CURSO: CMP 5002 - DATA MINING
COLEGIO: POLITÉCNICO
Semestre: 1er Semestre 2023/2024

Proyecto 3: Ejercicio usando la técnica *TextMining*

Problema:

1. La industria cinematográfica se encarga de producir proyectos tales como documentales, filmes de larga y corta duración, entre otros. Todos estos productos son recolectados y establecidos en el portal IMDB (*internet movies database*) para el consumo de cinéfilos y críticos alrededor del mundo. El siguiente proyecto se centra en el uso de una colección de *reviews* realizados a distintos productos cinematográficos dentro de la base de datos IMDB. Por tanto, se desea lo siguiente:
   a) Descargar la base de datos aquí: P3- Text mining. En esta carpeta encontrarás do ficheros: *reviews* - 12500 ficheros textos con la crítica escrita y el fichero *reviews_url* – con 12500 URLs referentes a cada producto cinematográfico. Cada URL contiene el identificador del producto cinematográfico.
   Nota: El orden de los ficheros en cada carpeta debe conservarse para que la URL y el *review* pertenezcan al mismo producto cinematográfico. Además, para conocer el título del producto se debe usar la URL hasta el identificador numérico y no hasta la sección de comentarios.
   b) Seleccionar una muestra aleatoria de 100 *reviews* de la base de datos IMDB. (**1 punto**)
   c) De los *reviews* seleccionados en el literal b), determinar las palabras más utilizadas por producto cinematográfico (top 10 palabras) y mostrar una tabla con todos los índices de cálculo por términos: *tf, df, idf, tf-idf*. (**2 puntos**)
   d) De la tabla obtenida en el literal c), mostrar una nueva tabla con la normalización (conversión al espacio vectorial) de la métrica *tf-idf* obtenida por cada *review*. (**2 puntos**)
   e) Establecer un ranking por título cinematográfico, atendiendo al *query* introducido por el usuario. El criterio de similaridad debe ser basado en la métrica del coseno y sobre el espacio vectorial. (**3 puntos**)
   f) Haciendo uso de la métrica del coseno. Se desea calcular entre los títulos cinematográficos seleccionados, cuáles obtuvieron un *review* muy parecido (posible intercepción de filmes). Mostrar los resultados de similaridad en una tabla. (**2 puntos**)

Adicionalmente:

I. Cargar al D2L los códigos implementados (fichero compactado) dentro del plazo de entrega.

II. +1

- Implementar una idea de *query* semántico ☺ (Ejemplo: cuáles títulos cinematográficos recibieron críticas positivas o negativas dependiendo de su contexto)

Nota: En cada fase de evaluación el profesor aplicará puntos de chequeo sobre el código implementado y basado en la trazabilidad por literal.