Pontificia Universidad JAVERIANA Bogotá

Introducción a la Minería de Texto

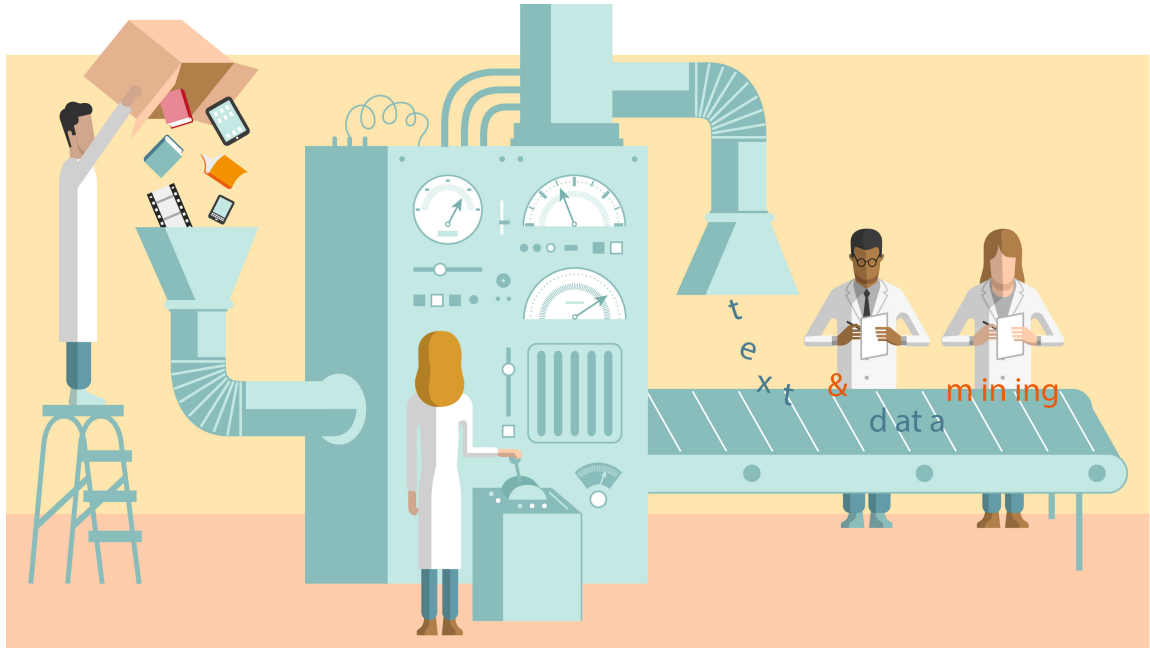
ING. HERIBERTO FELIZZOLA JIMENEZ

Educación **Continua**

Generamos experiencias educativas

Minería de Texto / Analítica de Texto

- La minería de texto es el proceso de extraer conocimiento e información de grandes cantidades de datos de texto.
- Estos datos pueden provenir de una variedad de fuentes, como sitios web, redes sociales, documentos, etc.
- La minería de texto es una herramienta importante para la analítica de datos porque nos permite comprender mejor el contenido de los datos de texto.
- Esto puede ayudarnos a:
  - Identificar patrones y tendencias en el texto.
  - Agrupar documentos según sus similitudes o tópicos
  - Realizar segmentación de clientes
  - Clasificar documentos en diferentes categorías.



Aplicaciones de la Mineria de Texto

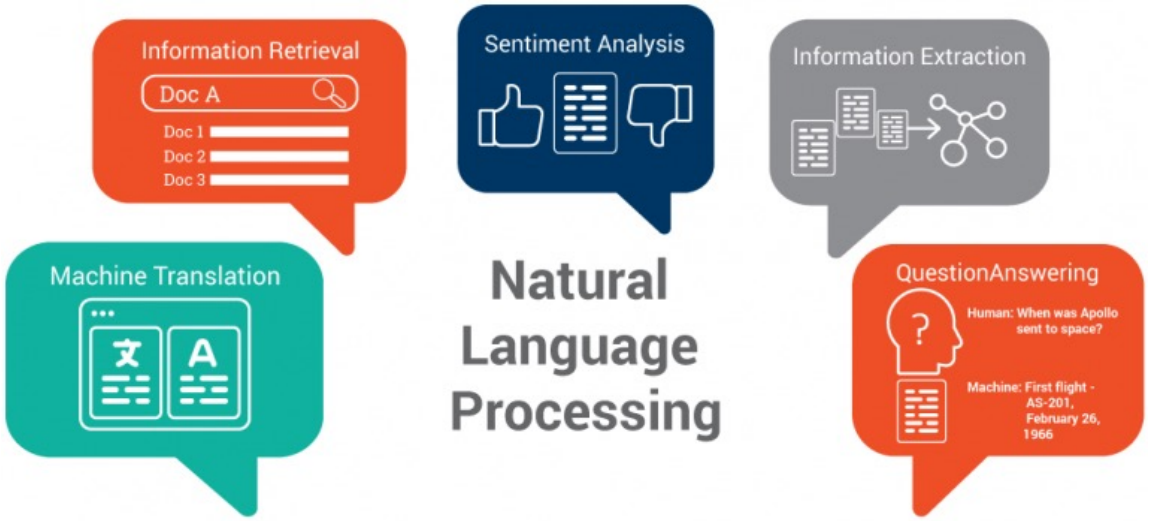
- Servicio al cliente: la minería de texto se puede utilizar para analizar opiniones y comentarios de los clientes para identificar problemas comunes y áreas de mejora.
- Marketing: la minería de texto se puede utilizar para analizar datos de redes sociales para identificar tendencias y comprender el sentimiento de los clientes. Esta información se puede utilizar para desarrollar campañas de marketing más efectivas.
- Servicios médicos: la minería de textos se puede utilizar para extraer información de registros médicos y datos de ensayos clínicos para identificar nuevos tratamientos y mejorar la atención al paciente.
- Investigación académica: la minería de textos se puede utilizar para analizar grandes volúmenes de artículos y libros académicos para identificar nuevas direcciones y tendencias de investigación.

Procesamiento de Lenguaje Natural (NLP)

El procesamiento del lenguaje natural (NLP) es una disciplina de la informática que se ocupa de la interacción entre las computadoras y el lenguaje humano. El objetivo del PNL es desarrollar sistemas informáticos que puedan entender y generar texto humano.

El NLP se puede dividir en dos subcampos principales:

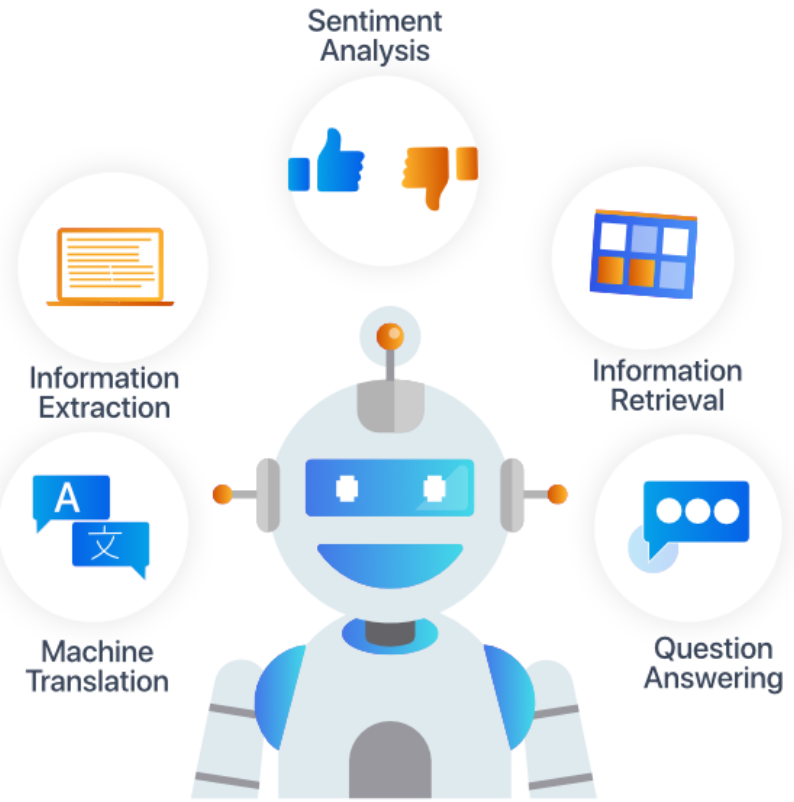
- **Comprensión del lenguaje natural (NLU)**: La NLU se centra en la comprensión del significado del lenguaje humano. Esto puede incluir tareas como la traducción automática, la clasificación de texto y la respuesta a preguntas.
- **Generación del lenguaje natural (NLG)**: La NLG se centra en la generación de texto humano. Esto puede incluir tareas como la escritura de artículos, la creación de chatbots y la traducción automática inversa.



Aplicaciones del NLP

El NLP se utiliza en una variedad de aplicaciones, como:

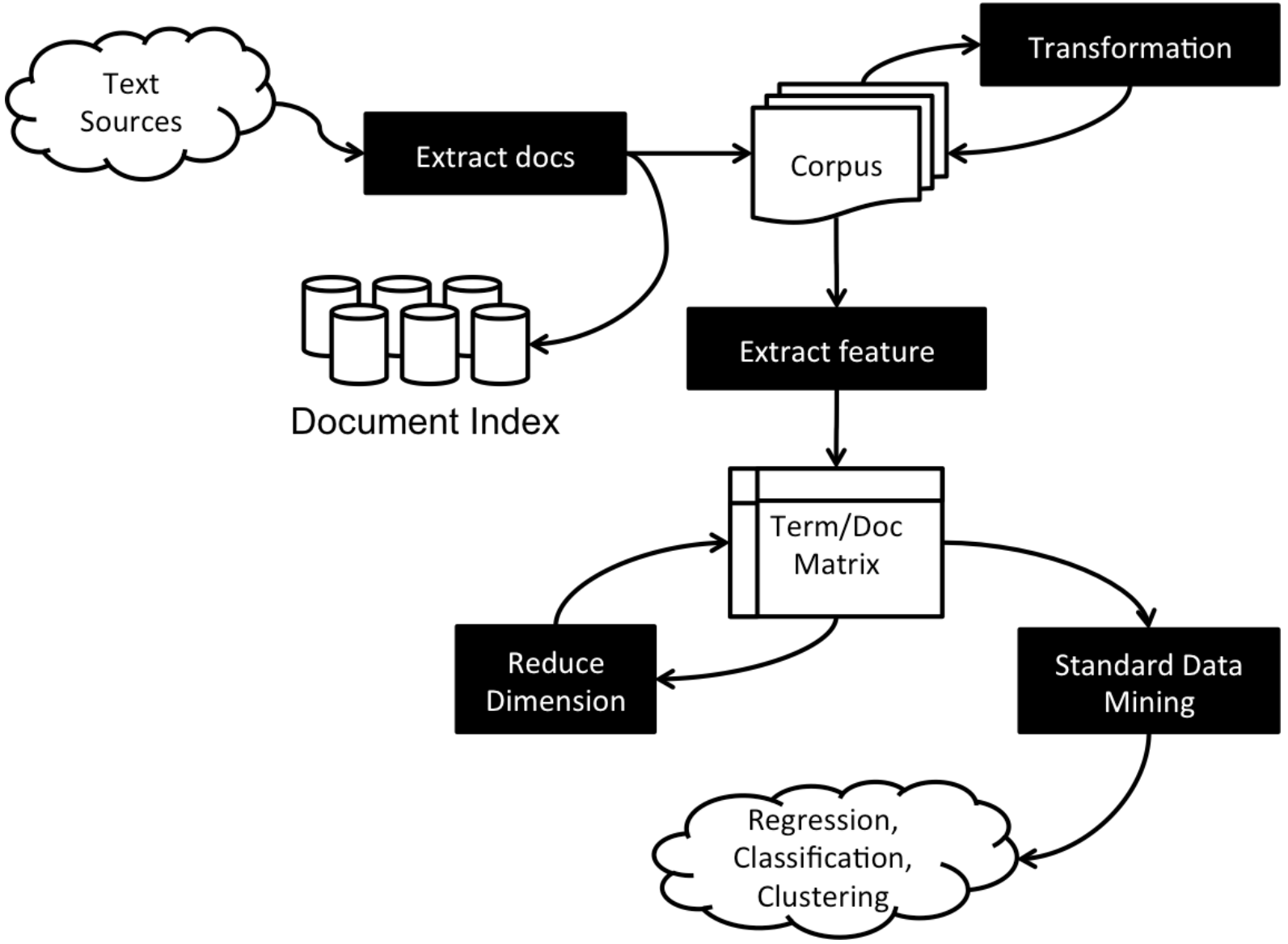
- **Traducción automática:** Las máquinas pueden traducir idiomas utilizando el NLP.
- **Búsqueda de información:** Las máquinas pueden comprender el significado de las consultas de búsqueda utilizando el NLP.
- **Asistentes virtuales:** Los asistentes virtuales, como Siri y Alexa, utilizan el NLP para entender y responder a las solicitudes de los usuarios.
- **Inteligencias artificiales generativas:** permiten generar documentos, textos, códigos de programación. Estas incluyen aplicaciones tales como Chatgpt, Bard, Github, Bing, Copilot.



Mineria de Texto Vs NLP

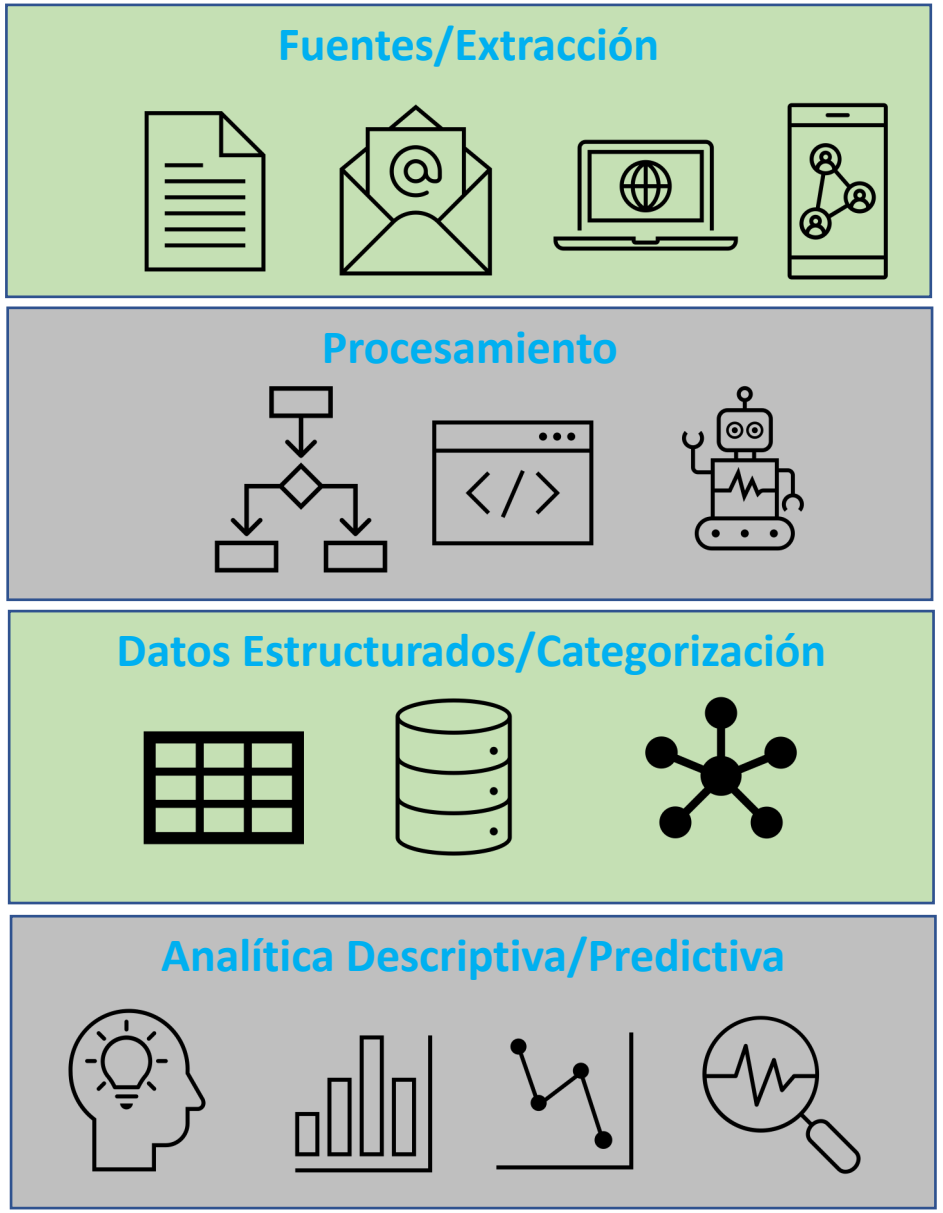
Característica	Minería de texto	Procesamiento del lenguaje natural
Definición	La minería de texto es el proceso de extraer información significativa de grandes cantidades de texto.	El procesamiento del lenguaje natural (NLP) es la disciplina de la informática que se ocupa de la interacción entre las computadoras y el lenguaje humano.
Objetivo	El objetivo de la minería de texto es extraer información relevante de grandes cantidades de texto.	El objetivo del PNL es desarrollar sistemas informáticos que puedan entender y generar texto humano.
Técnicas	Las técnicas de minería de texto incluyen el análisis de frecuencia de palabras, el análisis de temas y el análisis de sentimientos.	Las técnicas de NLP incluyen el reconocimiento de palabras, el análisis sintáctico y el análisis semántico.
Aplicaciones	La minería de texto se utiliza en una variedad de aplicaciones, como la búsqueda de información, la clasificación de texto, el análisis de tópicos y sentimientos.	El PNL se utiliza en una variedad de aplicaciones, como la traducción automática, la comprensión del lenguaje natural y la generación de texto.
Relación	La minería de texto es un subcampo del procesamiento del lenguaje natural.	La minería de texto se basa en el procesamiento del lenguaje natural para extraer información de texto.

Proceso de Minería de Texto



Insumos/Tareas de Minería de Texto

- Fuentes: principalmente documentos de texto que pueden ser informes, páginas web, libros, artículos, chats, comentarios redes, opiniones, reseñas, o transcripciones de audio.
- Extracción: consiste en recolectar los documentos de sus fuentes y llevarlos a texto plano que permitan su procesamiento. Ejemplo: web scraping, Web Crawling, APIs para redes sociales, bases de datos estructuradas, digitalización de documentos.
- Procesamiento: transformar los documentos en estructuras que permitan y faciliten la aplicación de técnicas de análisis de datos. Incluye la limpieza, normalización, estructuración y categorización del texto.
- Datos estructurados: son representaciones de los documentos en tablas, vectores y otro tipo de representación que permita su análisis.
- Analítica de texto:
  - Analítica descriptiva: frecuencia de palabra, longitud de texto.
  - Clasificación: análisis de sentimiento, clasificación de correos, notificas falsas, categorización de documentos.
  - Clusterización: segmentación de documentos, modelado de tópico.



Tareas Básicas de Procesamiento de Texto

¿Qué es un Corpus?

- En el contexto de la analítica de texto, un corpus es una colección de textos que se utilizan para entrenar y evaluar modelos de procesamiento de lenguaje natural.
- Los corpus pueden ser de cualquier tamaño, desde unos pocos documentos hasta millones de documentos.
- Los corpus pueden estar compuestos de una variedad de tipos de texto, como noticias, artículos científicos, libros, correos electrónicos, etc.
- El tipo de texto que se utiliza en un corpus depende de la tarea de analítica de texto que se esté realizando.
- Por ejemplo:
  - si se está realizando un análisis de sentimiento, el corpus podría estar compuesto de reseñas de productos o servicios.
  - Si se está realizando una clasificación de texto, el corpus podría estar compuesto de noticias de diferentes categorías.

¿Cómo convertimos datos no estructurados (informes, correos, comentarios de clientes, historias clínicas, artículos, información de una pagina web) en datos extructurados útiles para analizar?

Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to

Normalización

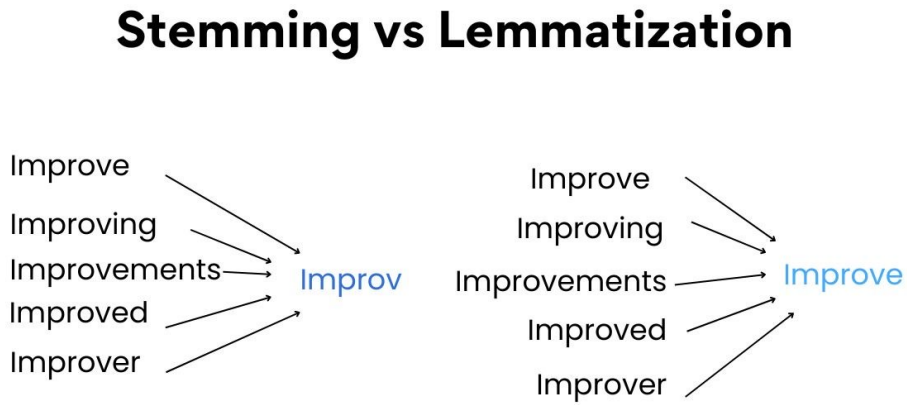
¿Qué es la normalización?

Es el proceso de transformar texto para que tenga una forma estandarizada.

Esto puede implicar una serie de tareas, como la corrección de errores ortográficos y gramaticales, la unificación del uso de mayúsculas y minúsculas, la eliminación de caracteres no deseados, y la conversión de palabras a un formato estándar.

Técnicas de normalización

- **Conversión a minúsculas:** Convertir todas las letras a minúsculas ayuda a garantizar que las palabras se comparen correctamente, independientemente de su caso.
- **Lematización:** Reducir las palabras a su forma base, independientemente de su conjugación o forma flexionada. Por ejemplo, "comer", "comiendo" y "comido" se convierten en "comer".
- **Stemming:** Recortar los sufijos de las palabras para reducirlas a sus formas fundamentales. Por ejemplo, "comer", "comiendo" y "comido" se convierten en "comer".



Eliminación de Caracteres Especiales

En el preprocesamiento de texto, la "Eliminación de Caracteres Especiales" es una tarea esencial para limpiar y preparar datos de texto para análisis.

¿Por qué se deben eliminar caracteres especiales?

Caracteres no alfabéticos, como signos de puntuación, números y símbolos, pueden agregar ruido innecesario a los datos de texto, dificultando su análisis efectivo.

Tareas en la Eliminación de Caracteres Especiales:

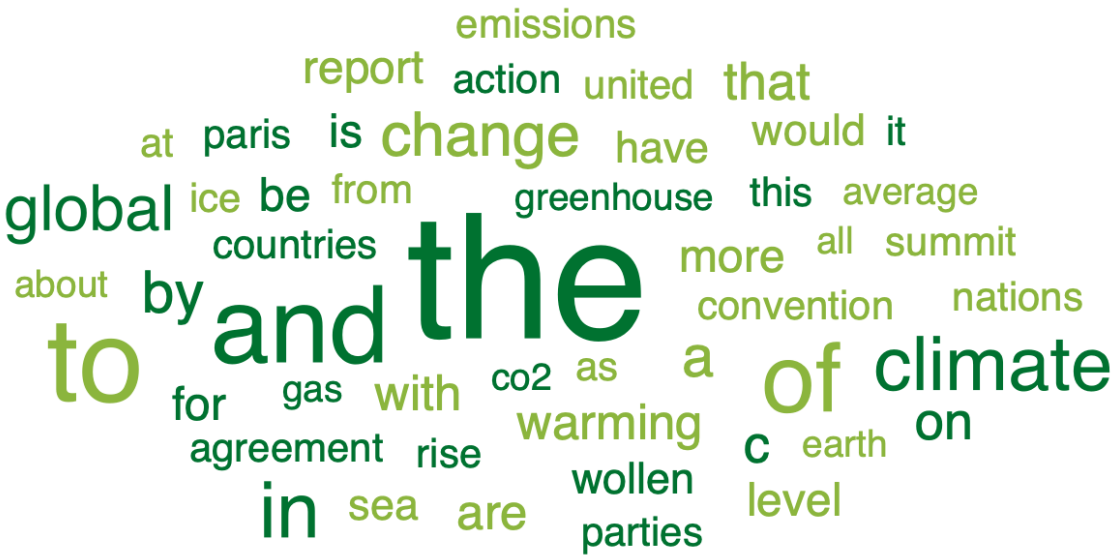
- Identificación de Caracteres no Alfabéticos: Se buscan caracteres que no sean letras o espacios en blanco en el texto.
- Eliminación de Caracteres no Deseados: Los caracteres no alfabéticos, como comas, puntos, números y otros símbolos no deseados, se eliminan del texto.

Herramientas Útiles: Expresiones Regulares (Regex)

- Las expresiones regulares son patrones textuales que permiten identificar y reemplazar caracteres y patrones específicos en el texto de manera eficiente.
- Ejemplo:
  - Texto original: texto = **"Hola, ¿cómo estás? Esto es un ejemplo de eliminación de caracteres especiales!!!"**
  - Regex(python) : **texto_limpio = re.sub(r'[^\w\s]', '', texto)**
  - La expresión regular [^\w\s] coincide con cualquier caracter que no sea una letra o un espacio en blanco. El reemplazo se realiza por una cadena vacía, lo que efectivamente elimina esos caracteres especiales.
  - Texto limpio: “Hola cómo estás Un ejemplo de eliminación de caracteres especiales”

Eliminación de Stop Words

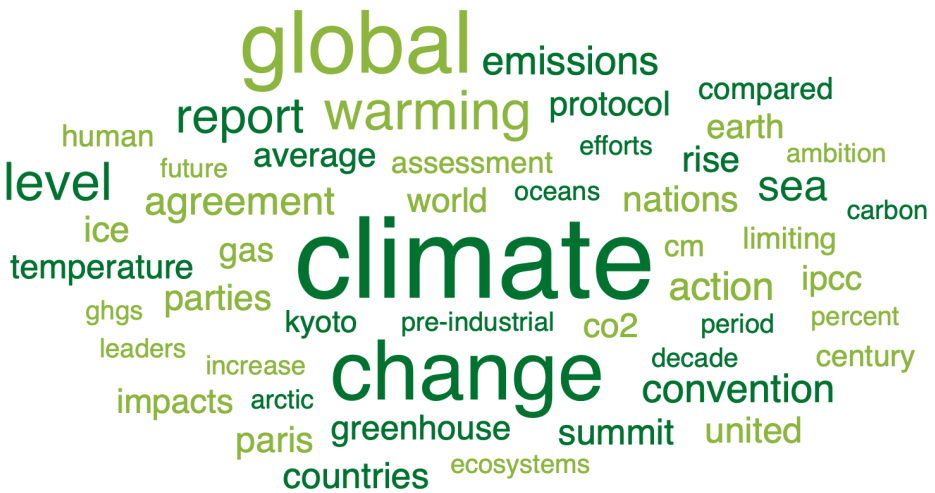
- En el preprocesamiento de texto, la "Eliminación de Stop Words" es una tarea esencial que implica la eliminación de palabras comunes que no aportan un significado sustancial al texto.
- Las "Stop Words" son palabras que aparecen con alta frecuencia en el texto, pero que generalmente no aportan información relevante para el análisis de contenido.



- Ejemplos de "Stop Words" incluyen palabras como "el," "la," "y," "en," "un," "una," "de," entre otras.

¿Cómo se eliminan?

- Identificación de "Stop Words": Se crea una lista de palabras que se consideran "Stop Words" en el idioma de trabajo.
- Eliminación de "Stop Words": Estas palabras se eliminan del texto para reducir el ruido y mejorar la calidad del análisis de texto.



Tokenización

¿Qué es la tokenización?

La tokenización es una tarea de procesamiento de lenguaje natural que consiste en dividir un texto en unidades más pequeñas, llamadas tokens.

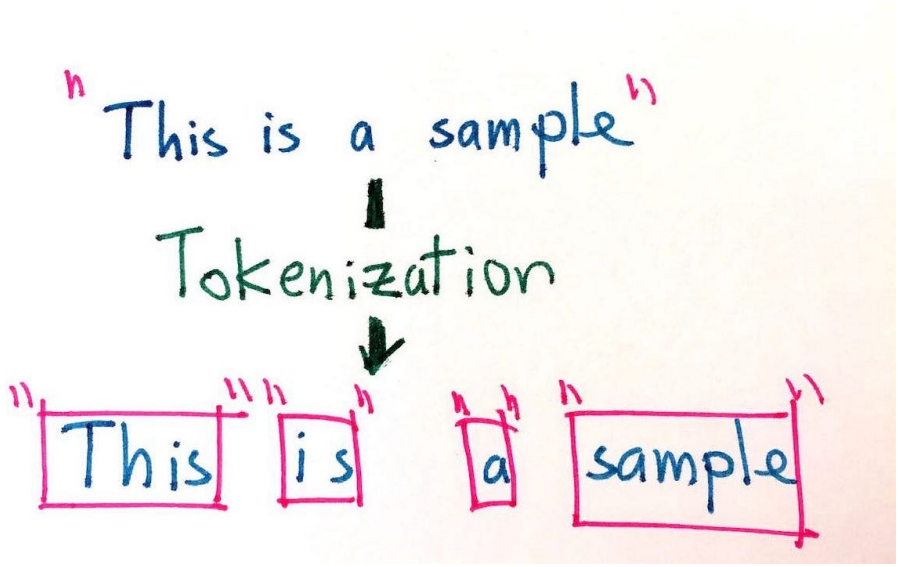
Los tokens pueden ser palabras, frases o incluso caracteres.

Al dividir el texto en tokens, se facilita el análisis del texto y la extracción de información.

Tipos de tokenización

Existen diferentes tipos de tokenización, que pueden clasificarse según el criterio utilizado para dividir el texto en tokens.

- **Tokenización léxica:** Se basa en la separación del texto en palabras, utilizando criterios como el espacio en blanco, los signos de puntuación o las reglas gramaticales.
- **Tokenización morfológica:** Se basa en la separación del texto en unidades morfológicas, como palabras, raíces, prefijos y sufijos.
- **Tokenización semántica:** Se basa en la separación del texto en unidades semánticas, como palabras, frases o incluso oraciones.



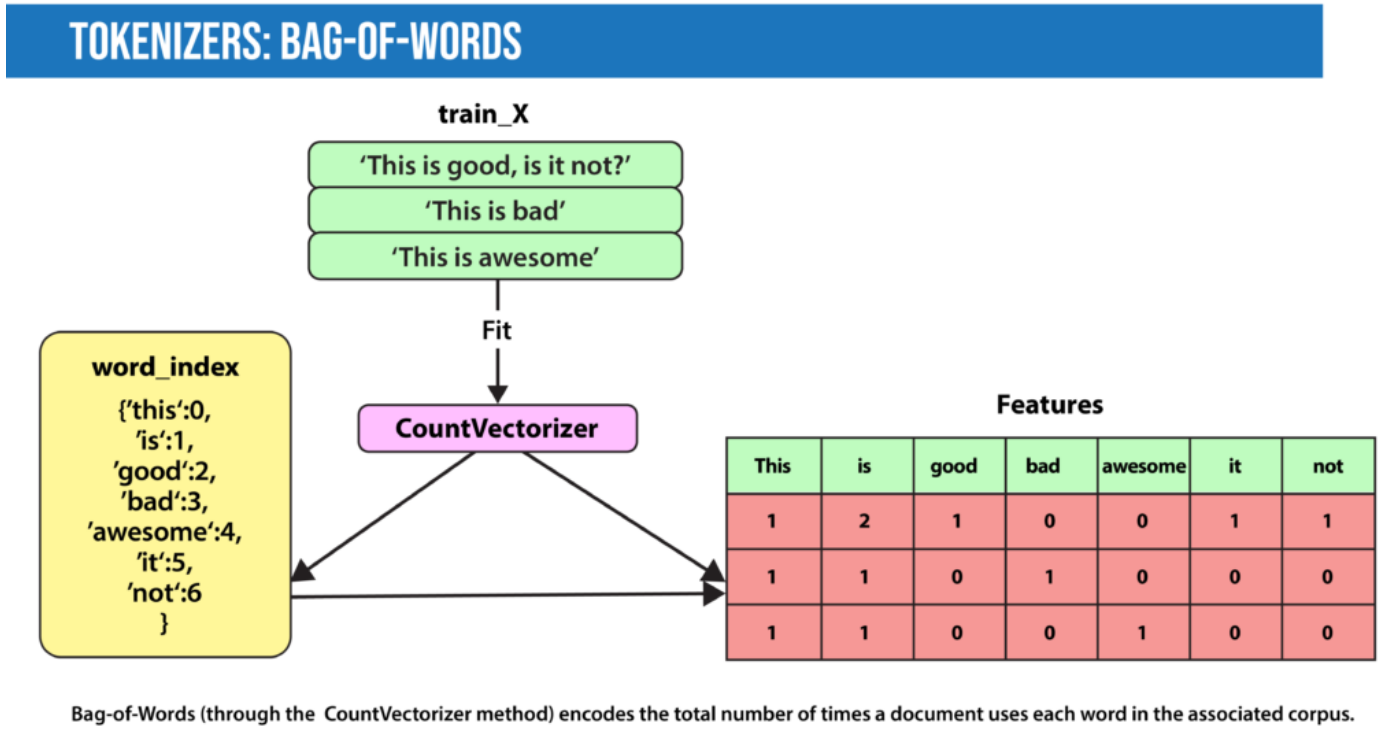
Tokenización

Text_id	Text	tokens
23	All roads lead to Rome	(All, roads, lead, to, Rome)
21	All work and no play makes Jack a dull buy	(All, work, and, no, play, makes, Jack, a, dull, buy)
22	Any port in a storm	(Any, port, in, a, storm)
21	Avoid a questioner, for he is also a tattler	(Avoid, a, questioner, ,, for, he, is, also, a, tattler)

Tareas Avanzadas de Procesamiento de Texto

Bolsa de Palabras (Bag-of-Words)

- Representa un documento como una colección no ordenada de palabras o n-grama.
- Ignora la estructura gramatical y el orden de las palabras en el texto.
- Cuenta la frecuencia de aparición de cada palabra única en el documento.
- Por ejemplo, en el siguiente caso, el modelo de Bolsa de Palabras genera una representación numérica del conjunto de datos basada en la frecuencia de aparición de cada palabra presente en el índice de palabras dentro del documento.



Fuente. deeplearning.ai

Frecuencia de Término – Frecuencia Inversa de Documento

- **TF-IDF** (Term Frequency-Inverse Document Frequency) es una técnica que asigna un peso a cada palabra en función de su importancia.
- Para evaluar la importancia de una palabra, se consideran dos factores clave: **Frecuencia de Término (TF)** y **Frecuencia Inversa de Documento (IDF)**.
- **TF** mide cuánto una palabra es importante en un documento específico, calculando el número de veces que aparece en ese documento en relación con el número total de palabras en el mismo.
- **IDF** mide cuánto una palabra es importante en el contexto del corpus entero, teniendo en cuenta cuántos documentos contienen la palabra en relación con el número total de documentos en el corpus.
- La combinación de TF y IDF se utiliza para determinar el peso de una palabra en un documento en comparación con su importancia en el corpus.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

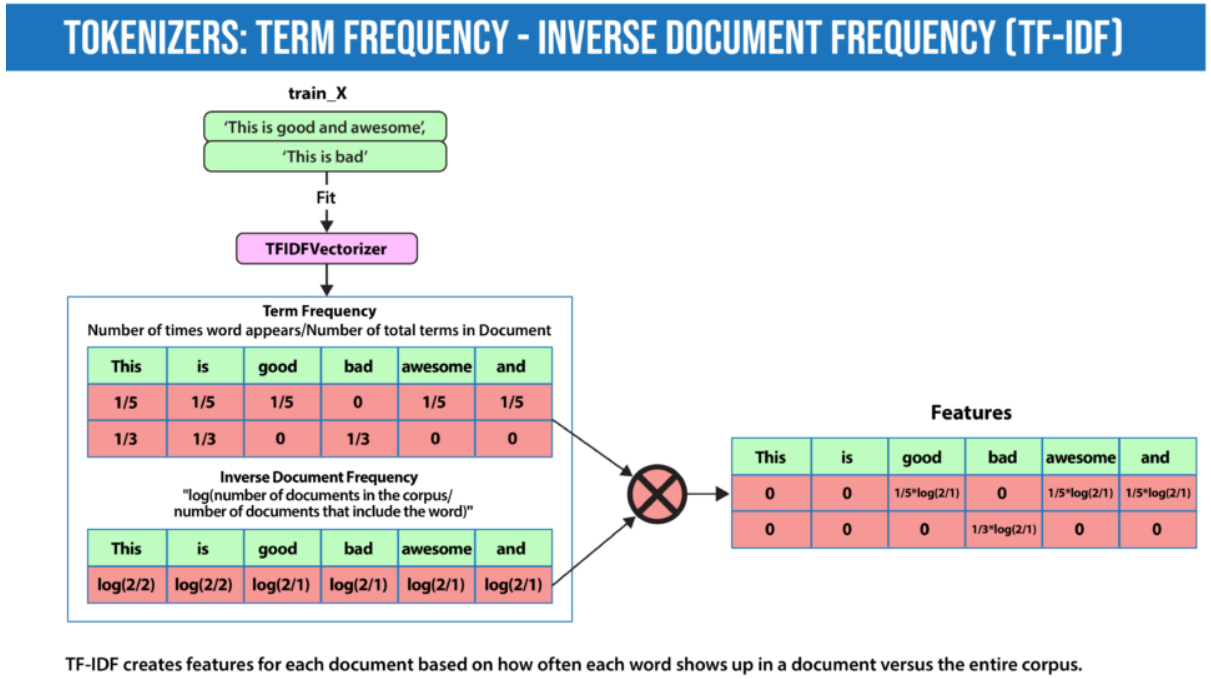
TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents



Recuerde que log(1) = 0

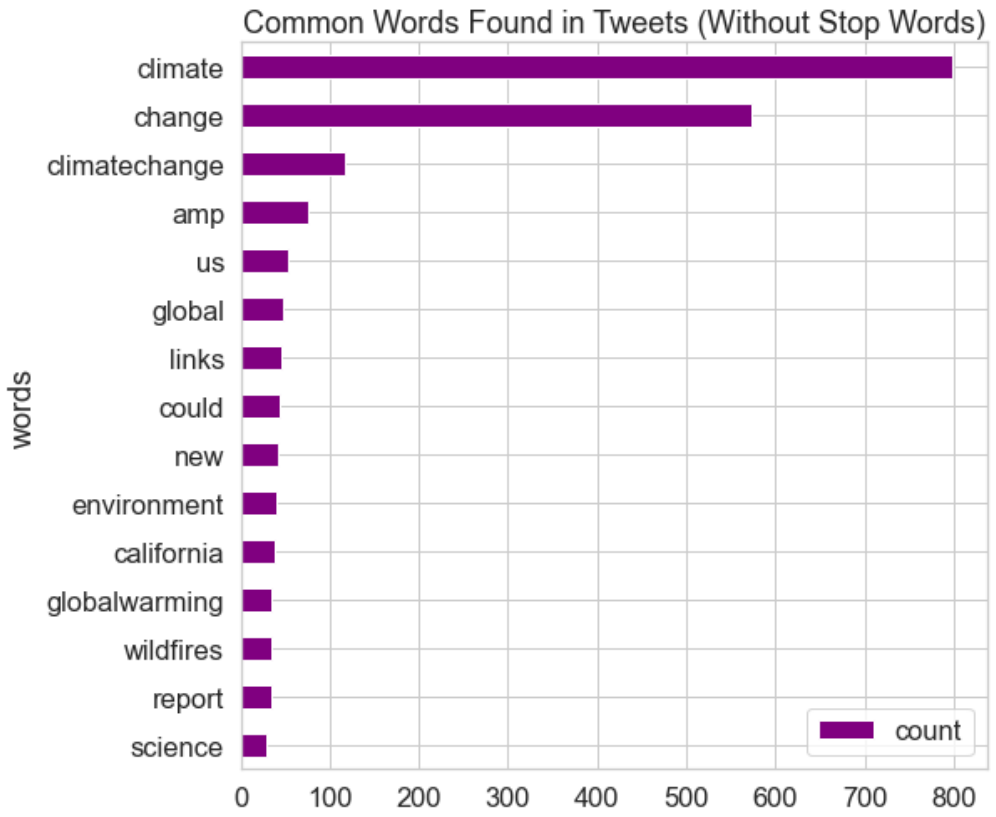
Analítica Descriptiva de Texto

Análisis de frecuencia de palabras

- El análisis de frecuencia de palabras es una tarea de procesamiento de lenguaje natural que consiste en calcular la frecuencia de ocurrencia de las palabras en el texto.
- La frecuencia de una palabra se define como el número de veces que aparece en el texto.
- Al calcular la frecuencia de las palabras, se puede identificar las palabras clave y términos más relevantes en el texto.

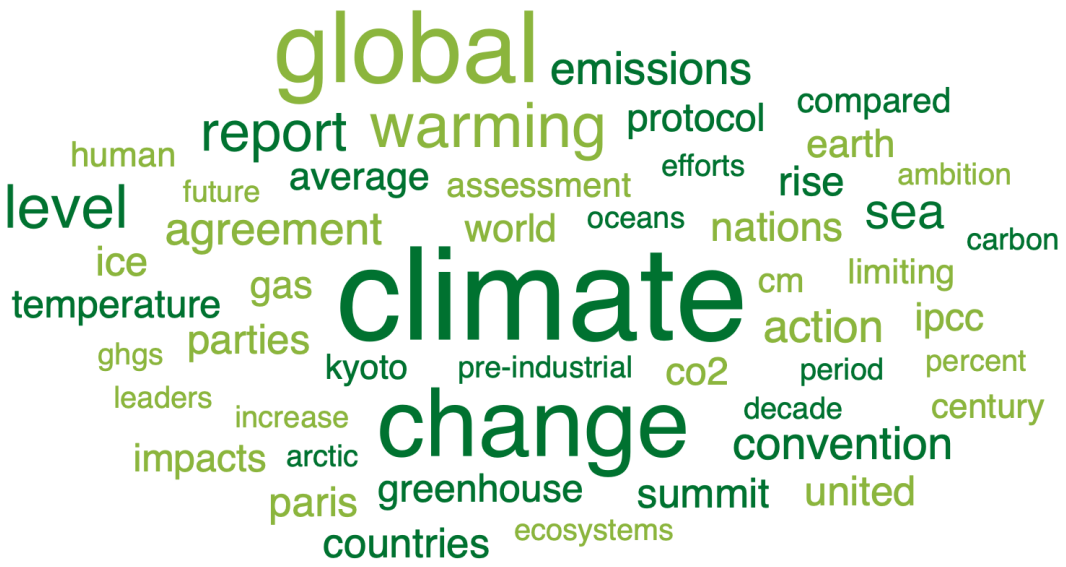
El análisis de frecuencia de palabras tiene los siguientes beneficios:

- **Identificación de palabras clave y términos relevantes:** Esto puede ser útil para tareas como la búsqueda de información, la clasificación de textos y la extracción de información.
- **Comprensión del tema del texto:** Esto puede ser útil para tareas como la generación de resúmenes y la traducción automática.
- **Análisis de tendencias:** Esto puede ser útil para tareas como la monitorización de redes sociales y la investigación de mercado.



Nube de Palabras (Wordcloud)

- Una nube de palabras es una representación visual de la frecuencia de las palabras en un texto.
- Las palabras más frecuentes se representan con un tamaño mayor, mientras que las palabras menos frecuentes se representan con un tamaño menor.
- Las nubes de palabras se pueden utilizar para una variedad de propósitos, como:
  - **Comprender el tema de un texto:** Las nubes de palabras pueden ayudar a identificar las palabras clave y términos más relevantes en un texto.
  - **Generar ideas:** Las nubes de palabras pueden ayudar a generar ideas para nuevos proyectos o productos.
  - **Comunicar información de forma visual:** Las nubes de palabras pueden ser una forma eficaz de comunicar información de forma visual.

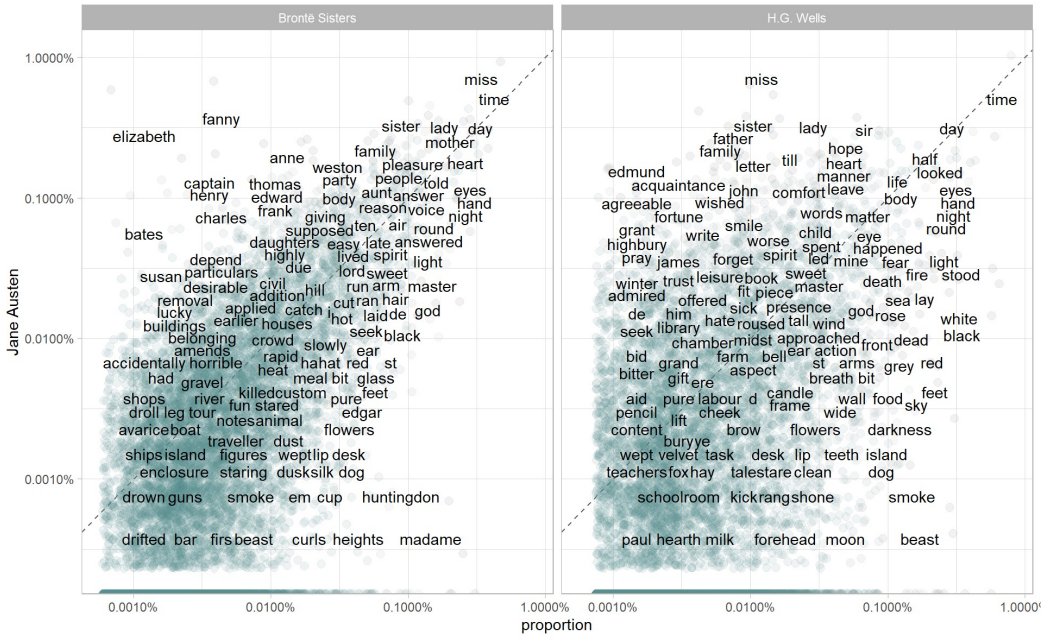


Análisis de correlación de corpus

El análisis de correlación de corpus es una técnica que se utiliza para medir la relación entre dos corpus de texto. El análisis se basa en la frecuencia de las palabras en los dos corpus.

El análisis de correlación de corpus se puede utilizar para una variedad de propósitos, como:

- **Comprender la relación entre dos temas:** puede ayudar a comprender la relación entre dos temas, comparando la frecuencia de las palabras que se relacionan con cada tema.
- **Detectar similitudes y diferencias entre dos corpus:** El análisis de correlación de corpus puede ayudar a detectar similitudes y diferencias entre dos corpus, comparando la frecuencia de las palabras que se encuentran en ambos corpus.
- **Identificar fuentes de texto:** El análisis de correlación de corpus puede ayudar a identificar fuentes de texto, comparando la frecuencia de las palabras que se encuentran en las fuentes.



Aprendizaje Supervisado/No Supervisado

Análisis de sentimientos

El análisis de sentimientos es una técnica de procesamiento de lenguaje natural que se utiliza para identificar y clasificar los sentimientos expresados en el texto.

El análisis de sentimientos se puede utilizar para una variedad de propósitos, como:

- Comprender la opinión pública: El análisis de sentimientos puede ayudar a comprender la opinión pública sobre un tema específico, analizando los sentimientos expresados en los medios de comunicación, las redes sociales y otros canales.
- Mejorar la experiencia del cliente: El análisis de sentimientos puede ayudar a mejorar la experiencia del cliente, analizando los sentimientos expresados por los clientes en reseñas, comentarios y otros medios.

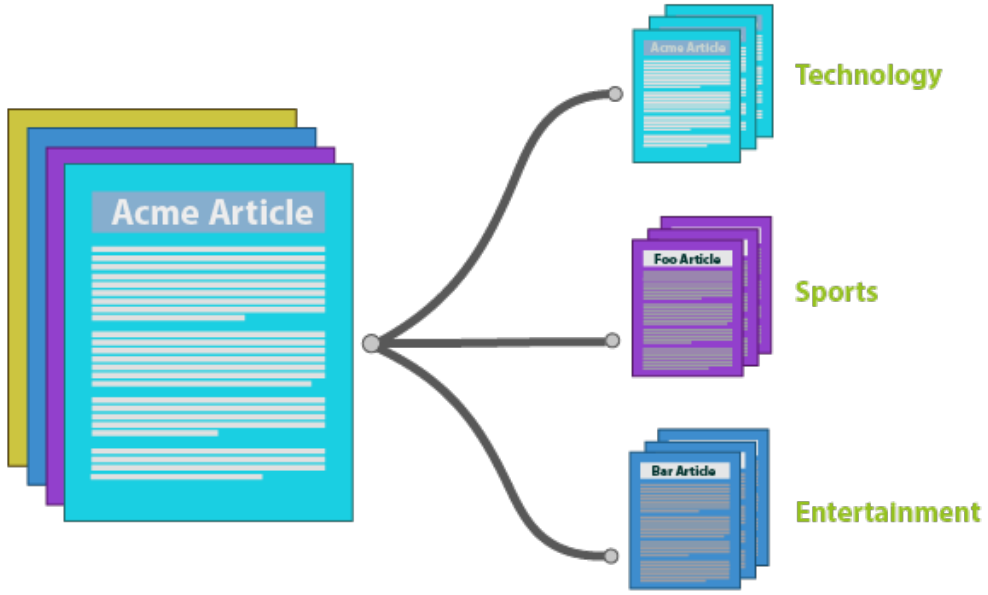
Clasificación de Documentos

La clasificación de documentos es una tarea de minería de texto que consiste en asignar a cada documento un conjunto de etiquetas o categorías.

Las etiquetas pueden ser temáticas, como "noticias", "deportes" o "ciencia"; o pueden ser más específicas, como "coronavirus", "elección presidencial" o "nuevo producto de Apple".

La clasificación de documentos se puede utilizar en una variedad de aplicaciones, como:

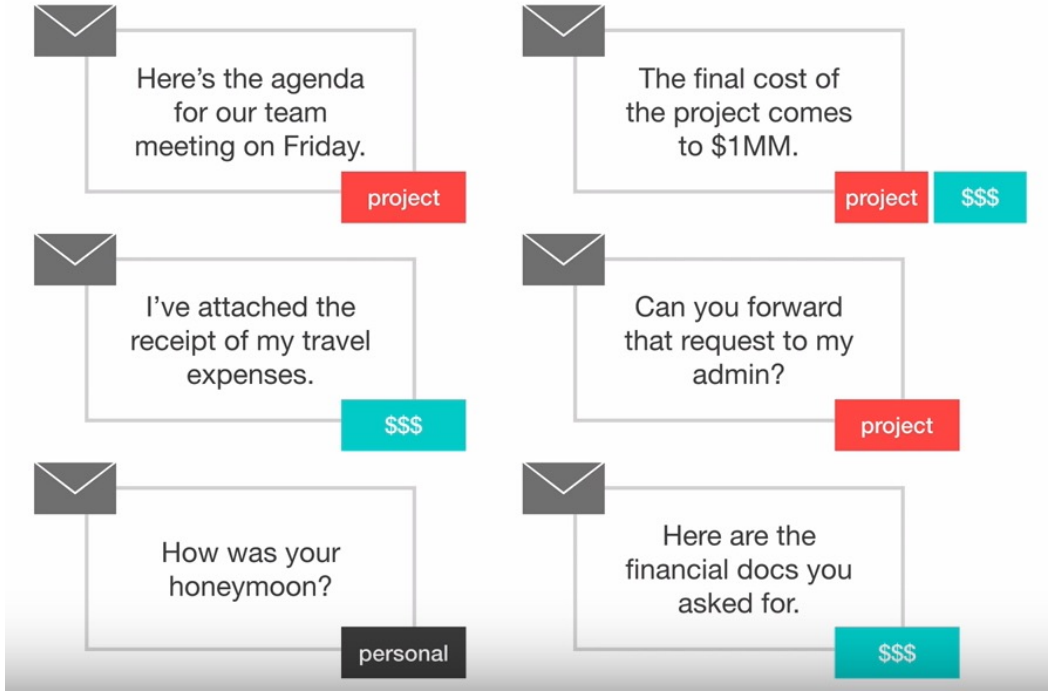
- **Organización de información:** La clasificación de documentos puede ayudar a organizar grandes cantidades de información, lo que facilita su búsqueda y recuperación.
- **Filtrado de información:** La clasificación de documentos puede ayudar a filtrar información irrelevante o no deseada.
- **Análisis de datos:** La clasificación de documentos puede ayudar a analizar datos para identificar tendencias y patrones.



Métodos de clasificación de documentos

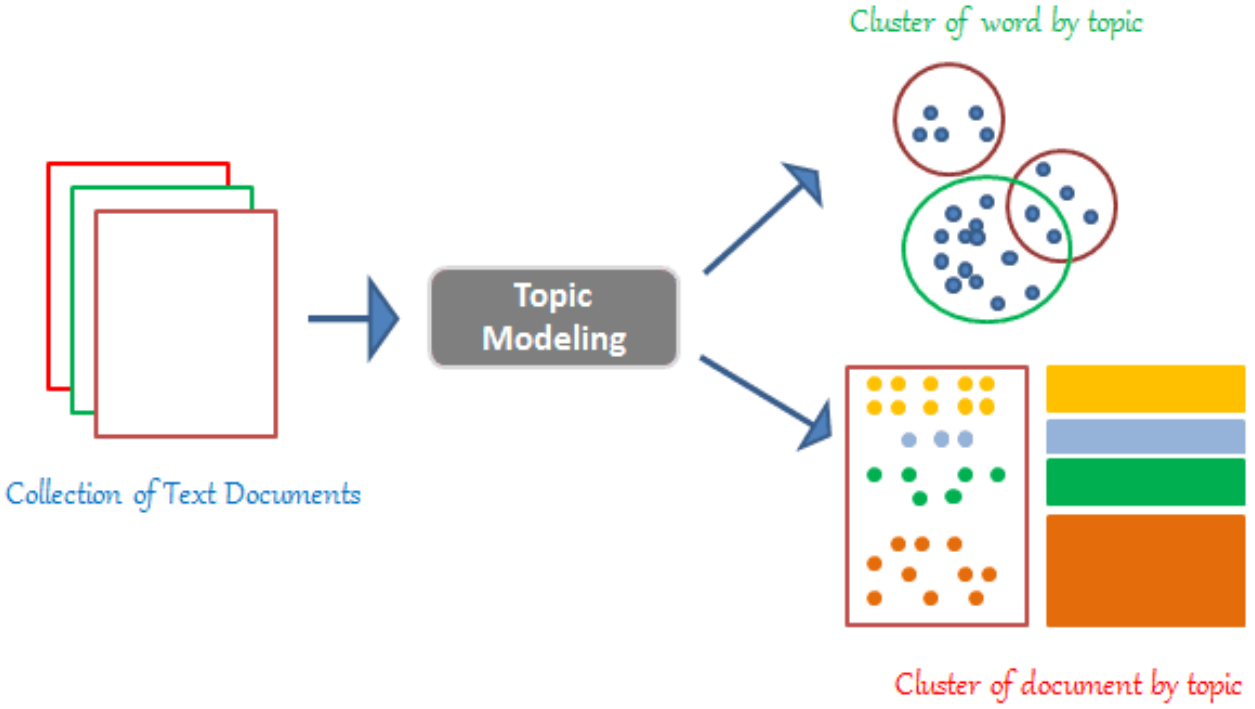
Existen varios métodos diferentes que se pueden utilizar para clasificar documentos. Algunos de los métodos más comunes son:

- Clasificación supervisada: En la clasificación supervisada, se utiliza un conjunto de datos de entrenamiento para entrenar un clasificador. El conjunto de datos de entrenamiento incluye documentos etiquetados con las categorías correctas.
- Clasificación no supervisada: En la clasificación no supervisada, no se utiliza un conjunto de datos de entrenamiento. El clasificador debe aprender a identificar las categorías de los documentos por sí mismo.



Modelado de Tópicos (Topic Modeling)

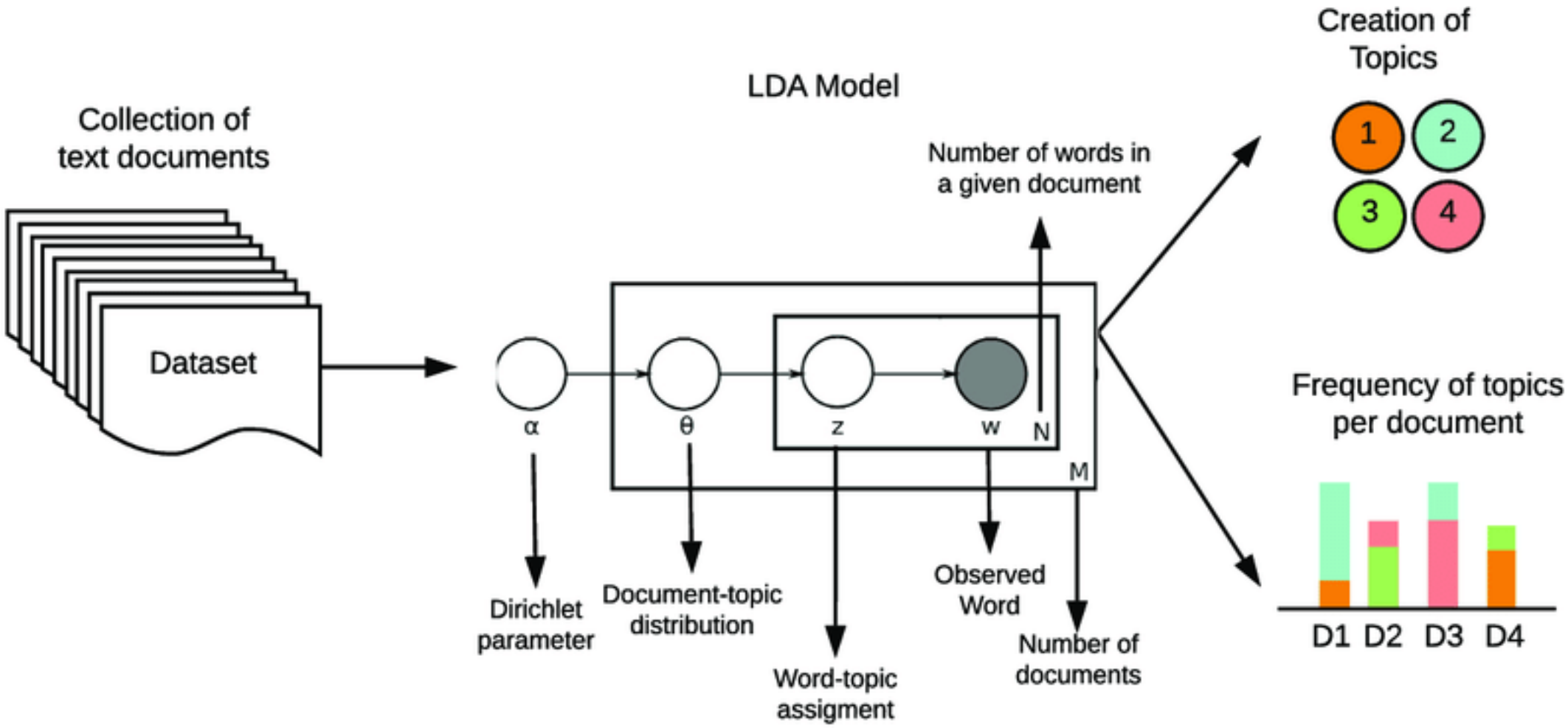
- El modelado de temas es un tipo de modelado estadístico que utiliza el aprendizaje automático no supervisado para identificar clusters o grupos de palabras similares dentro de un texto.
- El "Topic Modeling" se basa en la idea de que los documentos están compuestos por una mezcla de tópicos subyacentes.
- Cada tópico está formado por un conjunto de palabras clave que están relacionadas entre sí y son más frecuentes en los documentos que tratan ese tópico en particular.



- Latent Dirichlet Allocation (LDA) es una técnica de "Topic Modeling" ampliamente utilizada en procesamiento de lenguaje natural (NLP) y minería de texto.
- Fue propuesta por David Blei, Andrew Ng y Michael Jordan en 2003.
- La idea principal detrás de LDA es que cada documento es una mezcla de varios tópicos y que cada tópico es una mezcla de palabras clave.

Latent Dirichlet Allocation

¿Cómo funciona?



Latent Dirichlet Allocation

El topic modeling permite identificar los de términos o palabras que mayor relevancia tienen en cada tópico, con lo cual se puede hacer un identificación de los temas latentes.