PRÁCTICA: SIMILITUD DE DOCUMENTOS CON BAG-OF-WORDS Y SIMILITUD DEL COSENO

OBJETIVO:

Adquirir experiencia práctica en el cálculo de la similitud entre documentos utilizando el método de bag-of-words y la métrica de similitud del coseno.

CONTEXTO:

Se proporciona un conjunto de 12 documentos relacionados con tecnología móvil y vehículos eléctricos. En el conjunto, algunos documentos tratan solo de tecnología móvil, otros solo de vehículos eléctricos, y hay documentos que abordan ambos temas simultáneamente.

INSTRUCCIONES:

1. Preprocesamiento de datos:

- Convertir todos los textos a minúsculas.

- Eliminar signos de puntuación y otros caracteres no alfabéticos.

- (Opcional) Realizar una eliminación de palabras comunes o "stop words".

2. Construcción de Bag-of-Words:

- Crear un vocabulario global basado en todas las palabras únicas presentes en los documentos.

- Representar cada documento como un vector en el espacio del vocabulario. La posición de cada palabra en el vector corresponderá a su frecuencia en el documento.

3. Cálculo de la Similitud del Coseno:

- Calcular la similitud del coseno entre todos los pares de documentos utilizando sus representaciones vectoriales.

4. Análisis:

- Identificar qué documentos son más similares entre sí y cuáles son menos similares.

- Observar la relación entre la similitud y los temas tratados en los documentos. Por ejemplo, determinar si los documentos que tratan sobre el mismo tema tienden a ser más similares entre sí que aquellos que tratan temas diferentes.

ENTREGABLES:

- Código fuente utilizado para el preprocesamiento, construcción del bag-of-words y cálculo de la similitud del coseno.

- Un breve informe que detalle los hallazgos en cuanto a la similitud entre documentos.