. Si la secuencia es de dos palabras se le denomina bigrama y es de tres palabras se denomina trigrama. La disponibilidad de n-gramas proporciona la probabilidad de que aparezca

en un texto cierta secuencia de palabras. Esta información se puede usar por ejemplo para completar frases u oraciones ya que los n-gramas también se pueden interpretar como la probabilidad de proximidad de dos o más palabras.

8.3. Técnicas básicas de procesamiento de lenguaje natural

Vídeo *Técnicas de procesado del lenguaje natural*.



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=0ea94a50-744e-43d8-a6c8-b15d00aa264c

Como ya se comentó en la sección anterior, el procesamiento del lenguaje natural tiene ciertas dificultades. Es por ello que se necesitan técnicas especializadas que permitan su análisis. Estas son algunas de las técnicas de análisis básico para el procesamiento del lenguaje natural:

- Tokenización o análisis léxico.
- Normalización de los términos.

- Etiquetado gramatical y análisis sintáctico.
- Reconocimiento de entidades.
- Creación de índices invertidos.

Análisis léxico o tokenización

Dado un texto fuente que es una secuencia de caracteres alfanuméricos (letras, número y posibles símbolos), el análisis léxico extrae las palabras o Tokens que componen ese texto. El texto pasa a ser una secuencia de Tokens. Se prefiere el vocablo Token ya que se quiere enfatizar que es una generalización de las palabras. Por ejemplo, una fecha (1972) aunque no es estrictamente una palabra sí puede ser un Token resultante de un texto.

Tokenización

Dado el texto de ejemplo:

«1492 es el año del descubrimiento de américa».

La salida en tokens puede ser un fichero con este aspecto:

<sentence>

<word>1942</word>

<word>es</word>

<word>el</word>

<word>año</word>

<word>del</word>

<word>descubrimiento</word>

```
<word>de</word>

<word>américa</word>

</sentence>
```

Es importante indicar que en el análisis léxico hay que tomar ciertas decisiones que pueden condicionar los procesos posteriores. Básicamente hay que establecer qué consideramos palabras relevantes del texto. Por ejemplo, hay que decidir si convertimos en tokens ciertas cifras o números (fechas, edades, identificadores, etc.). También hay que decidir si se incluyen en los tokens o no algunos los símbolos especiales, tales como símbolos de puntuación, tildes, guiones, etc. Diferenciar tokens con mayúsculas y minúsculas puede ser interesante en algunos casos para diferenciar por ejemplo apellidos de un nombre común. Por último, existen términos que están compuestos por varias palabras. A veces puede ser interesante y muy útil considerarlos un único token.

Otra de las decisiones a tomar es si se eliminan palabras que pueden tener una importancia relativa en el texto, por ejemplo, artículos, preposiciones, conjunciones, o incluso algunos verbos, adverbios y adjetivos. Por otro lado, en la mayoría de los documentos, las palabras que más se repiten no suelen aportar información importante en el texto, por lo que también se podrían eliminar.

La **tokenización de un texto** permite darle la primera estructura básica. Es la base sobre la que aplicar el resto de técnicas de procesamiento de lenguaje natural.

Normalización del texto

La **normalización del texto** se refiere a la simplificación de familias de palabras que en esencia tienen el mismo significado y que por tanto pueden resumirse en un

mismo token.

Uno de los pasos a realizar con los tokens es la unificación de forma.Existen palabras que pueden presentar diferentes formas (por ejemplo, NY, New York, New York) pero que se refieren a la misma entidad y que por tanto debieran ser el mismo token.

Teniendo en cuenta que la flexión de una palabra es su modificación mediante la utilización de morfemas (posiblemente prefijos y sufijos), el análisis morfológico o lematización consiste en la localización de estas palabras flexionadas y su sustitución por la palabra que por convenio se considera la representante de esa familia flexionada o también denominada lema. Por ejemplo, las palabras «gato»,«gatito», «gata», «gatos», «gatas»se pueden sustituir por una única palabra que en este caso podría ser «gato». En la figura 1 se pueden ver ejemplos de lematización.

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

Figura 1. Ejemplo de lematización. Fuente: https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/

Otro paso hacia la normalización consiste en la reducción de las palabras a su forma básica o raíz de la que deriva (también denominado *stemming*). Básicamente consiste en hacer a los tokens independientes de género, conjugación, número gramatical, etc. Existen algoritmos que eliminan las partes finales de las palabras (por ejemplo, el algoritmo de Porter). La figura 2 muestra ejemplos de *stemming*.

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Figura 2. Ejemplo de stemming. Fuente: https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/

La diferencia entre la lematización y el stemming es que **la lematización** reduce la familia de palabras a su forma (palabra) de diccionario, mientras que **el stemming** reduce las palabras a su raíz, que no tiene por qué ser una palabra es sí misma, sino una parte común a esa familia de palabras.

Las técnicas de normalización del texto permiten disminuir el número de tokes, con lo que el procesado del texto se hace más sencillo. Sin embargo, un abuso de la normalización puede hacer perder riqueza expresiva en la tokenización resultante.

Etiquetado gramatical y análisis sintáctico

El etiquetado gramatical, también denominado Part of Speech Tagging o POS Tagging, consiste en asociar a cada palabra su categoría gramatical. La figura 3 muestra un ejemplo de etiquetado.

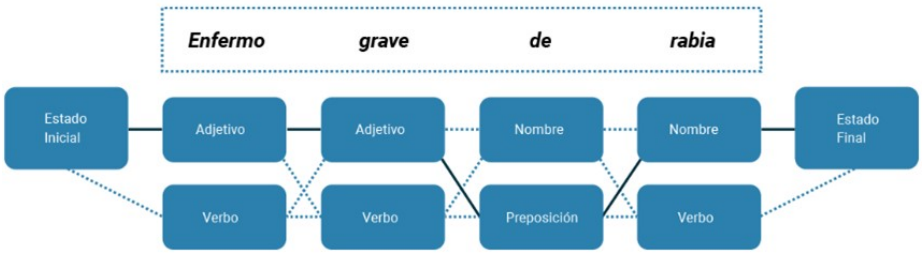


Figura 3. Ejemplo de etiquetado gramatical. Fuente: https://medium.com/soldai/etiquetado-gramatical-a418278e115c

Una vez realizado el etiquetado gramatical el siguiente paso puede ser el análisis sintáctico que puede proporcionar el árbol sintáctico de cada oración. El etiquetado permite diferenciar las funciones de los tokens (sustantivos, adjetivos, verbos, etc.)

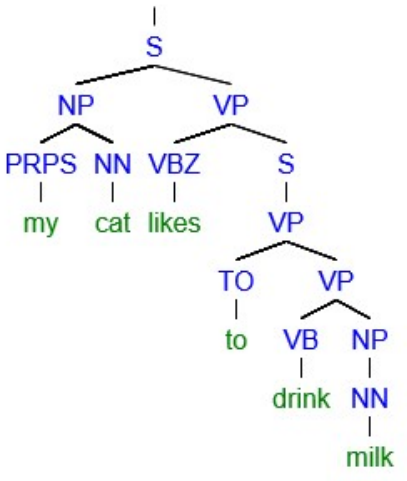


Figura 4. Ejemplo de árbol sintáctico. Fuente: https://www.ionos.es/digitalguide/online-marketing/vender-en-internet/como-funciona-el-natural-language-processing/

La figura 4 muestra el árbol de análisis sintáctico para la frase: «My cat likes to drink milk». La oración (S) se divide en sintagma nominal (NP) y sintagma verbal (VP) y este a su vez se subdivide en otros sintagmas. La creación de árboles sintácticos permite definir dependencias y grados de importancia de los distintos tokens. No es lo mismo un token que actúa como sujeto que como complemento, por ejemplo.

Reconocimiento de entidades

El reconocimiento de entidades (Named Entity Recognition, NER) consiste en la identificación de secuencias de caracteres que se refieren a nombres de personas, lugares, organizaciones, expresiones de fecha y hora, cantidades, valores monetarios, porcentajes, etc. El reconocimiento de dichas entidades se puede basar en reglas gramaticales codificadas manualmente o bien mediante aprendizaje automático y colecciones de entrenamiento anotadas.

Creación de índices invertidos

Una de las acciones más habituales en los documentos de texto es la búsqueda de términos o palabras. Por ejemplo, el usuario proporciona una palabra y desea saber en qué documentos aparece dicho término. La solución obvia es realizar un recorrido secuencial de todos los documentos de texto y devolver la localización de las repeticiones del texto. Sin embargo, esta solución solo es operativa hasta cierto tamaño de los documentos. A partir de ese tamaño crítico, el tiempo de búsqueda puede ser inaceptable. Una solución a este problema es disponer de un fichero que guarda un conjunto de índices invertidos. A groso modo este fichero contiene todos los términos o palabras incluidas en los documentos, y sus localizaciones. Por tanto, cuando el usuario realiza una búsqueda, no se accede a los ficheros originales, sino el fichero de índices para obtener la respuesta.

En la figura 5 se puede ver un ejemplo simplificado de la creación de un fichero de índices. Primero se realiza una tokenización de los documentos originales, manteniendo únicamente los términos deseados. Finalmente se crea el fichero, que no es más que una tabla, donde aparecen los términos de interés y los documentos donde aparecen. Por ejemplo, la palabra «Blue» que es la segunda entrada de la tabla, aparece en los documentos 1 y 3.

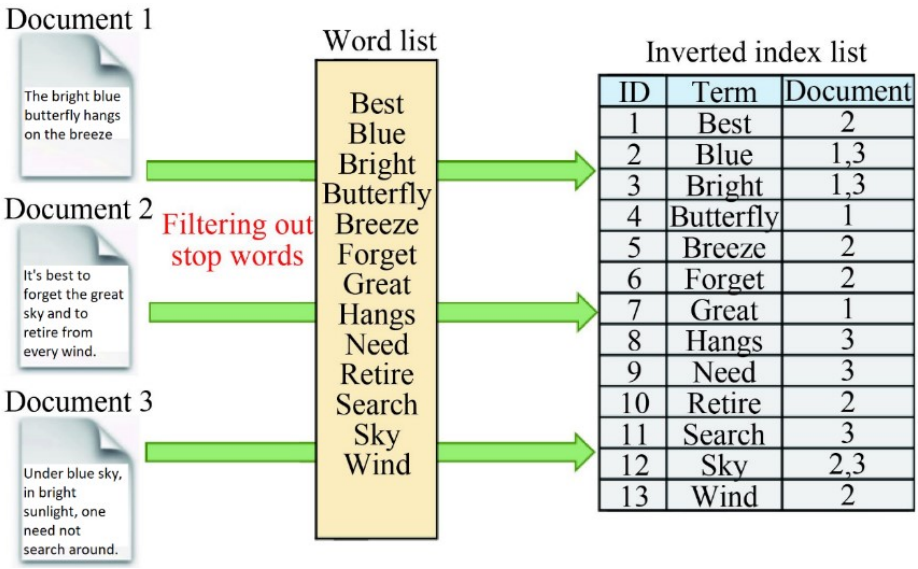


Figura 5. Ejemplo fichero de índice invertido. Fuente: http://www.xml-data.org/DZKJDXXBYWB/html/20170208.htm#zz

Se pueden crear índices más precisos, que por ejemplo indiquen la posición dentro de un documento o bien el número de ocurrencias de esa palabra en cada documento. En resumen, los ficheros de índices es información estructurada obtenida a partir de texto no estructurado que acelera ciertas acciones como por ejemplo la búsqueda de términos.

Extracción de la información

Dado un conjunto de documentos de texto no estructurado, es de gran interés extraer la información relevante que contienen y proporcionarle un formato más estructurado. La estructuración del texto permite el almacenamiento en bases de datos relaciones y facilita su procesamiento mediante máquinas computadoras.

La figura 6 presenta un ejemplo de aplicado de la extracción de la información. Se parte de un conjunto de informes realizados por la policía. De esos documentos se extrae la información relevante que contienen, en este caso persona, lugar, tipo de delito víctima y fecha y se almacena en una estructura de datos tipo tabla. Una vez que la información textual está estructurada, se puede analizar y visualizar, tal y como aparece en la última de las gráficas.

La extracción de la información también es clave en los buscadores de página web. Tener los contenidos de las páginas web almacenados con algún tipo de estructura permite la realización de búsquedas de alta velocidad.

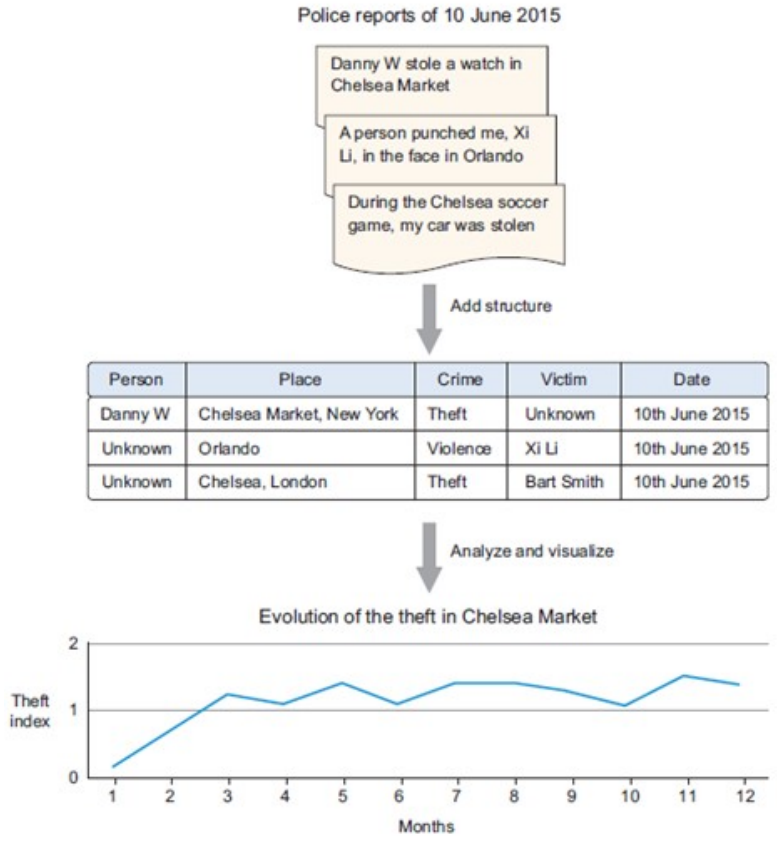


Figura 6. Ejemplo de extracción de la información es texto no estructurado. Fuente: (Cielen y otros, 2016)

8.4. Minería de textos. Aplicaciones

Las técnicas de minería de texto, en general, tratan de descubrir nueva información incluida en los textos no estructurados pero desconocida antes de su procesado. La minería de textos utiliza las técnicas básicas de procesamiento del lenguaje natural y las técnicas de aprendizaje automático (clasificación, clusterización, etc.) con el fin de extraer información no trivial de un conjunto de textos no estructurados. Esta información podrá ser utilizada por analistas o directivos para tomar decisiones en el ámbito de una organización (empresa o institución).

A nivel general, **el proceso de la minería de textos está dividido en dos grandes fases, el procesamiento y el aprendizaje automático.** En la fase de procesamiento el objetivo es obtener información estructura a partir de texto no estructurado. En esta fase se pueden aplicar técnicas básicas de análisis de texto o de procesamiento de lenguaje natural. La información estructurada obtenida se puede utilizar como entrada de la segunda fase donde se aplicarán técnicas de aprendizaje automático. Si se aplican técnicas de clasificación (aprendizaje supervisado) el objetivo será crear una aplicación informática capaz de discriminar contenidos textuales. Si se aplican técnicas de aprendizaje no supervisado (clusterización o reducción de la dimensión) el objetivo será descubrir relaciones latentes existentes en la información contenida en los textos.

Aplicaciones de la minería de textos

Esta sección presenta algunas aplicaciones prácticas de la minería de texto. No pretende ser una lista exhaustiva, sino, un conjunto de ejemplos para que el lector tenga cierta visión de la potencialidad de la minería de textos.

Clasificación o categorización automática de documentos

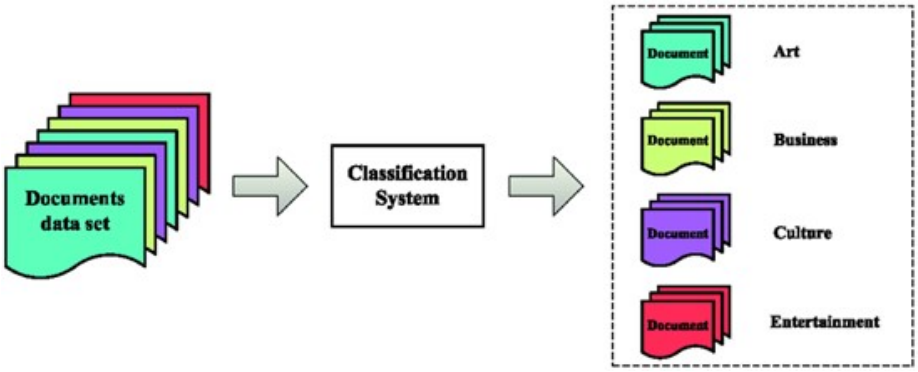


Figura 7. Categorización de textos. Fuente: https://www.researchgate.net/figure/A-simple-example-algorithm-framework-for-text-categorization_fig11_314161864

Dado un conjunto de documentos de texto, la categorización de texto trata de asignarle una categoría a cada uno de ellos (ver figura 7). Una de las principales utilidades de la clasificación de documentos es el filtrado o identificación de correos no interesantes o spam. La categorización se realiza a partir de los contenidos de los documentos y las etiquetas se seleccionan a partir de un conjunto de categorías previamente establecidas. Se suelen usar técnicas de aprendizaje supervisado para realizar la clasificación.

Un ejemplo de proceso para la creación de un sistema de clasificación sería el siguiente. Dado un conjunto de documentos de texto no estructurados, se aplicarían técnicas básicas de procesamiento del lenguaje natural para extraer información estructurada delos textos. A esta información estructurada se la puede considerar las variables independientes del sistema clasificador. A cada una de las observaciones de las variables independientes se le asignaría una etiqueta o clase, siendo ese conjunto de etiquetas la variable dependiente. Estos son los ejemplos con los que se entrenaría un clasificador (basados en alguna de las técnicas mostradas en el tema «Técnicas de aprendizaje supervisado», por ejemplo, una red neuronal o un regresor lineal). Una vez entrenado el clasificador, este se puede usar para clasificar nuevos

documentos.

Por supuesto, cada nuevo documento a clasificar hay que aplicarle el preprocesamiento que extraiga las observaciones de las variables dependientes que le corresponden a ese documento.

Agrupamiento o clusterización de documentos

Dado un conjunto de documentos de texto, puede ser interesante agruparlos por ejemplo por temas comunes que traten. La **clusterización de documentos** consiste en dividir un conjunto de documentos en subgrupos con el objeto de entender cómo se organiza la información de dichos documentos. La clusterización de documentos se diferencia de la clasificación en que los subgrupos o categorías no están predefinidos. Por tanto, hay que aplicar técnicas de aprendizaje no supervisado. La figura 8 presenta un ejemplo de aplicación de la clusterización. Se ha realizado una búsqueda de página con la palabra «jaguar». El resultado son 208 documentos. Sin embargo, la palabra «jaguar» tiene cierta ambigüedad semántica. Con el objeto de ayudar al usuario en la parte superior izquierda se presenta el resultado de una clusterización de los documentos. Además, a cada subgrupo de documentos se le asigna un tema que dirige la búsqueda del usuario.

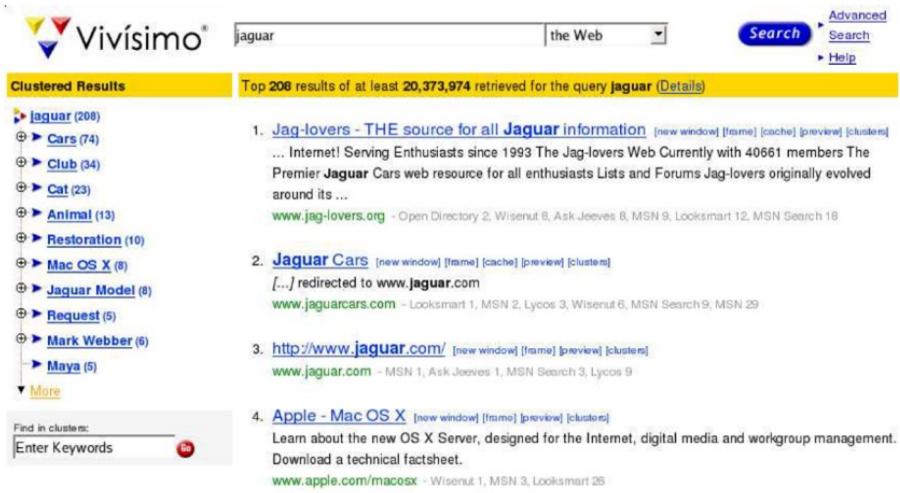


Figura 8. Clusterización de textos. Fuente: https://nlp.stanford.edu/IR-book/pdf/16flat.pdf

Desde el punto de vista operativo, normalmente las operaciones de clusterización necesitan de información estructurada. Por tanto, antes de realizar la clusterización en sí, es necesario aplicar técnicas de procesamiento de lenguaje natural para obtener información estructurada del texto no estructurado.

Resumen automático

Dada una fuente de información (documento o conjunto de documentos), el **resumen automático** consiste en extraer y presentar al usuario el contenido más importante de la fuente de información, de forma condensada y adaptado a las necesidades de la aplicación o del usuario.

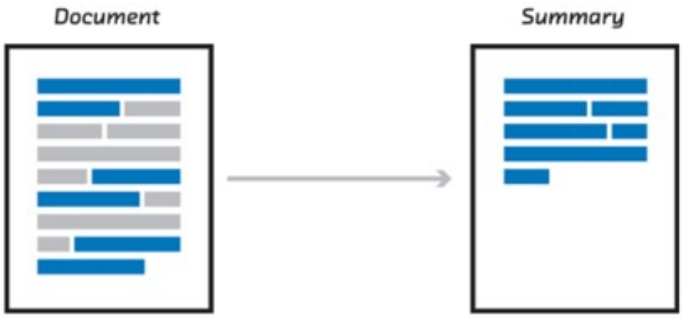


Figura 9. Ejemplo de resumen automático. Fuente: (Cielen y otros, 2016)

Las dos principales técnicas de creación de resúmenes automáticos son las extractivas y las abstractivas. Las **técnicas extractivas** se basan en seleccionar un conjunto de frases u oraciones del texto original (ver figura 9). El objetivo es eliminar aquellas frases que puedan ser de algún modo redundantes en el texto. Las **técnicas abstractivas** se basan en técnicas de aprendizaje automático que en base a ejemplos logran simplificar textos.

Análisis de sentimientos/minería de opinión

Dado un texto, el análisis de sentimiento intenta clasificar las opiniones expresadas en lenguaje natural. Se intenta identificar la polaridad: positivo, negativo, neutro. Una utilidad clara del análisis de sentimiento es que permite **identificar la actitud de los consumidores ante una marca, producto o servicio.** La figura 10 muestra un conjunto de fases típicas de una aplicación de análisis de sentimiento, se comienza con el texto de entrada al que se le aplica técnicas básicas de procesamiento de lenguaje natural (tokenización, eliminación de palabras sin información, etc.). Las últimas fases consisten en una clasificación del texto una vez simplificado. La figura 11 muestra un ejemplo del resultado de un análisis de sentimiento para una marca.

Figura 10. Fases aplicadas al análisis de sentimiento. Fuente: https://medium.com/@tomyuz/a-sentiment-analysis-approach-to-predicting-stock-returns-d5ca8b75a42

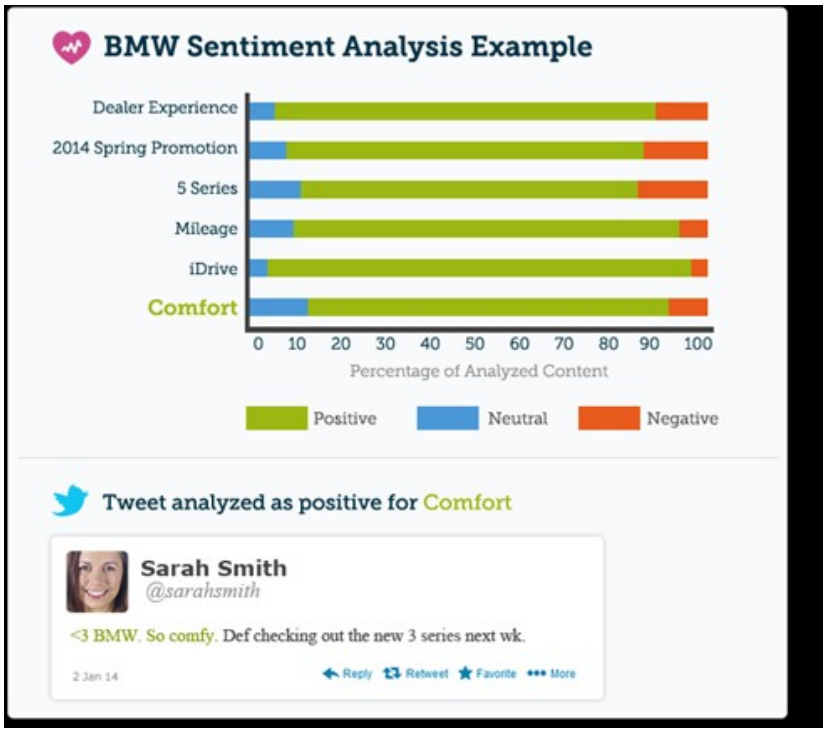


Figura 11. Ejemplo de salida de un estudio basado en análisis de sentimiento. Fuente: https://www.crowdsource.com/solutions/content-moderation/sentiment-analysis/

Con la expansión de las redes sociales (webs de opinión, mensajerías, etc.) ha

crecido exponencialmente la cantidad de opiniones y el interés de las organizaciones por averiguar qué se dice de ellas. La minería de opinión es una disciplina que estudia la extracción de opiniones usando técnicas de procesamiento de lenguaje natural. La minería de opinión trata de extraer las opiniones y sentimientos existentes en un texto. La diferencia entre minería de opinión y análisis de sentimiento puede ser una cuestión técnica.

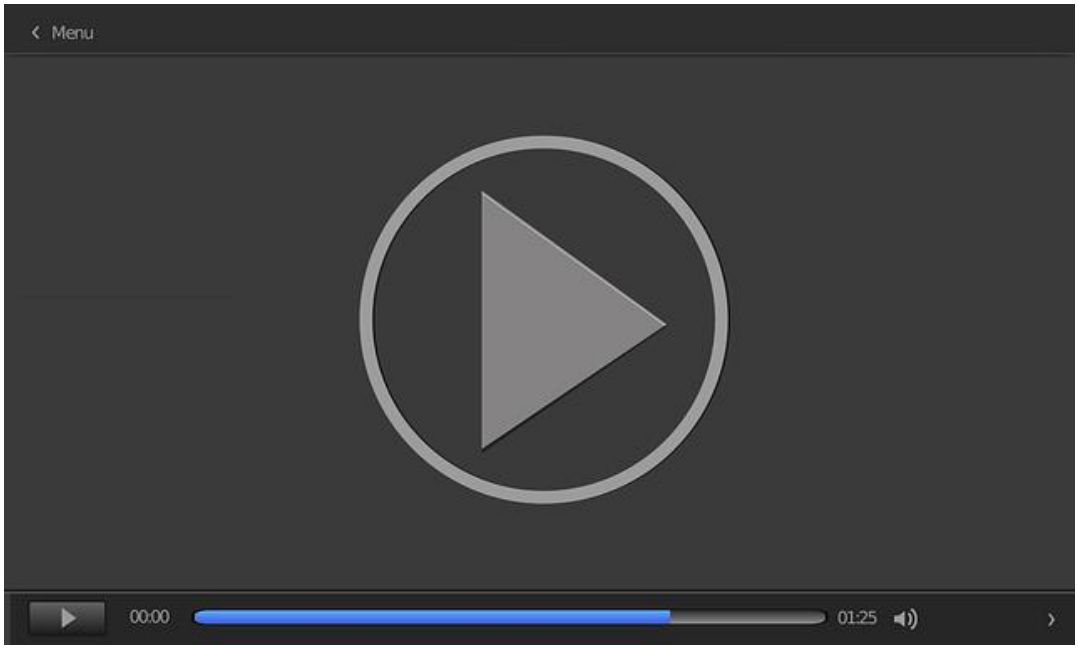
8.5. Referencias bibliográficas

Cielen D., Meysman A. D. B. y Ali M. (2016). *Introducing data science. Big data, machine learning, and more, using python tools.* Manning Publications Co.

Técnicas de procesamiento de lenguaje natural

Edureka (15 octubre 2018). *Natural Language Processing In 10 Minutes | NLP Tutorial For Beginners | NLP Training | Edureka* [Archivo de vídeo].

El vídeo recomendado resume en 10 minutos y de forma adecuada las técnicas de procesamiento de lenguaje natural.



Accede al vídeo:

https://www.youtube.com/embed/5ctbvkAMQO4

Analítica de textos

Monkeylearn (s. f.). Text Analysis.

Web interactiva muy completa que trata sobre la analítica de textos.

Accede al contenido a través del aula virtual o desde la siguiente dirección web: https://monkeylearn.com/text-analysis/

Test

1. La minería de textos busca fundamentalmente:

   A. Traducir textos.

   B. Contabilizar la frecuencia de aparición de palabras en el texto.

   C. Obtener información a partir de documentos de texto no estructurado.

   D. Ninguna de las anteriores es correcta.

2. El análisis léxico o tokenización de un documento de texto consiste en:

   A. Dividir la secuencia de caracteres en palabras o tokens.

   B. Eliminar palabras irrelevantes como artículos o preposiciones.

   C. Etiquetar cada palabra con su categoría sintáctica.

   D. Crear el árbol sintáctico de las oraciones del documento.

3. La lematización y el *stemming:*

   A. Son técnicas totalmente equivalentes.

   B. Pueden generar salidas diferentes para un mismo texto.

   C. Se aplica a textos estructurados.

   D. Ninguna de las anteriores es correcta.

4. La categorización o clasificación automática de documentos consiste en:

   A. Aplicar minería de texto para descubrir nuevas categorías.

   B. Asignar una categoría predefinida a cada documento.

   C. Obtener un resumen del documento.

   D. Ninguna de las anteriores es correcta.

5. Un fichero de índices inversos:

A. Se utiliza para invertir el orden de los textos.

B. Permite realizar búsquedas de términos no incluidos en los textos.

C. Permite acelerar las búsquedas en textos.

D. Ninguna de las anteriores es correcta.

6. ¿Cuál de los siguientes es un obstáculo importante para el procesamiento del lenguaje natural?

A. La falta de datos.

B. La ambigüedad y complejidad del lenguaje humano.

C. El exceso de puntuación en los textos.

D. La longitud variable de los documentos.

7. ¿Qué técnica permite reducir la cantidad de tokens en un texto sin perder demasiado significado?

A. Tokenización.

B. Creación de árboles sintácticos.

C. Normalización del texto.

D. Análisis de sentimiento.

8. ¿Qué tipo de resumen extrae directamente frases del documento original?

A. Abstractivo.

B. Automático.

C. Estadístico.

D. Extractivo.

9. ¿Cuál es una aplicación típica del análisis de sentimientos?

A. Clasificar documentos por tema.

B. Detectar la actitud positiva o negativa hacia un producto en redes sociales.

C. Agrupar documentos con contenido similar.

D. Generar índices de búsqueda.

10. ¿Qué estructura se crea al etiquetar las funciones gramaticales de los tokens?

A. Token.

B. Lema.

C. Árbol sintáctico.

D. Vector de frecuencias.