Practica 8

El objetivo de esta práctica es determinar la similaridad de varias documentos usando una medida de similitud

Datos: A modo de ejemplo use los siguientes documentos

- Documento1: noticias1.txt
  A civil society organization on Tuesday said there was an urgent need to improve government schools across the country so that a maximum number of children can get quality education instead of joining the army of unskilled labor force as they grow up.

- Documento2: noticias2.txt
  Sometimes you just don't want to think. You only want mindless entertainment where nothing is logical and it doesn't matter.

- Documento3: noticias3.txt
  Google announced its StreetView trekker has plunged under the waters of the Galapagos and tracked across its islands, bringing to the internet 360-degree images of the isolated landscape and world's largest living tortoises that inspired Darwin to develop his theory of evolution.

- Documento4: noticias4.txt
  Pakistan skipper Misbah-ul-Haq believes his team's prospects in the Champions Trophy will depend on his players' ability to adapt to English conditions.

- Documento5: noticias5.txt (texto repetido del Documento4-noticias4.txt)
  Pakistan skipper Misbah-ul-Haq believes his team's prospects in the Champions Trophy will depend on his players' ability to adapt to English conditions.

Con este objetivo se usará los siguientes operadores

Process Document from Data: Use este operador para leer información de varios documentos. Entonces configure este operador para determinar cada etiqueta de clase y mencionar el directorio de origen. Dentro de este operador crear un subproceso que permita ejecutar las siguientes acciones

1. Dividir las frases en tokens
2. Filtras los tokens por tamaño
3. Transformar a mayúsculas
4. Eliminar Palabras Vacías
5. Reducir palabras similares a sus bases (stem).

En este proceso, configure el método de creación de vectores a ser utilizado. El operador a ser usado es TF-IDF, el cual permite asignar pesos a cada una de las palabras que conforman el documento.

Tareas:

- Identifique el operador necesario para calcular la similaridad entre documentos usando una función de distancia. Pruebe de usar la distancia euclidiana numérica y la medida del coseno.

- En cada caso identifique los valores que permiten determinar cuándo dos documentos son similares