Grado en Ingeniería Informática

Gestión del conocimiento en las organizaciones

Sistemas de recomendación. Modelos basados en el contenido

Lorenzo Román Luca de Tena, Mariam Laaroussi Ramos

ÍNDICE

1. Descripción del Análisis

El objetivo principal del análisis realizado es calcular la similaridad entre documentos a partir de la frecuencia de términos (TF), la frecuencia inversa de documentos (IDF) y el valor combinado TF-IDF. El análisis se centró en la implementación de un algoritmo que procesa documentos textuales, calcula estos valores y evalúa la similitud entre ellos mediante el uso de la similitud coseno.

El sistema está compuesto por dos archivos principales de código:

1. **main.cc**: Contiene la función principal del programa y coordina las tareas del sistema.

2. **tools.cc**: Contiene funciones auxiliares utilizadas en el proceso de análisis de los documentos.

El proceso comienza con la lectura de los documentos, las palabras vacías (stopwords) y las lematizaciones desde archivos proporcionados por el usuario. Luego, se calculan los valores de TF, IDF y TF-IDF para cada término en los documentos. Finalmente, se calculan las similitudes coseno entre los documentos, las cuales se imprimen junto con los resultados obtenidos.

El algoritmo utiliza un enfoque de preprocesamiento para optimizar los cálculos, como la eliminación de palabras vacías y la lematización, lo que mejora la precisión de los resultados al reducir la variabilidad en los términos utilizados en los documentos.

2. Conclusiones Extraídas

A partir de los resultados obtenidos, se puede concluir que el cálculo de las similitudes entre documentos es una tarea eficiente para evaluar la relación entre textos. Los pasos de preprocesamiento, como la eliminación de las palabras vacías y la lematización, contribuyen significativamente a mejorar la precisión de los resultados. Al utilizar el valor de TF-IDF, se logra un análisis más representativo del contenido relevante de cada documento.

En cuanto a la similitud coseno, este método resultó ser efectivo para medir la relación entre los documentos, proporcionando una medida cuantitativa de similitud entre los textos. Un valor cercano a 1 indica que los documentos son muy similares, mientras que un valor cercano a 0 indica que son muy diferentes.

Es importante destacar que los valores calculados de TF, IDF y TF-IDF son fundamentales para cualquier

sistema de búsqueda o análisis de textos, ya que permiten identificar los términos más relevantes en un conjunto de documentos. El uso de estas métricas es común en tareas de minería de texto, recuperación de información y análisis semántico.

Finalmente, el algoritmo proporciona un enfoque robusto para calcular y visualizar la similitud entre los documentos, lo que puede ser utilizado en una variedad de aplicaciones, desde motores de búsqueda hasta sistemas de recomendación basados en contenido.