Trabajo Práctico Nº 6

Tema: Unidad 3 y 4

Fecha Inicio: 21/05/2024 **Fecha de Entrega:** 04/06/2024

Actividades:

1) Para los documentos procesados en el punto N° 3 del TP N° 1 (5 documentos) y Punto N° 2 del TPN ° 4 (10 documentos/noticias), por separado, cree una representación TF-IDF de los mismos (RepA primer caso y RepB segundo caso). Las representaciones deben ser de n-1 documentos, dejando 1 (un) documento para realizar una consulta y obtener el top 3 de documentos más parecidos/similares. Debe mostrar como resultado para cada representación el score obtenido y el texto del documento. Las representaciones a realizar deben tomar:
   a) El texto original
   b) Eliminando stop-words
   c) Con bi-gramas

   Se produce algún cambio en los resultados?

   NOTA: Ejemplos empleando el esquema TF-IDF en Scikit-Learn:
   https://markhneedham.com/blog/2016/07/27/scitkit-learn-tfidf-and-cosine-similarity-for-computer-science-papers/
   https://www2.cs.duke.edu/courses/spring14/compsci290/assignments/lab02.html
   https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

2) Empleando el corpus inaugural de NLTK (que contiene los discursos inaugurales de los presidentes de EEUU), utilice el archivo '2009-Obama.txt' y realice lo siguiente.
   Obtenga un resumen del texto entre 7 y 10 oraciones, mediante 2 métodos distintos de resúmenes extractivos, compare los resultados.