Universidad de La Laguna

GCO.

Sistemas de recomendación.

Modelos basados en el contenido

Víctor Rodríguez Dorta
alu0101540153

ALEJANDRO RODRÍGUEZ MEDEROS
alu0101413938

MARIO GUERRA PÉREZ
alu0101395036

Introducción:

En la era de la información, la cantidad de datos generados y disponibles crece de forma exponencial. Ante este escenario, los sistemas de recomendación se han convertido en herramientas esenciales para filtrar, organizar y personalizar la información que reciben los usuarios. Estos sistemas se aplican en múltiples ámbitos, como el comercio electrónico, las plataformas de streaming o los motores de búsqueda, con el objetivo de ofrecer sugerencias relevantes basadas en los intereses o comportamientos previos del usuario.

Dentro de las diferentes aproximaciones existentes, los modelos basados en el contenido (content-based) destacan por centrarse en el análisis de las características de los propios ítems (documentos, productos, películas, etc.) en lugar de depender del comportamiento de otros usuarios. Este enfoque se fundamenta en la representación de los elementos mediante descriptores o vectores de características, lo que permite medir la similaridad entre ellos y generar recomendaciones a partir de esa comparación.

El propósito de esta práctica es implementar un sistema de recomendación basado en el contenido, aplicando técnicas de procesamiento del lenguaje natural (NLP) y modelado vectorial. Para ello, se parte de un conjunto de documentos en texto plano y se procesan mediante distintas etapas: eliminación de palabras vacías (stopwords), lematización de términos y cálculo de métricas estadísticas como TF (Term Frequency), IDF (Inverse Document Frequency) y su producto TF-IDF, que cuantifica la relevancia de los términos en cada documento. Finalmente, se calcula la similaridad coseno entre los vectores resultantes para identificar qué documentos son más parecidos entre sí.

El software desarrollado puede ejecutarse tanto en línea de comandos como en una interfaz web, y genera como salida una tabla de términos con sus respectivos valores de TF, IDF y TF-IDF, además de la matriz de similaridades entre documentos.

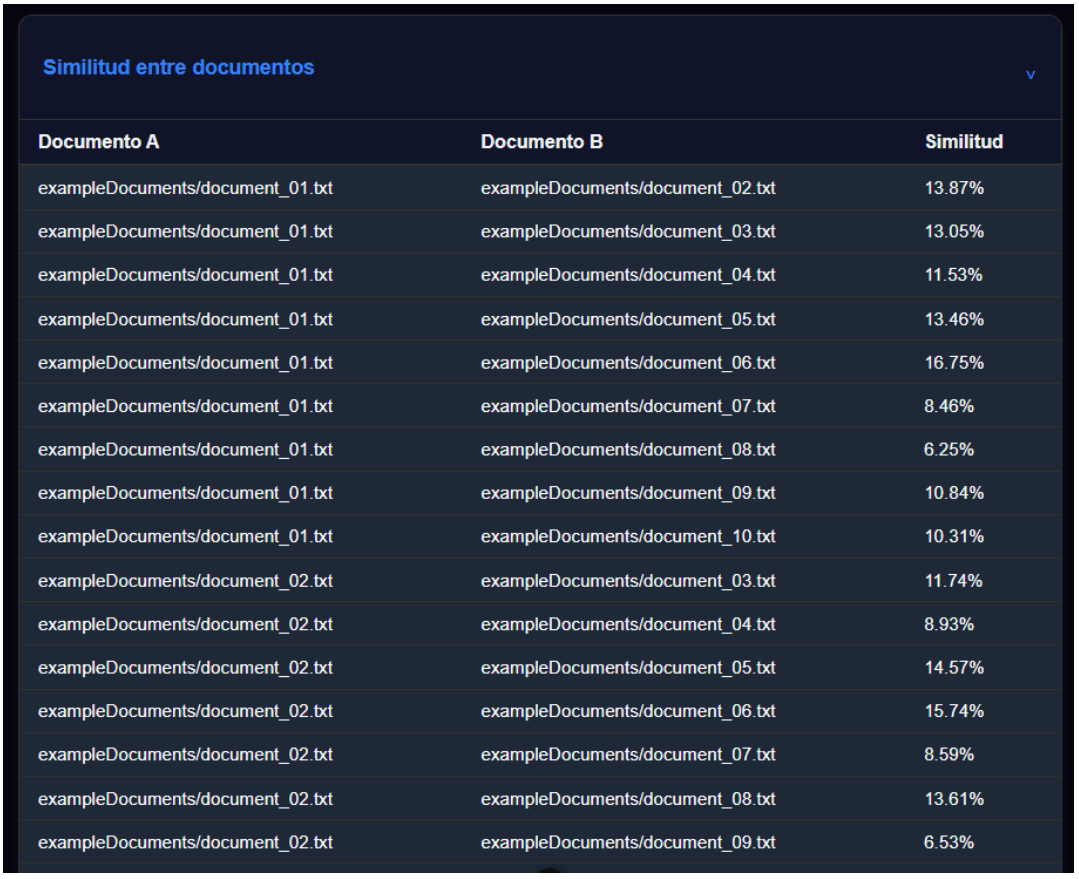
Resultados obtenidos:

Una vez implementado el sistema, se ejecutó con los documentos de ejemplo de la asignatura. El sistema generó tablas TF-IDF para cada documento, donde se calcularon las siguientes métricas:

- **TF (Term Frequency)**: Frecuencia normalizada del término en el documento
- **IDF (Inverse Document Frequency)**: Especificidad del término en toda la colección
- **TF-IDF**: Relevancia ponderada del término

En las tablas generadas se identifican los términos más relevantes de cada documento. Por ejemplo, en el documento 05, los términos con mayor TF-IDF fueron "field" (1.0217), "slow" (1.0217), "grasses" (0.7133), "swayed" (0.7133), "steady" (0.7133) y "watching" (0.7133). Los términos con IDF más alto (~0.51 para "field" y "slow") son los más específicos de este documento, mientras que términos con IDF menor (~0.36) aparecen en algunos otros documentos de la colección, reduciendo su poder discriminativo.

La matriz de similaridad coseno muestra valores entre 0 (sin similitud) y 1 (idénticos). En los documentos de ejemplo, las similaridades oscilan entre 11.66% y 20.24%, indicando una colección diversa. Las similaridades más altas se observan entre document_04 y document_08 (20.24%), document_05 y document_02 (19.34%), y document_05 y document_04 (19.14%), sugiriendo que estos pares comparten más términos relevantes. Por el contrario, similaridades bajas como document_04 y document_10 (11.66%) o document_05 y document_07 (12.35%) indican contenidos temáticos diferentes. Se adjunta a continuación captura de la matriz de similaridad:



Similitud entre documentos

Documento A	Documento B	Similitud
exampleDocuments/document_01.txt	exampleDocuments/document_02.txt	13.87%
exampleDocuments/document_01.txt	exampleDocuments/document_03.txt	13.05%
exampleDocuments/document_01.txt	exampleDocuments/document_04.txt	11.53%
exampleDocuments/document_01.txt	exampleDocuments/document_05.txt	13.46%
exampleDocuments/document_01.txt	exampleDocuments/document_06.txt	16.75%
exampleDocuments/document_01.txt	exampleDocuments/document_07.txt	8.46%
exampleDocuments/document_01.txt	exampleDocuments/document_08.txt	6.25%
exampleDocuments/document_01.txt	exampleDocuments/document_09.txt	10.84%
exampleDocuments/document_01.txt	exampleDocuments/document_10.txt	10.31%
exampleDocuments/document_02.txt	exampleDocuments/document_03.txt	11.74%
exampleDocuments/document_02.txt	exampleDocuments/document_04.txt	8.93%
exampleDocuments/document_02.txt	exampleDocuments/document_05.txt	14.57%
exampleDocuments/document_02.txt	exampleDocuments/document_06.txt	15.74%
exampleDocuments/document_02.txt	exampleDocuments/document_07.txt	8.59%
exampleDocuments/document_02.txt	exampleDocuments/document_08.txt	13.61%
exampleDocuments/document_02.txt	exampleDocuments/document_09.txt	6.53%

Resultados obtenidos con nuestros documentos:

Se elaboró un corpus experimental de 10 documentos divididos temáticamente: documentos 1-5 sobre inteligencia artificial y documentos 6-10 sobre historia. Los documentos de IA incluyen términos como “algoritmo, modelo, entrenamiento, datos, redes neuronales, aprendizaje y predicción”, mientras que los históricos contienen vocabulario específico como “imperio, civilización, batalla, reino, cultura, conquista y dinastía”.

Los resultados de la matriz de similaridad confirman la capacidad del sistema para diferenciar y agrupar documentos por temática:

Similaridad intra-grupo (IA - documentos 1-5):

- Ejemplos: doc1-doc2 (8.65%), doc1-doc3 (5.34%), doc4-doc5 (1.24%)
- **Promedio: 3.44%**

Similaridad intra-grupo (Historia - documentos 6-10):

- Ejemplos: doc8-doc10 (8.36%), doc6-doc8 (7.94%), doc7-doc9 (6.13%), doc6-doc7 (5.61%), doc6-doc10 (5.11%)
- **Promedio: 5.35%**

Similaridad inter-grupo (IA vs Historia):

- Ejemplos: doc5-doc9 (2.68%), doc4-doc6 (1.92%), doc3-doc6 (1.65%), doc2-doc6 (1.53%), doc3-doc9 (1.40%), doc1-doc10 (0.35%)
- **Promedio: 1.70%**

Los resultados demuestran que las similaridades intra-grupo son consistentemente superiores a las inter-grupo. Sobretodo, los documentos históricos muestran mayor cohesión temática (5.35%) comparado con los de IA (3.44%), posiblemente debido a un vocabulario histórico más específico y compartido. La diferencia más significativa se observa entre la similaridad intra-Historia (5.35%) y la inter-grupo (1.70%), con una diferencia de 3,65%, confirmando que el modelo TF-IDF captura efectivamente las características léxicas distintivas de cada dominio temático.

El sistema clasifica correctamente documentos nuevos según su contenido: un texto sobre machine learning obtendría mayor similaridad con los documentos 1-5, mientras que uno sobre civilizaciones antiguas se asociaría con los documentos 6-10. Se adjunta a continuación la matriz de similaridad:

newExampleDocs/doc3.txt	newExampleDocs/doc4.txt	4.39%
newExampleDocs/doc3.txt	newExampleDocs/doc5.txt	9.02%
newExampleDocs/doc3.txt	newExampleDocs/doc6.txt	0.29%
newExampleDocs/doc3.txt	newExampleDocs/doc7.txt	0.29%
newExampleDocs/doc3.txt	newExampleDocs/doc8.txt	0.92%
newExampleDocs/doc3.txt	newExampleDocs/doc9.txt	1.26%
newExampleDocs/doc4.txt	newExampleDocs/doc5.txt	8.94%
newExampleDocs/doc4.txt	newExampleDocs/doc6.txt	0.11%
newExampleDocs/doc4.txt	newExampleDocs/doc7.txt	0.35%
newExampleDocs/doc4.txt	newExampleDocs/doc8.txt	1.93%
newExampleDocs/doc4.txt	newExampleDocs/doc9.txt	0.41%
newExampleDocs/doc5.txt	newExampleDocs/doc6.txt	0.70%
newExampleDocs/doc5.txt	newExampleDocs/doc7.txt	0.85%
newExampleDocs/doc5.txt	newExampleDocs/doc8.txt	1.48%
newExampleDocs/doc5.txt	newExampleDocs/doc9.txt	1.08%
newExampleDocs/doc6.txt	newExampleDocs/doc7.txt	5.24%
newExampleDocs/doc6.txt	newExampleDocs/doc8.txt	9.15%
newExampleDocs/doc6.txt	newExampleDocs/doc9.txt	3.62%
newExampleDocs/doc7.txt	newExampleDocs/doc8.txt	6.01%
newExampleDocs/doc7.txt	newExampleDocs/doc9.txt	4.73%
newExampleDocs/doc8.txt	newExampleDocs/doc9.txt	4.90%

Conclusión:

Esta práctica nos permitió comprender el funcionamiento de los sistemas de recomendación basados en contenido mediante la implementación práctica de TF-IDF y similaridad coseno. Los experimentos con los documentos de la asignatura y nuestro corpus experimental validaron la efectividad del modelo para representar y comparar documentos textuales.

Los resultados demostraron claramente la capacidad del sistema para agrupar documentos por temática. Las similaridades intra-grupo (3.44% para IA, 5.35% para Historia) fueron significativamente superiores a las inter-grupo (1.70%), con una diferencia de hasta 3.65 puntos porcentuales. Esto confirma que el modelo identifica correctamente las características léxicas diferentes de cada tema.

Se identificaron algunas limitaciones del enfoque: dependencia estricta del vocabulario compartido (sin considerar sinónimos o relaciones semánticas), sensibilidad a la calidad del preprocesamiento (stop words y lematización), y valores de similaridad relativamente bajos incluso entre documentos del mismo tema. Esto se debe a que cada documento, aunque trate sobre la misma temática, utiliza palabras bastante únicas y específicas, lo que hace que la similaridad no sea tan alta como se podría esperar.

A pesar de estas limitaciones, el sistema ofrece una base sólida para aplicaciones de clasificación y recomendación documental.