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 de un script[[1]](#footnote-1) se han descargado cerca de 6000 letras de canciones presentes en la DB de Spotify y que fuera utilizada en el TP01.

Luego, se ha generado[[2]](#footnote-2) un dataset con el subconjunto de las letras que corresponden al idioma español y se ha realizado un mongoexport[[3]](#footnote-3) con la información. Se solicita trabajar, en función de esos datos, sobre las siguientes consignas:

1. **SOBRE LOS DATOS**
   1. Cargue y explore el dataset *lyrics-spanish.json*: explique en qué consiste el mismo y qué características posee.
   2. Genere el corpus de documentos y explore la instrucción *inspect().*¿Qué información brinda?
2. **PREPROCESAMIENTO DE TEXTO**
   1. 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:
      1. Convierta el texto a minúsculas.
      2. Elimine valores numéricos.
      3. Elimine palabras vacías.
      4. 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.
      5. Elimine los espacios en blanco adicionales.
      6. Finalmente, elimine los acentos.
3. **GENERACIÓN DE FEATURES A PARTIR DE TEXTO**
   1. Una vez preprocesado el texto, genere la Matríz Término-Documento y explore el resultado. ¿Qué observa a simple vista?
   2. ¿Cuáles son los términos que más aparecen?
4. **REPRESENTACIÓN GRÁFICA DE FEATURES TEXTUALES**
   1. Genere la nube de palabras (wordcloud) con los términos más frecuentes.
   2. 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>

1. Disponible en <https://raw.githubusercontent.com/dm-uba/dm-uba.github.io/master/2021/laboratorios/LAB08/scripts/download-lyrics-v2.r> [↑](#footnote-ref-1)
2. Disponible en <https://raw.githubusercontent.com/dm-uba/dm-uba.github.io/master/2021/laboratorios/LAB08/scripts/filter-spanish-lyrics.R> [↑](#footnote-ref-2)
3. Comando: “mongoimport -h localhost -d DMUBA\_SPOTIFY -c lyrics --file=.\lyrics-dm.json” [↑](#footnote-ref-3)