SISTEMAS DE RECOMENDACIÓN BASADOS EN CONTENIDO:

Gestión del Conocimiento en las Organizaciones

Daniel Pérez Rodríguez (alu0101541006@ull.edu.es)
Saúl Ruiz Fernández (alu0101540917@ull.edu.es)
Ismael Rojas Torres (alu0101539393@ull.edu.es)

Índice

ANÁLISIS:

Caso: Texto Quijote

Primero probaremos el funcionamiento del programa empleando un único archivo, que será el texto del Quijote. Para ello, lo seleccionamos en los ficheros de entrada y escogemos los ficheros con palabras de parada y de lematización para español.

RESULTADOS:

Documento 1

Índice	Término	TF	IDF	TF-IDF
11	que	0.072	0.000	0.000
2	de	0.066	0.000	0.000
28	a	0.057	0.000	0.000
15	y	0.056	0.000	0.000
23	no	0.020	0.000	0.000
73	se	0.017	0.000	0.000

Similitud coseno

	Doc 1
Doc 1	0.000

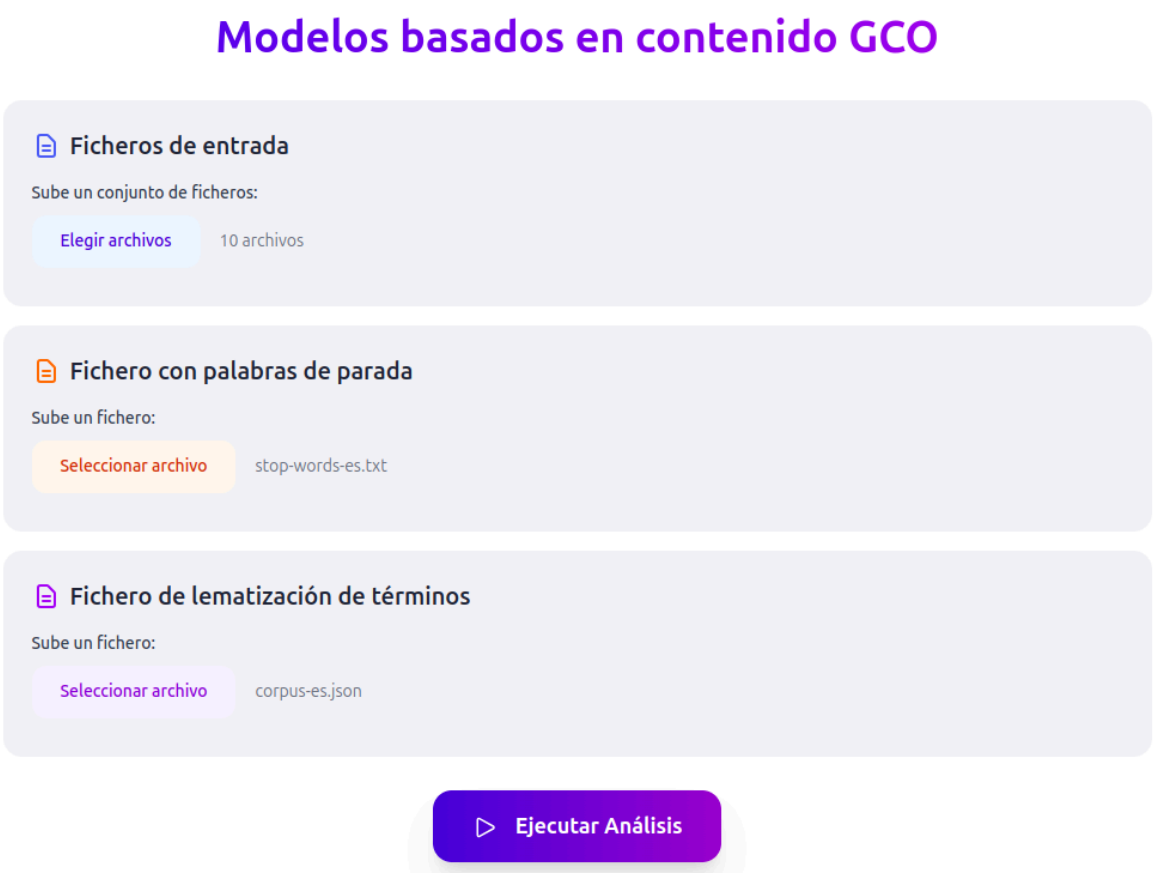
Conclusiones:

Como podemos observar el programa nos muestra correctamente el resultado para un único fichero, en la primera tabla vemos los términos ordenados según su frecuencia de aparición (la frecuencia de aparición TF mide cuántas veces aparece un término respecto al tamaño del documento), y además nos muestra el IDF (Que mide lo importante que es una palabra en el conjunto de documentos) y la resta de TF-IDF. Pero como sólo hay un fichero de datos el IDF no tiene significado.
En lo que respecta a la similitud coseno, como sólo hay un fichero el IDF es 0 y por tanto log(1)= 0 lo que provoca que la similitud coseno sea 0 ya que ambos vectores son [0,0,0...,0].

Caso: Textos 11-20

Ahora cargaremos nuestros diez ficheros personalizados a la vez, formados por textos periodísticos y fragmentos de novelas de la literatura española. Una vez más, usaremos las palabras de parada y el fichero de lematización para español.

RESULTADOS:



Documento 1

Índice	Término	TF	IDF	TF-IDF
2	de	0.106	0.000	0.000
8	y	0.062	0.000	0.000
32	a	0.031	0.000	0.000
63	que	0.030	0.000	0.000
3	soledad	0.014	1.609	0.023
1	años	0.013	0.693	0.009

Documento 2

Índice	Término	TF	IDF	TF-IDF
2	de	0.107	0.000	0.000

Similitud coseno

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Doc 1	1.000	0.015	0.019	0.012	0.013	0.008	0.106	0.012	0.024	0.032
Doc 2	0.015	1.000	0.072	0.070	0.016	0.009	0.007	0.012	0.006	0.008
Doc 3	0.019	0.072	1.000	0.126	0.011	0.017	0.013	0.010	0.004	0.007
Doc 4	0.012	0.070	0.126	1.000	0.016	0.012	0.009	0.007	0.013	0.008
Doc 5	0.013	0.016	0.011	0.016	1.000	0.039	0.004	0.101	0.057	0.015
Doc 6	0.008	0.009	0.017	0.012	0.039	1.000	0.006	0.032	0.021	0.013
Doc 7	0.106	0.007	0.013	0.009	0.004	0.006	1.000	0.013	0.005	0.029
Doc 8	0.012	0.012	0.010	0.007	0.101	0.032	0.013	1.000	0.030	0.023
Doc 9	0.024	0.006	0.004	0.013	0.057	0.021	0.005	0.030	1.000	0.016

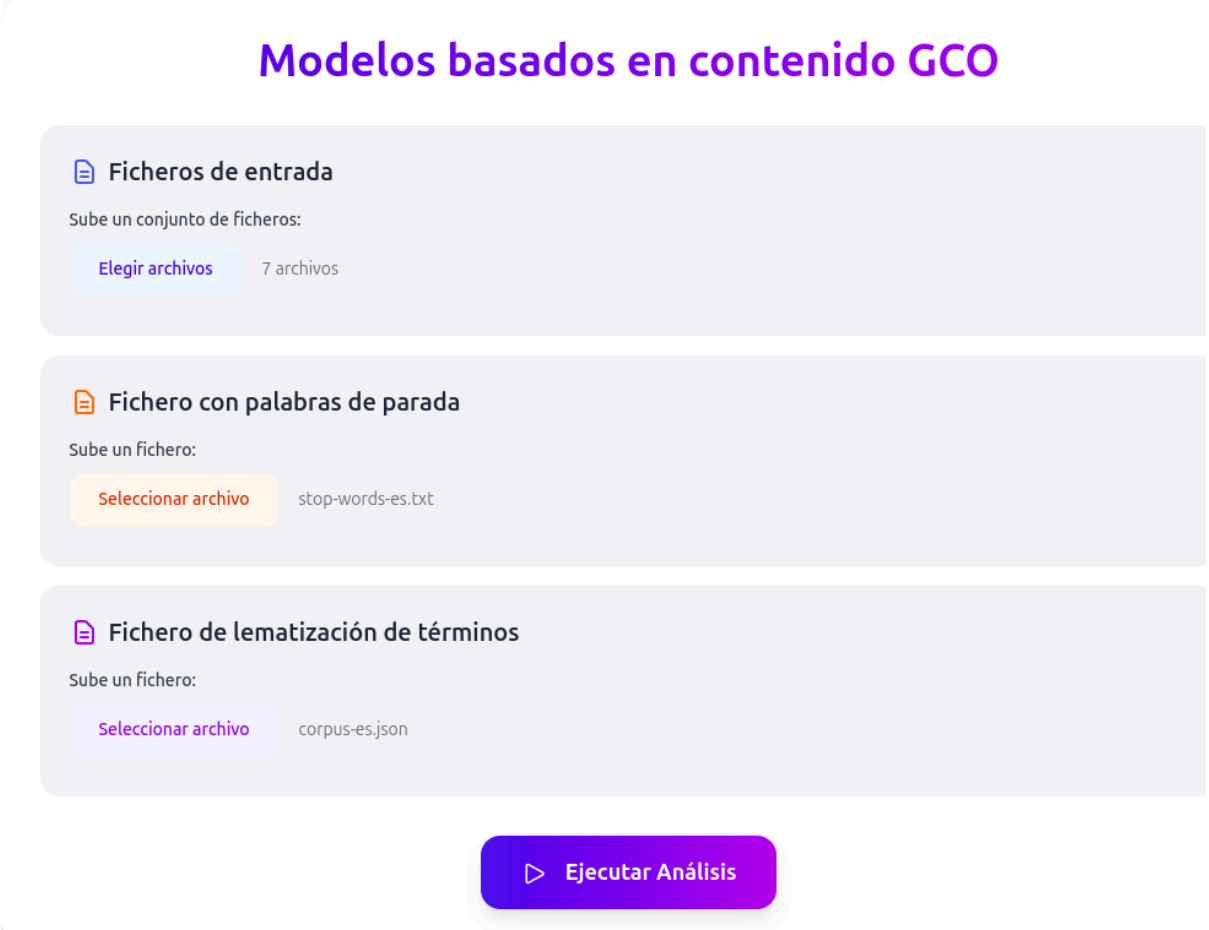
Conclusiones:

Ahora, gracias a que hemos puesto múltiples ficheros, podemos observar la utilidad real de la aplicación, ya que ahora la TF e IDF de cáda término tienen valores significativos. Los términos con IDF de 0 indican que están presentes en todos los documentos y por tanto su valor es bajo, mientras que los que tienen valores más altos indican que se trata de palabras más importantes. Además podemos ver los grados de similitud entre los diferentes textos gracias a la similitud coseno, pudiendo así apreciar temáticas similares o un vocabulario parecido entre los diferentes autores.

Caso: Textos 15-20 + Quijote

Ahora añadiremos el Quijote a la lista de antes para ver la similitud en la forma de escribir de Cervantes comparada con autores y periodistas más modernos.

RESULTADOS:



Documento 1

Índice	Término	TF	IDF	TF-IDF
13	de	0.102	0.000	0.000
15	kael	0.029	1.792	0.052
22	que	0.027	0.000	0.000
3	y	0.022	0.000	0.000
41	no	0.018	0.000	0.000
163	lyra	0.018	1.792	0.033

Documento 2

Índice	Término	TF	IDF	TF-IDF
13	de	0.068	0.000	0.000

Similitud coseno

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6
Doc 1	1.000	0.003	0.022	0.013	0.008	0.023
Doc 2	0.003	1.000	0.008	0.002	0.019	0.012
Doc 3	0.022	0.008	1.000	0.024	0.013	0.039
Doc 4	0.013	0.002	0.024	1.000	0.013	0.036
Doc 5	0.008	0.019	0.013	0.013	1.000	0.019
Doc 6	0.023	0.012	0.039	0.036	0.019	1.000

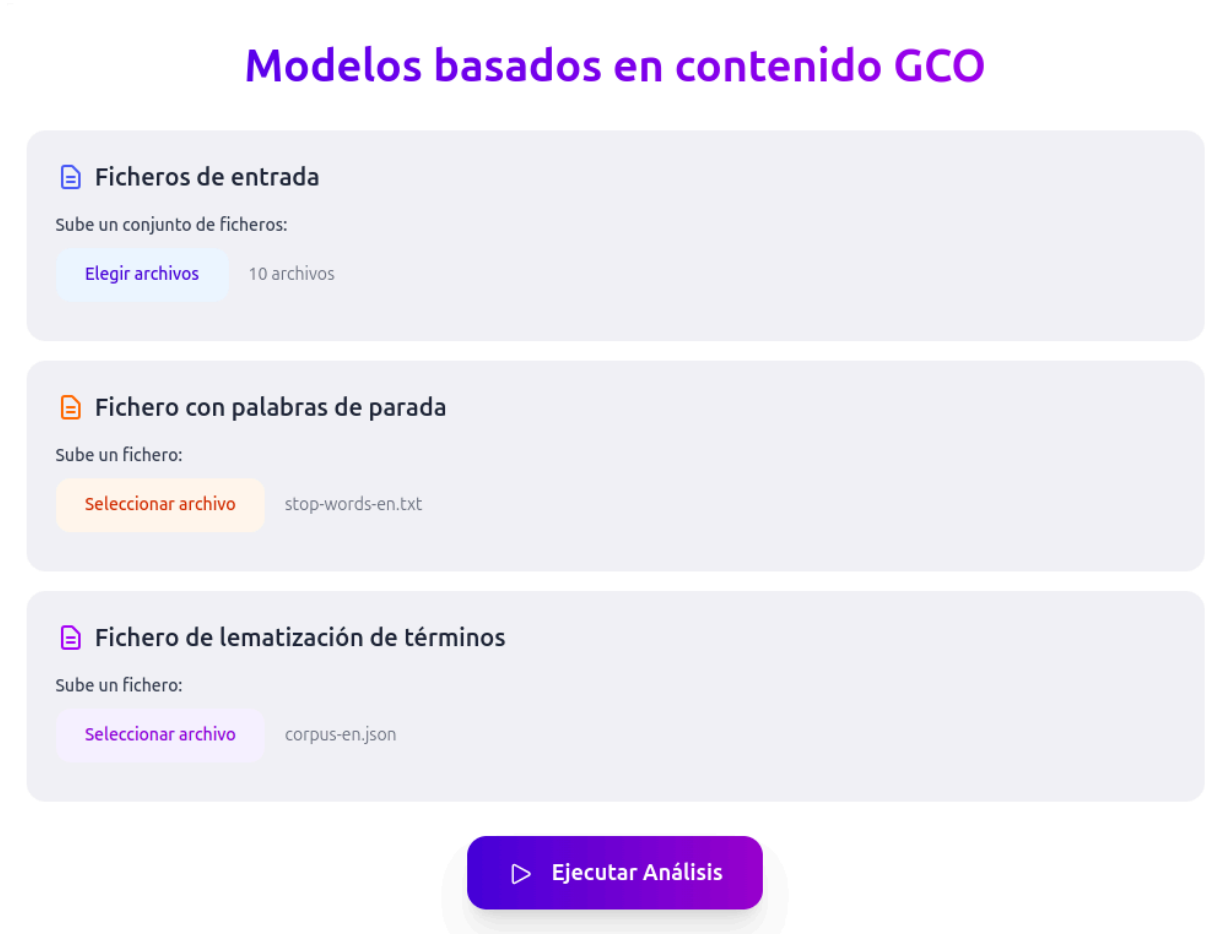
Conclusiones:

Como podemos apreciar el Quijote (DOC1) tiene grados de similitud coseno muy bajos comparados con el resto de textos, esto muy probablemente se debe a que usan un vocabulario muy distinto pues provienen de diferentes épocas y entornos.

Caso: Textos 1-10

Ahora probaremos el funcionamiento del programa con textos en inglés. Para ello tendremos que seleccionar un fichero de términos de parada especializado así como un fichero de lematización de términos específico para el inglés.

RESULTADOS:



20	ground	0.005	0.000	0.000
38	alive	0.005	0.223	0.001
45	quietly	0.005	0.357	0.002
49	branch	0.005	0.916	0.004
56	part	0.005	0.916	0.004
58	dirt	0.005	0.105	0.000
60	side	0.005	1.204	0.006
61	open	0.005	0.000	0.000

Documento 2

Índice	Término	TF	IDF	TF-IDF
3	i	0.092	0.000	0.000

Similitud coseno

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Doc 1	1.000	0.138	0.127	0.112	0.142	0.167	0.084	0.063	0.113	0.103
Doc 2	0.138	1.000	0.124	0.087	0.157	0.163	0.088	0.138	0.077	0.099
Doc 3	0.127	0.124	1.000	0.128	0.152	0.151	0.138	0.132	0.137	0.141
Doc 4	0.112	0.087	0.128	1.000	0.135	0.157	0.114	0.203	0.148	0.139
Doc 5	0.142	0.157	0.152	0.135	1.000	0.156	0.088	0.141	0.125	0.162
Doc 6	0.167	0.163	0.151	0.157	0.156	1.000	0.080	0.138	0.137	0.137
Doc 7	0.084	0.088	0.138	0.114	0.088	0.080	1.000	0.154	0.119	0.159
Doc 8	0.063	0.138	0.132	0.203	0.141	0.138	0.154	1.000	0.157	0.174
Doc 9	0.113	0.077	0.137	0.148	0.125	0.137	0.119	0.157	1.000	0.138

Conclusiones:

Como podemos ver, obtenemos resultados similares a las pruebas anteriores, con la diferencia de que la similitud entre los textos es incluso menor que antes. También sería interesante considerar la posibilidad de añadir palabras como “I” o “a” a la lista de palabras de parada, con el objetivo de obtener resultados más significativos al analizar textos en inglés, pese a que ya el IDF y TF-IDF nos indican que tienen poca importancia.