



CURSO: MINERÍA DE DATOS
MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

LABORATORIO VIII: Ingeniería de atributos textuales

INTRODUCCIÓN

Esta práctica de laboratorio tiene como objetivo avanzar sobre algunos tópicos de ingeniería de features textuales, trabajando con técnicas de preprocesamiento de atributos textuales y algunas estrategias de representación y ponderación de los términos que componen los atributos textuales.

Para la exploración de estos temas, se utilizará el IDE R-Studio del lenguaje de programación R, a efectos de ejercitar los conceptos abordados en las clases teóricas.

CONSIGNAS

A partir del dataset que contiene las letras de algunas de las canciones de Spotify utilizadas en el TP01, se solicita trabajar sobre las siguientes consignas:

1. SOBRE LOS DATOS

- a. Cargue¹ y explore el dataset: explique en qué consiste el mismo y qué características posee.
- b. Genere el corpus de documentos y explore la instrucción *inspect()*. ¿Qué información brinda?

2. PREPROCESAMIENTO DE TEXTO

- a. Utilizando la librería *tm*, ejecute las siguientes tareas de pre-procesamiento sobre los datos y verifique en cada paso si se reduce la cantidad de términos del corpus de documentos:
 - i. Convierta el texto a minúsculas.
 - ii. Elimine valores numéricos.
 - iii. Elimine palabras vacías.

¹ Explore la instrucción *read.table()*.



CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

- iv. Elimine signos de puntuación. ¿Quedan signos de puntuación sin eliminar por parte de la librería *tm*? Explore el resultado y sirvase de la función *gsub()* en estos casos.
- v. Elimine los espacios en blanco adicionales.
- vi. Finalmente, elimine los acentos.

3. GENERACIÓN DE FEATURES A PARTIR DE TEXTO

- a. Una vez preprocesado el texto, genere la Matriz Término-Documento y explore el resultado. ¿Qué observa a simple vista?
- b. ¿Cuáles son los términos que más aparecen?

4. REPRESENTACIÓN GRÁFICA DE FEATURES TEXTUALES

- a. Genere la nube de palabras (wordcloud) con los términos más frecuentes.
- b. Verifique gráficamente el cumplimiento de la Ley de Zipf.

Referencias sugeridas:

Text Mining Package: <https://cran.r-project.org/web/packages/tm/tm.pdf>

Ingeniería de Features textuales: https://rpubs.com/jumafernandez/text_features