Universidad de La Laguna

IMPLEMENTACIÓN DE UN SISTEMA DE RECOMENDACIÓN BASADO EN CONTENIDOS

Luis Chinea Rangel (alu0101118116@ull.edu.es)
Adrián León Díaz (alu0101495668@ull.edu.es)
Marlon Eduardo Salazar Amador (alu0101433943@ull.edu.es)

ÍNDICE

1. Introducción

En este informe se documenta el diseño, implementación y evaluación de un sistema de recomendación basado en el contenido. El objetivo principal ha sido desarrollar una aplicación web que procese documentos en texto plano, aplique preprocesamiento (stop words y lematización), calcule métricas de representación (TF, IDF, TF–IDF) y obtenga medidas de similitud entre documentos mediante la similitud del coseno.

Se describirán las decisiones de diseño, la estructura del código, las instrucciones para ejecutar la aplicación y los resultados obtenidos sobre el conjunto de documentos de ejemplo proporcionado en el repositorio indicado. Además, se presenta un subconjunto propuesto de 10 nuevos documentos y se ilustra cómo el sistema genera recomendaciones y tablas por documento (índice de término, término, TF, IDF, TF–IDF) y la matriz de similitud entre pares. Finalmente, el informe sintetiza conclusiones, limitaciones y posibles mejoras futuras.

2. Fundamentos teóricos

En este apartado se describen los conceptos y fórmulas esenciales para la creación de un sistema de recomendación basado en el contenido sobre documentos de texto. Se explica cómo se representa un documento numéricamente (modelo vectorial), las medidas de importancia de término (TF, IDF, TF–IDF) y la medida de similitud utilizada (similitud del coseno).

2.1. Modelo vectorial de documentos

Cada elemento (ítem/documento) se representa como un vector de atributos en un espacio n-dimensional. Las características de un usuario se representan también como un vector. El producto escalar entre dos vectores se define como;

$$< x, y > \; = \sum_{i=1}^{d} x_i \, y_i$$

A partir de estos vectores se evalúa qué ítems deben recomendarse comparando características compartidas.

2.2. Preprocesamiento de texto

Antes de construir vectores, es habitual aplicar transformaciones sobre el texto para reducir ruido y tamaño del vocabulario:

- Stop words: eliminación de palabras muy frecuentes y poco informativas (artículos, preposiciones, etc.)
- Lematización: reducir palabras a su forma canónica (por ejemplo: "corriendo" → "correr").

Estas operaciones mejoran la calidad de los vectores y reducen su dimensión.

2.3. Frecuencia de término (TF)

La frecuencia de término mide cuántas veces aparece una palabra clave x en un documento y. Para determinar la frecuencia ponderada del término calculamos:

$$w_{t,d} = \{1 + log_{10} tf_{t,d} \ \ si \ tf_{t,d} > 0 \ ; \ 0 \ en \ otro \ caso\}$$

2.4. Frecuencia inversa de documento (IDF)

La IDF reduce el peso de términos muy frecuentes en el corpus:

$$IDF(x) = log\left(\frac{N}{df_x}\right)$$

donde N es el número total de documentos y df_x es el número de documentos en los que aparece x.

2.5. TF–IDF (ponderación combinada)

La medida TF–IDF combina TF e IDF para obtener pesos por término en cada documento, codificando los documentos en un espacio euclídeo multidimensional (vector de términos ponderados).

2.6. Normalización de vectores

Una vez calculados los valores TF–IDF para los atributos de un documento, se calcula la longitud del vector como la raíz cuadrada de la suma de los cuadrados de los valores de sus atributos. A continuación cada componente se divide por dicha longitud para obtener el vector normalizado.

2.7. Similitud coseno

Para determinar la similitud entre documentos, se emplea el modelo del espacio vectorial y se calculan los ángulos entre vectores. Tras normalizar los vectores, la similitud entre artículos se obtiene mediante el coseno del ángulo (suma del producto de las componentes correspondientes).

3. Diseño e implementación

Nuestro sistema compara documentos según su contenido, calculando qué tan parecidos son entre sí. Para lograrlo, dividimos el trabajo en varias partes que se ejecutan de forma ordenada dentro del programa.

3.1. Funcionamiento del programa

1. Lectura de documentos: se cargan los textos que se van a analizar desde una carpeta del proyecto.
2. Preprocesamiento: se limpia cada texto eliminando palabras vacías, signos de puntuación y se transforman las palabras a su forma base mediante lematización.
3. Cálculo de medidas: se calculan los valores TF, IDF y TF–IDF, que indican la importancia de cada palabra dentro de un documento y en el conjunto total.
4. Cálculo de similitud: se comparan los documentos entre sí usando la similitud del coseno, que devuelve un valor entre 0 y 1 según cuánto se parecen.
5. Visualización de resultados: el programa muestra en pantalla las tablas generadas y la matriz de similitud, que también pueden guardarse para analizarlas más fácilmente.

3.2. Herramientas utilizadas

Para el desarrollo se emplearon las siguientes herramientas y tecnologías principales:

- Angular: framework principal usado para la estructura de la aplicación, la gestión de componentes y la lógica del sistema de recomendación.
- TypeScript: Lenguaje base de Angular para mejorar la organización del código.
- HTML y CSS/SCSS: Usados para el diseño y la interfaz web.
- Node.js y npm: Necesarios para instalar dependencias y ejecutar el entorno de desarrollo.

- Angular CLI: Herramienta que facilita la creación, compilación y despliegue del proyecto.

3.3. Pruebas y verificación

En nuestro programa hicimos varias pruebas para asegurarnos de que cada parte del programa funcionaba correctamente. Primero se probó con pocos documentos, y luego con los conjuntos completos, comprobando que los valores de similitud coincidían con lo esperado.

4.Conjunto de datos empleados

4.1. Documentos procesados (repositorio)

- `document-01.txt` … `document-10.txt`: 10 relatos en inglés.
- `el_quijote.txt`: texto de El Quijote (español).

4.2. Ficheros auxiliares de preprocesado

- `stop-words-en.txt` (stop words inglés)
- `stop-words-es.txt` (stop words español)
- `corpus-en.json` (lematización inglés)
- `corpus-es.json` (lematización español)

4.3. Ficheros propuestos por nuestro grupo

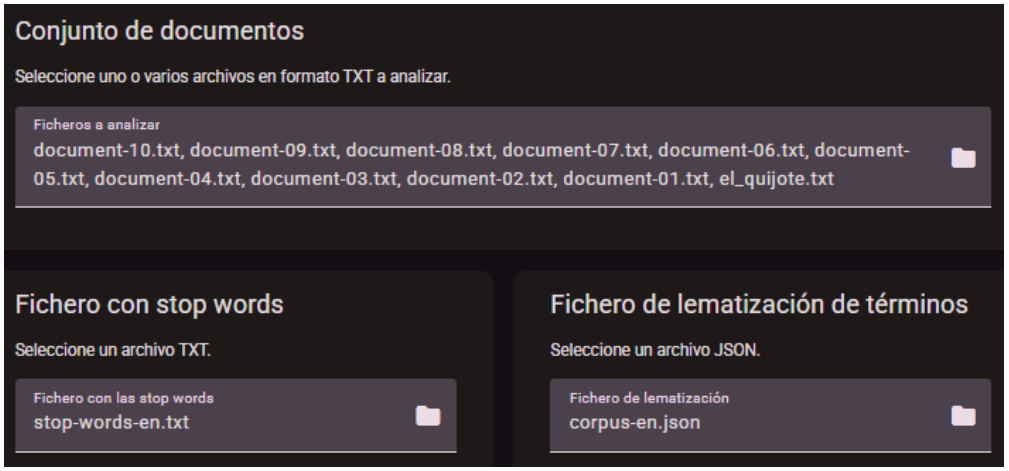
- `Prueba-1.txt` … `Prueba-10.txt`: 10 textos genéricos en inglés.

5. Resultados

En este apartado se presentan los resultados obtenidos tras la ejecución del sistema de recomendación desarrollado sobre los ejemplos proporcionados por el profesor y los ejemplos elaborados por nuestro equipo. Las capturas de pantalla que se ven a continuación muestran el cálculo de las métricas TF, IDF y TF–IDF, así como las matrices de similitud entre documentos. Cabe resaltar que por cada tabla de cada documento se mostrará un número limitado de resultados debido a la extensión de los

datos. Sin embargo, tanto los resultados al completo como los ficheros utilizados pueden encontrarse en las referencias al final de este informe.

5.1. Resultados con los documentos de ejemplo del profesor



document-01.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.3617	0	0
2	accept	1	2.3979	2.3979
3	acceptance	1	0.0953	0.0953
4	afraid	1.301	1.0116	1.3161
5	air	1.6021	0.0953	0.1527

document-02.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.301	0	0
2	acceptance	1	0.0953	0.0953
3	achieve	1	2.3979	2.3979
4	admire	1	1.7047	1.7047
5	air	1.699	0.0953	0.1619

document-03.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.3979	0	0
2	acceptance	1	0.0953	0.0953
3	achievement	1	2.3979	2.3979
4	activity	1	2.3979	2.3979
5	air	1.6021	0.0953	0.1527

document-04.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.3802	0	0
2	acceptance	1	0.0953	0.0953
3	air	1.4771	0.0953	0.1408
4	alive	1	0.3185	0.3185
5	allow	1.301	0.3185	0.4143

document-05.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.301	0	0
2	absence	1	1.7047	1.7047
3	acceptance	1	0.0953	0.0953
4	afraid	1	1.0116	1.0116
5	air	1.4771	0.0953	0.1408

document-06.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.2304	0	0
2	acceptance	1	0.0953	0.0953
3	afraid	1	1.0116	1.0116
4	air	1.6021	0.0953	0.1527
5	alive	1	0.3185	0.3185

document-07.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.301	0	0
2	acceptance	1	0.0953	0.0953
3	air	1.7782	0.0953	0.1695
4	allow	1	0.3185	0.3185
5	ask	1	0.6061	0.6061

document-08.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.2788	0	0
2	acceptance	1	0.0953	0.0953
3	air	1.4771	0.0953	0.1408
4	alive	1.301	0.3185	0.4143
5	allow	1.301	0.3185	0.4143

document-09.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.2788	0	0
2	acceptance	1	0.0953	0.0953
3	afraid	1	1.0116	1.0116
4	air	1.6021	0.0953	0.1527
5	alive	1	0.3185	0.3185

document-10.txt

Índice	Término	TF	IDF	TF-IDF
1	a	2.2553	0	0
2	absence	1	1.7047	1.7047
3	acceptance	1	0.0953	0.0953
4	air	1.6021	0.0953	0.1527
5	alive	1	0.3185	0.3185

el_quijote.txt

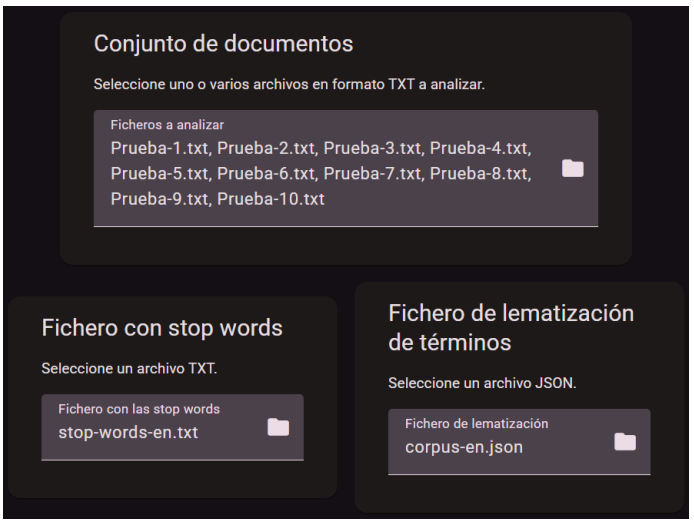
Índice	Término	TF	IDF	TF-IDF
56	abad	1	2.3979	2.3979
57	abadejo	1.301	2.3979	3.1197
58	abades	1	2.3979	2.3979
59	abadesa	1	2.3979	2.3979
60	abajarse	1.301	2.3979	3.1197

Similaridades entre documentos

Documento A	Documento B	Similaridad
document-01.txt	document-02.txt	0.1616
document-01.txt	document-03.txt	0.1559
document-01.txt	document-04.txt	0.1876
document-01.txt	document-05.txt	0.169
document-01.txt	document-06.txt	0.1845
document-01.txt	document-07.txt	0.1601
document-01.txt	document-08.txt	0.1519
document-01.txt	document-09.txt	0.1189
document-01.txt	document-10.txt	0.1086
document-01.txt	el_quijote.txt	0.0017

5.2. Resultados con los documentos de ejemplo de nuestro equipo

En este apartado se muestran los resultados obtenidos con los 10 documentos creados por nuestro grupo usando nuestro programa. Cada texto trataba un tema distinto, como astronomía, inteligencia artificial, evolución o los océanos. A continuación, se muestran las imágenes de las tablas con los resultados:



Prueba-1.txt

Índice	Término	TF	IDF	TF-IDF
16	accelerate	1.301	0.9163	1.1921
17	active	1.301	2.3026	2.9957
18	activity	1.301	0.9163	1.1921
19	advance	1	0.5108	0.5108
20	affect	1	0.5108	0.5108

Prueba-2.txt

Índice	Término	TF	IDF	TF-IDF
6	address	1	0.6931	0.6931
7	ads	1	2.3026	2.3026
8	advance	1	0.5108	0.5108
9	affect	1.4771	0.5108	0.7546
10	aging	1	2.3026	2.3026

Prueba-3.txt

Índice	Término	TF	IDF	TF-IDF
11	act	1	2.3026	2.3026
12	action	1.301	1.6094	2.0939
13	activates	1	2.3026	2.3026
14	activating	1	2.3026	2.3026
15	activity	1	0.9163	0.9163

Prueba-4.txt

Índice	Término	TF	IDF	TF-IDF
1	abnormalities	1	2.3026	2.3026
2	accelerate	1	0.9163	0.9163
3	accessibility	1	2.3026	2.3026
4	accountability	1	2.3026	2.3026
5	accuracy	1.6021	2.3026	3.6889

Prueba-5.txt

Índice	Término	TF	IDF	TF-IDF
16	a	1	0.5108	0.5108
17	absence	1	1.6094	1.6094
18	abyssal	1.301	2.3026	2.9957
19	accumulate	1	2.3026	2.3026
20	acidification	1	1.6094	1.6094

Prueba-6.txt

Índice	Término	TF	IDF	TF-IDF
11	accelerate	1	0.9163	0.9163
12	accurately	1	2.3026	2.3026
13	achieve	1	1.6094	1.6094
14	activities	1	1.204	1.204
15	adapt	1	1.204	1.204

Prueba-7.txt

Índice	Término	TF	IDF	TF-IDF
1	a	1	0.5108	0.5108
2	ability	1	2.3026	2.3026
3	access	1	1.6094	1.6094
4	achieve	1	1.6094	1.6094
5	adaptation	1	0.9163	0.9163

Prueba-8.txt

Índice	Término	TF	IDF	TF-IDF
11	adapt	1	1.204	1.204
12	adaptation	1.699	0.9163	1.5568
13	address	1.301	0.6931	0.9018
14	affect	1.9031	0.5108	0.9721
15	agricultural	1.4771	2.3026	3.4012

Prueba-9.txt

Índice	Término	TF	IDF	TF-IDF
1	acceptable	1.301	2.3026	2.9957
2	access	1.9031	1.6094	3.0629
3	activities	1.301	1.204	1.5664
4	address	1	0.6931	0.6931
5	adherence	1	2.3026	2.3026

Prueba-10.txt

Índice	Término	TF	IDF	TF-IDF
1	3d	1.7782	2.3026	4.0943
2	4d	1	2.3026	2.3026
3	accessories	1	2.3026	2.3026
4	add	1	2.3026	2.3026
5	additive	1.4771	2.3026	3.4012

Documento A	Documento B	Similaridad
Prueba-1.txt	Prueba-10.txt	0.1263
Prueba-1.txt	Prueba-2.txt	0.0324
Prueba-1.txt	Prueba-3.txt	0.0353
Prueba-1.txt	Prueba-4.txt	0.054
Prueba-1.txt	Prueba-5.txt	0.0244
Prueba-1.txt	Prueba-6.txt	0.0463
Prueba-1.txt	Prueba-7.txt	0.0354
Prueba-1.txt	Prueba-8.txt	0.0252
Prueba-1.txt	Prueba-9.txt	0.0281
Prueba-10.txt	Prueba-2.txt	0.0657

6. Conclusiones

El sistema desarrollado cumple con los objetivos planteados: procesa documentos en texto plano, realiza un preprocesado consistente (eliminación de stop-words y lematización), calcula las métricas TF, IDF y TF–IDF según las fórmulas descritas, normaliza los vectores resultantes y obtiene medidas de similitud mediante la similitud del coseno. Las tablas generadas y la matriz de similitud muestran coherencia con la implementación teórica, sobre el conjunto evaluado las similitudes entre documentos son, en su mayoría, bajas a moderadas, lo que refleja una diversidad temática del conjunto y confirma que el método captura coincidencias léxicas relevantes pero no relaciones semánticas profundas. Asimismo, el sistema distingue correctamente textos en distinto idioma, por lo que no recomienda entre documentos en inglés y en español.

A pesar de su validez académica, el prototipo presenta oportunidades claras de mejora. Entre los aspectos a mejorar estaría:

- Reforzar el preprocesado (se evalúan por ejemplo muchos números cuya importancia para lo que nos compete es relativa y se podría ajustar)
- Implementar algún método para detectar el idioma e informar de lo poco idóneo que es usar un determinado archivo de lematización o de stop words en ese caso.
- Incluir técnicas más avanzadas que tengan en cuenta la semántica.
- Mejorar la interfaz visual.

En general, el proyecto nos ayudó a comprender de forma práctica cómo funcionan los sistemas de recomendación basados en contenido y a aplicar conceptos teóricos como TF–IDF y la similitud del coseno con buenos resultados.

7. Bibliografía

- Sistemas de recomendación. Google Slides.
https://docs.google.com/presentation/d/1G9dtDODvXBQpYzOCBI_7VCBKcOD4SVZZZrVUBLxGHnE/edit?slide=id.ge76dc206a814a53_191#slide=id.ge76dc206a814a53_191

- Repositorio del sistema de recomendación basado en el contenido.
https://github.com/LuisChineaRangel/sistemas_recomendacion_basado_contenido

- Cristofer Juan Expósito Izquierdo. Documentos de ejemplo. Github.
https://github.com/ull-cs/gestion-conocimiento/tree/main/recommeder-systems/examples-documents

- Cristofer Juan Expósito Izquierdo. Stop words. Github.
https://github.com/ull-cs/gestion-conocimiento/tree/main/recommeder-systems/stop-words

- Cristofer Juan Expósito Izquierdo. Lematización. Github.
https://github.com/ull-cs/gestion-conocimiento/tree/main/recommeder-systems/corpus

- Resultados de las pruebas realizadas con el sistema de recomendación. Google Drive.
https://drive.google.com/drive/folders/1mtT4rnoYuFUhiV1JP0aZyQxPgbn1JXtf?usp=sharing

- Ficheros de ejemplo propuestos por nuestro grupo de trabajo. Google Drive.
https://drive.google.com/drive/folders/1Krhf_6KPe_6pvWqTib3N0eRRu1t4oBHi?usp=sharing