Minería de texto

¿Qué es la minería de texto?

Según @Kwartler2017 esta se define como:

> **La minería de texto es el proceso de destilación de información procesable del texto**

Minería de texto puede ser sinónimo de *análisis de texto*, sin embargo, el uso de minería de texto describe de forma más adecuada el descubrimiento de ideas (**KDD**) y el uso de algoritmos específicos más allá del análisis estadístico básico.

Dónde se aplica la minería de texto?

> En cualquier parte donde se genere texto

¿Por qué es importante?

- Las **redes sociales** continúan evolucionando y las empresas e instituciones participan en menor o mayor medida en estos espacios
- **Contenido en línea** de una organización, sus competidores y fuentes externas, como los blogs, sigue creciendo.
- La **digitalización** de los antiguos registros en papel se está produciendo en muchas industrias, como la salud.
- Las **nuevas tecnologías**, como la transcripción automática de audio, ayudan a capturar información del cliente.
- A medida que las fuentes textuales crecen en cantidad, complejidad y número de fuentes, el avance simultáneo en el **poder de procesamiento y almacenamiento** se ha traducido en una gran cantidades de texto que se almacenan en todo el lago de datos de una empresa.

Las consecuencias de ignorarlo

- Ignorar el texto no es una respuesta adecuada de un *esfuerzo analítico*. La exploración científica y analítica rigurosa requiere investigar fuentes de información que puedan explicar los fenómenos.
- No realizar minería de texto puede conducir a un **análisis o resultado falso** (sesgo de confirmación).
- Algunos problemas se basan **casi exclusivamente en texto**, por lo que no usar estos métodos significaría una reducción significativa en la efectividad o incluso no poder realizar el análisis.

Beneficios

- La **confianza** se genera entre las partes interesadas ya que se necesita poco o ningún muestreo para extraer información.
- Las **metodologías** se pueden aplicar **rápidamente**.
- El uso de **R** permite métodos auditables y repetibles.
- La minería de texto identifica **nuevas ideas** o **refuerza** las percepciones existentes basadas en toda la información relevante.

[Posibles usos] #(_fig/tm1.PNG)

Flujo de trabajo en la minería de texto

[Flujo de trabajo] #(_fig/tm2.PNG)

1. **Definir el problema** y establecer las metas
2. **Identificar el texto** que se quiere **recolectar**
3. **Organizar** el texto (corpus, colección de documentos)
4. **Extraer** características (modelado inicial)
5. **Analizar** el texto (modelado completo)
6. Llegar a una idea o una **recomendación**

Conceptos básicos de la minería de texto

Minería de texto en la práctica

> La minería de textos representa la capacidad de tomar grandes **cantidades de lenguaje** no estructurado y **extraer rápidamente información útil** y novedosa que puede afectar la **toma de decisiones** de las partes interesadas.

Tipos de Minería de Texto

- Bolsa de Palabras (**bag words**)
- Análisis sintáctico (**syntactic parsing**)

Bolsa de palabras

Trata **cada palabra**, o grupos de palabras, llamados **n-gramas**, como una **característica única** del documento. El **orden de las palabras** y el *tipo gramatical* de las palabras *no se capturan* en el análisis de una bolsa de palabras. Una ventaja de este enfoque es que, por lo general, **no es computacionalmente costoso** ni abrumadoramente técnico organizar los corpus para la minería de texto. Como resultado, el análisis de estilo de la bolsa de palabras a menudo se puede realizar rápidamente. Además, la bolsa de palabras **encaja** muy bien en los **marcos de ML/MD** porque proporciona una matriz organizada de observaciones y atributos (*Base de datos estructurada*)

Esta se organiza como:

- DTM: Filas documentos, columnas palabras (tokenización)
- TDM: Filas palabras (tokenización), columnas documentos

Nota: la colección de documentos de interés se llama corpus

- DTM: Document Term Matrix
- TDM: Term Document Matrix

> Actividad 1

- Buscar 3 noticias/tweets/facebook/tiktok y armar el DTM y TDM

Análisis sintáctico

Se basa en la sintaxis de las palabras. En su raíz, la sintaxis representa un conjunto de reglas que definen los componentes de una oración que luego se combinan para formar la oración misma (similar a los bloques de construcción).

Específicamente, el análisis sintáctico utiliza técnicas de etiquetado de parte del discurso (POS) para identificar las palabras mismas en un contexto gramatical o útil.

R tiene un paquete que se basa en el proyecto OpenNLP (procesamiento de lenguaje natural) para realizar estas tareas. Estas diversas etiquetas son atributos capturados como metadatos de la oración original

Recolección

La recolección de texto puede provenir de:

- *Base de datos en csv u otro similar*: Normalmente cada texto/documento se incorpora dentro de un vector
- *Colección de documentos:* Carpeta con archivos del mismo formato
- *Scraping Web:* Raspar información de páginas web
- *API:* Puertas de entrada que dejan algunos servicios

Tratamiento de texto

Librerías en R para texto

- stringi: Funciones para el tratamiento de texto
- stringr: Funciones para el tratamiento de texto
- qdap: Orientada a la minería de texto
- tm: text mining

Caracteres y sustitución

- Cantidad de caracteres: nchar
- Identificar patrones de texto y reemplazar: sub

Pegar, splits y extracciones

- Unir texto: paste
- Dividir texto según algún criterio: strsplit
- Extraer un sub texto de un texto: substr

Actividad 2

Del vector de noticias en mayúscula, extraer los últimas 2 caracteres del documento.

Actividad 3

Usando la EH23, la variable folio/UPM es una variable de identificación y de tipo texto/carácter, tiene la letra A o D que hace referencia al área Amanzanada o Dispersa. Se pide, extraer la letra y hacer una tabla de frecuencias.

Búsqueda de palabras clave

- grep
- grepl

Stringr and stringi

- str_detect
- stri_count

Pasos de preprocesamiento para minería de texto de bolsa de palabras

- Tener cuidado cuando el texto es reconocido como factor
- El enfoque de la bolsa de palabras consume **RAM** si el corpus es grande
- Tener cuidado con la configuración de los equipos en cuanto la codificación y reconocimiento de caracteres del **español**
- Los pasos dependerán del corpus del estudio

Una ruta sugerida:

- Configuración básica para los caracteres e idioma
- Cargar las librerías básicas
- Cargar el corpus de interés
- Limpieza de texto
  - Mayúsculas, minúsculas
  - Puntuación
  - Espacios
  - Números
  - Palabras específicas (stopwords)
  - Prefijos y sufijos
- Armado del DTM O TDM

Spellcheck (ORTOGRAFÍA)

- Se recomienda realizar este tratamiento antes de cargar los documentos a R.
- https://github.com/wooorm/dictionaries/tree/main/dictionaries

Lematización y etiquetado

Importación, corpus, TDM, DTM

Frecuencias

Asociaciones

Nubes de palabras

Análisis de sentimiento

El análisis de sentimientos es el proceso de extraer la intención emocional del autor de un texto.

Se debe tener en cuenta:

- Aspectos culturales
- diferencias demográficas
- texto con sentimientos compuestos

Hay varios **marcos de referencias** de emociones que se pueden considerar. Uno de los más usados es el creado en 1980 por *Robert Plutchik* (psicólogo), se establecen 8 emociones:

- (-) ira (anger)
- (-) miedo (fear)
- (-) tristeza (sadness)
- (-) asco (disgust)
- (+) sorpresa (surprise)
- (+) anticipación (anticipation)
- (+) confianza (trust)
- (+) alegría (joy)

[Espectro de emociones de Plutchik's a partir de las primarias] #(_fig/tm3.PNG)

La manera de aplicar el análisis de sentimiento en el enfoque de bolsas de palabras es utilizar un léxico con los sentimientos.

> Un léxico es como un diccionario que asocia cada término con una palabra

Una dificultad de estos léxicos son los modismos de cada región.

Lo que se recomienda en esos casos es ampliar el léxico.

Librería syuzhet

Redes

- **Nodo:** normalmente es una entidad, objeto o sujeto. Para nuestro caso los nodos son los términos o los documentos. Estos nodos tienen atributos, como el tamaño, la forma, el color, etc.
- **Conexiones:** se refieren a la manera de vincular los nodos, se representan por líneas y tienen diferentes atributos; tipo, grosor, dirección. A->B; A<-B; A<->B; A-B. En nuestro caso se utiliza A-B.

Para la minería de texto el insumo principal es matriz adjunta a partir de un TDM o DTM. Donde se cuenta con matrices simétricas que hacen referencia a los términos o documentos

$$Adj_A = A * A^t$$

Cluster

Sigue la misma lógica vista en los métodos de agrupamiento

Ejercicios de práctica

1. Usando los datos del archivo eco.RData calcule el porcentaje de documentos que tratan sobre “pobreza”
2. Usando los archivos pdf de la LAJED realice una nube de palabras global, por documentos y una red sobre las palabras 3 Usando la encuesta a hogares 2021, la pregunta s03a_05e respecto las razones por la que no se inscribió o matriculo realice la limpieza correspondiente, una nube de palabras.
3. Usando la base de datos sobre economia_tw.Rda, para cada gestión 2021, 2022 y 2023 identifique el texto, realice la limpieza y presente: + Nube de palabras + Análisis de sentimiento + Cluster jerárquico por documentos + Red para términos y usuarios
4. Recolecte un corpus de su interés y realice la limpieza y genere: + Nube de palabras + Análisis de sentimiento + Cluster jerárquico por documentos + Red para términos y usuarios