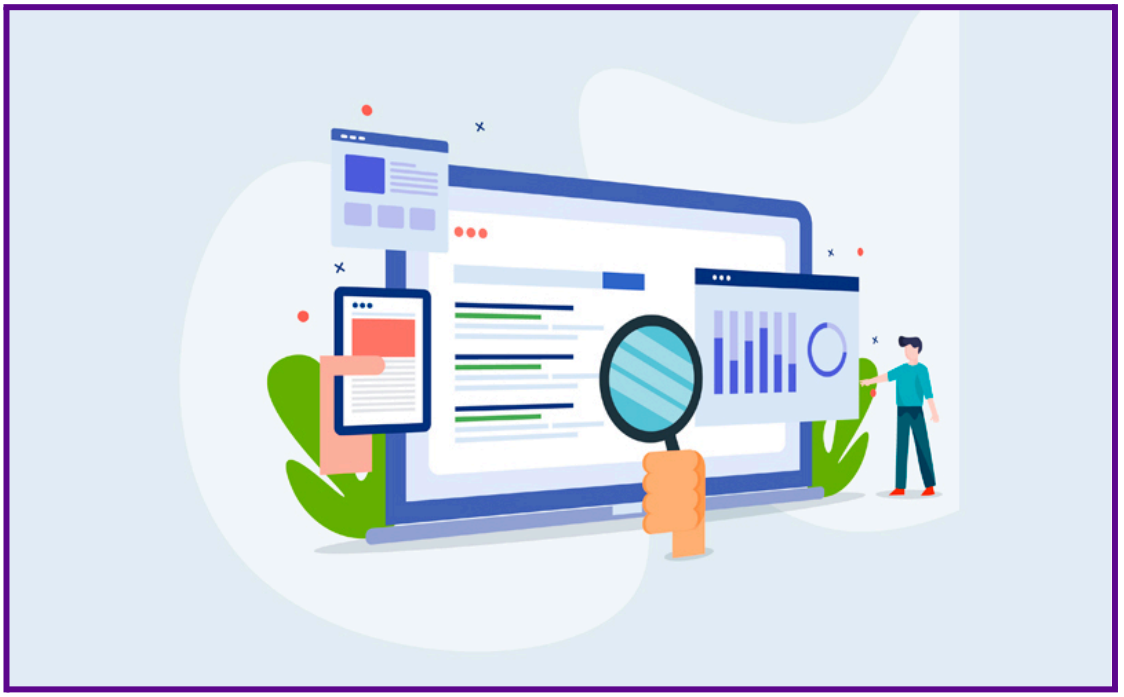
Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

25-26 GCO

Modelos basados en el contenido



C/ Padre Herrera s/n
38207 La Laguna
Santa Cruz de Tenerife. España

T: 900 43 25 26

ull.es

Jean Franco Hernández García - alu0101538853@ull.edu.es
Arun Daswani Lakhani - alu0101560410@ull.edu.es
Javier González Brito - alu0101548197@ull.edu.es

Tabla de contenidos

Introducción	2
Descripción del método	3
Desarrollo e implementación	3
Resultados	5
Conclusiones	5
Bibliografía / Fuentes	6

Introducción

Siguiendo la línea de la entrega anterior, en esta práctica se ha desarrollado un sistema de recomendación basado en el contenido. Este es un enfoque que centra su funcionamiento en el análisis de las características de los ítems, para generar con ello, recomendaciones personalizadas. A diferencia del filtrado colaborativo, el cual se apoyaba en las valoraciones de los usuarios, los modelos basados en el contenido utilizan la información descriptiva de los ítems para identificar similitudes y sugerir aquellos que comparten términos o temas relevantes entre sí.

El sistema implementado en esta ocasión procesa un conjunto de ficheros de texto, aplicando técnicas de preprocesamiento lingüístico como la eliminación de 'stopwords' y la lematización de términos, con el fin de normalizar el vocabulario y así mejorar y facilitar la calidad del análisis. Además de esto, también se calculan diferentes métricas las cuales nos permiten representar cada documento como un vector de características ponderadas. A partir de estos vectores, se determina la similaridad coseno entre los diferentes documentos, identificando aquellos más relacionados entre sí.

Teniendo en cuenta todo lo comentado hasta el momento, digamos que en esta práctica se busca comprender en profundidad el funcionamiento de los modelos basados en el contenido, su aplicación en el procesamiento de texto y su utilidad en la generación de recomendaciones basadas en el contenido.

Descripción del método

- Preprocesamiento del texto
    - Tokenización: la separación del texto es palabras individuales.
    - Eliminación de stopwords: el descarte de palabras muy frecuentes sin un valor como tal. (el, para, de, en).
    - Lematización: se reducen las palabras a su forma base (comida -> comer)
- Cálculo de TF-IDF
    - TF (Term Frequency): la frecuencia de un término en el texto
    - IDF (Inverse Document Frequency): se encarga de medir la rareza de un término en un conjunto total de documentos.
    - TF-IDF: es el producto de ambos valores, encargado de ponderar la relevancia del término.
- Cálculo similitud coseno
    - Cada documento es representado por un vector TF-IDF.
    - Cuanto más próximo a 1 se el valor al medir la similitud entre dos valores, más se parecen

Resultados

Los resultados obtenidos con la ejecución del programa se encuentran en dos directorios ubicados dentro del repositorio del proyecto.

El directorio denominado 'results' contiene los archivos CSV correspondientes a cada documento procesado de los que fueron proporcionados por el profesorado de la asignatura. De forma análoga, en el directorio 'resultados-textos-buscados' encontramos los archivos de resultado correspondientes a los documentos que fueron buscados por nosotros.

Cada archivo CSV contiene 5 columnas, en las que reflejan los siguientes datos:

- **term_index:** Índice del término dentro del vocabulario global.
- **term:** Palabra después de aplicar el preprocesamiento
- **TF (Term Frequency):** Frecuencia relativa del término en ese documento específico.
- **IDF (Inverse Document Frequency):** Mide la importancia del término en toda la colección de documentos.
- **TF-IDF:** Producto de TF × IDF. Representa la relevancia final del término para ese documento.

Por último encontramos un archivo CSV denominado 'similarities' el cual contiene la matriz de similitud y por tanto la similitud coseno entre todos los pares de documentos.

Conclusiones

Una vez ejecutado el sistema de recomendación, hemos podido comprobar que nuestro modelo basado en el contenido muestra un funcionamiento coherente y acorde a lo esperado. Los resultados indican que los documentos que contienen en común un mayor número de términos relevantes, despues del preprocesamiento de texto, obtienen valores de similitud coseno más altos.

Dentro de los conjuntos de pruebas que nos ha proporcionado el profesorado, se observa que tienen temáticas similares, mostrando un vocabulario similar y por ende presentando valores de similitud mayores a 0,8, mientras que los documentos con distinta temática apenas superan valores del 0,1-0,2. Esto hace que el programa gane fuerza y que el cálculo del TF-IDF permita correctamente diferenciar los textos.

Respecto a los diez documentos propuestos por nuestro grupo, también el sistema fue capaz de identificar de manera correcta la similitud de los textos viendo como es capaz de adaptarse a nuevos documentos.

Por otro lado, también se comprobó que la calidad del preprocesamiento de texto (stopwords y lematización) afectan de manera significativa en los resultados. Mismamente hay casos entre términos equivalentes como coche y automóvil los cuales al ser muy parecidos pueden ser uno mismo. Esto refleja una de las limitaciones que tiene este modelo.

En resumen, el sistema desarrollado por el grupo ha sido capaz de cumplir con los objetivos planteados, permitiendo al usuario poder analizar, comparar y recomendar documentos en función de su contenido, así ofreciendo resultados facilmente interpretables y coherentes.

Bibliografía / Fuentes

Incluye:

- MIT artículos - Artículos usados para la creación de nuestros documentos
- Documentación de NumPy - Manuales de usuario
- Documentación de Pandas - Guía de uso
- Enunciado de la práctica - Aula virtual de la asignatura