SISTEMAS DE RECOMENDACIÓN BASADOS EN CONTENIDO:

Gestión del Conocimiento en las Organizaciones

Daniel Pérez Rodríguez (alu0101541006@ull.edu.es)
Saúl Ruíz 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	-0.693	-0.050
2	de	0.066	-0.693	-0.046
28	a	0.057	-0.693	-0.039
15	y	0.056	-0.693	-0.039
23	no	0.020	-0.693	-0.014
73	se	0.017	-0.693	-0.012

Similitud coseno

	Doc 1
Doc 1	1.000

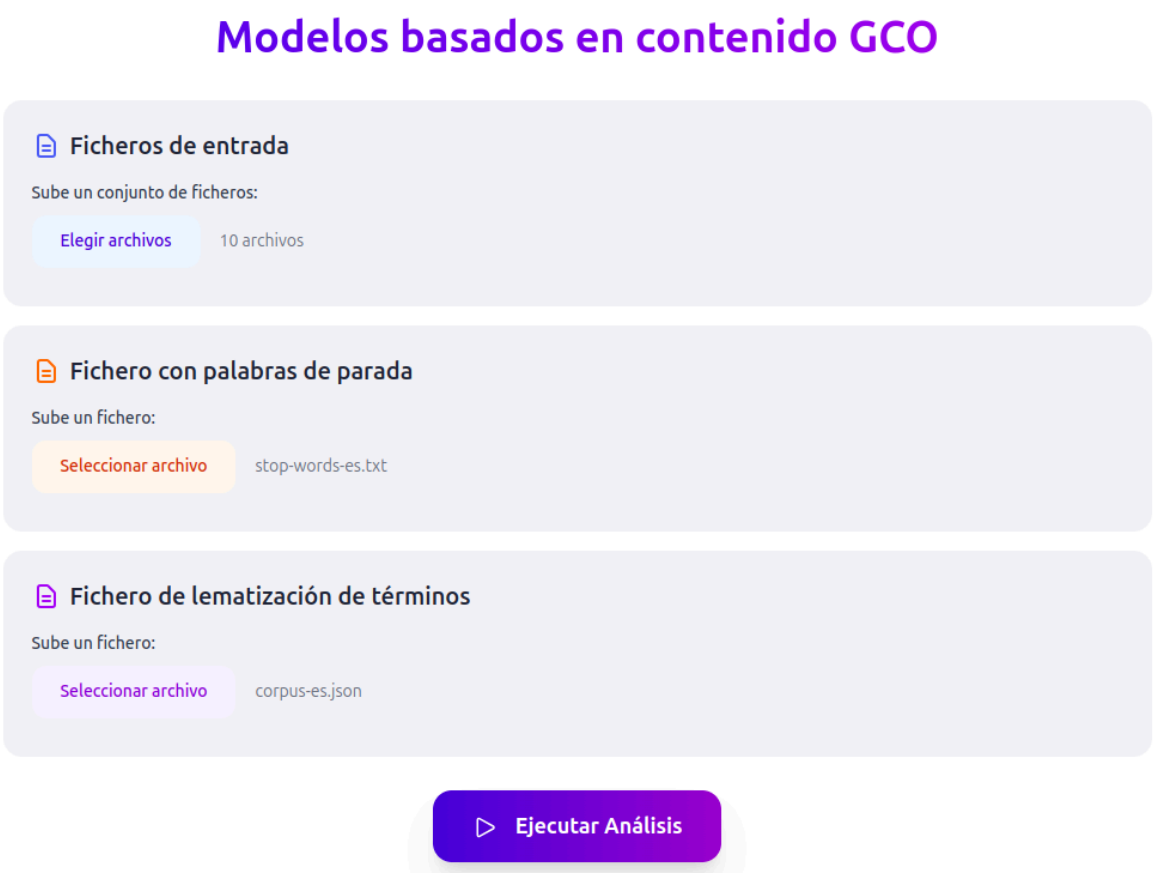
Conclusiones:

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, pero como sólo hay un fichero de datos el IDF da negativo, ya que según nuestra implementación al calcular el IDF sumamos 1 al denominador para evitar valores negativos, lo que en este caso resulta en log(1/(1+1)) = -0.693.
En lo que respecta a la similitud coseno, como sólo hay un fichero la similitud consigo mismo es perfecta, es decir, que vale 1.

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:



163	compromiso	0.002	1.609	0.004
165	constante	0.002	0.693	0.002
167	humanos	0.002	1.609	0.004
169	fundamentales	0.002	1.609	0.004
171	requiere	0.002	1.609	0.004
172	paciencia	0.002	1.609	0.004
173	da	0.002	1.609	0.004
174	sentido	0.002	1.204	0.003

Documento 2

Índice	Término	TF	IDF	TF-IDF
2	de	0.087	-0.095	-0.008

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	0.032	0.040	0.042	0.028	0.034	0.027	0.029	0.028	0.057
Doc 2	0.032	1.000	0.049	0.015	0.039	0.082	0.033	0.028	0.027	0.048
Doc 3	0.040	0.049	1.000	0.024	0.051	0.131	0.027	0.035	0.034	0.036
Doc 4	0.042	0.015	0.024	1.000	0.016	0.017	0.022	0.029	0.021	0.130
Doc 5	0.028	0.039	0.051	0.016	1.000	0.063	0.033	0.043	0.030	0.032
Doc 6	0.034	0.082	0.131	0.017	0.063	1.000	0.042	0.042	0.043	0.042
Doc 7	0.027	0.033	0.027	0.022	0.033	0.042	1.000	0.165	0.102	0.040
Doc 8	0.029	0.028	0.035	0.029	0.043	0.042	0.165	1.000	0.108	0.053
Doc 9	0.028	0.027	0.034	0.021	0.030	0.043	0.102	0.108	1.000	0.045

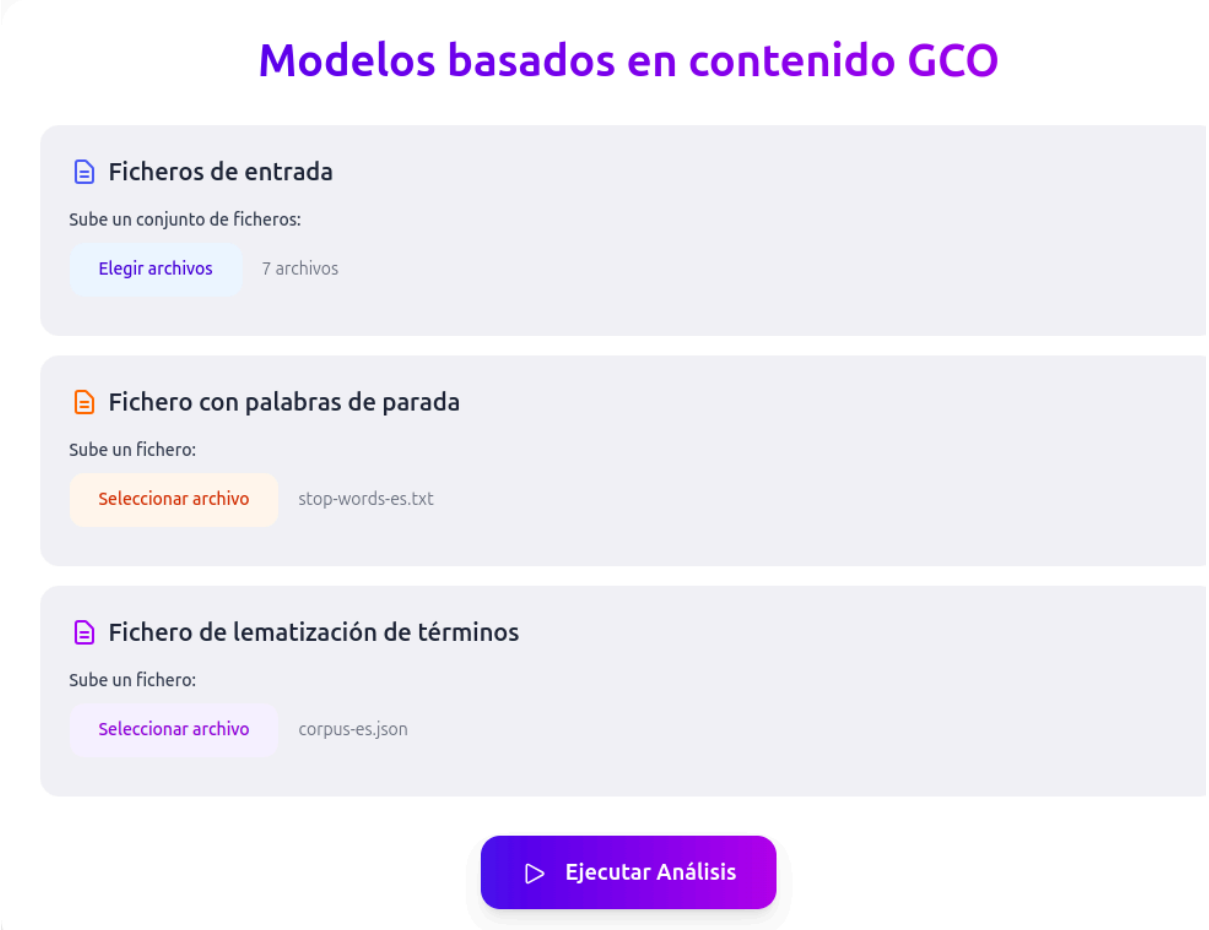
Conclusiones:

Ahora, gracias a que hemos puesto múltiples ficheros, podemos observar la utilidad real de la aplicación, ya que no sólo podemos ver la TF e IDF de cáda término en cada documento, sino que además podemos ver los grados de similitud entre los diferentes textos, pudiendo así apreciar temáticas similares entre los diferentes autores.

Caso: Textos 11-20 + 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:



1	quijote	0.005	1.253	0.007
132	l	0.005	1.253	0.007
236	sen	0.005	1.253	0.007
273	habi	0.005	1.253	0.007
168	tan	0.005	0.847	0.004
1610	sancho	0.004	1.253	0.006
333	ver	0.004	0.336	0.001
53	an	0.004	1.253	0.005

Documento 2

Índice	Término	TF	IDF	TF-IDF
2	de	0.074	-0.134	-0.010

Similitud coseno

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7
Doc 1	1.000	0.123	0.134	0.132	0.068	0.106	0.131
Doc 2	0.123	1.000	0.067	0.070	0.056	0.055	0.076
Doc 3	0.134	0.067	1.000	0.087	0.040	0.073	0.117
Doc 4	0.132	0.070	0.087	1.000	0.044	0.078	0.148
Doc 5	0.068	0.056	0.040	0.044	1.000	0.034	0.044
Doc 6	0.106	0.055	0.073	0.078	0.034	1.000	0.096
Doc 7	0.131	0.076	0.117	0.148	0.044	0.096	1.000

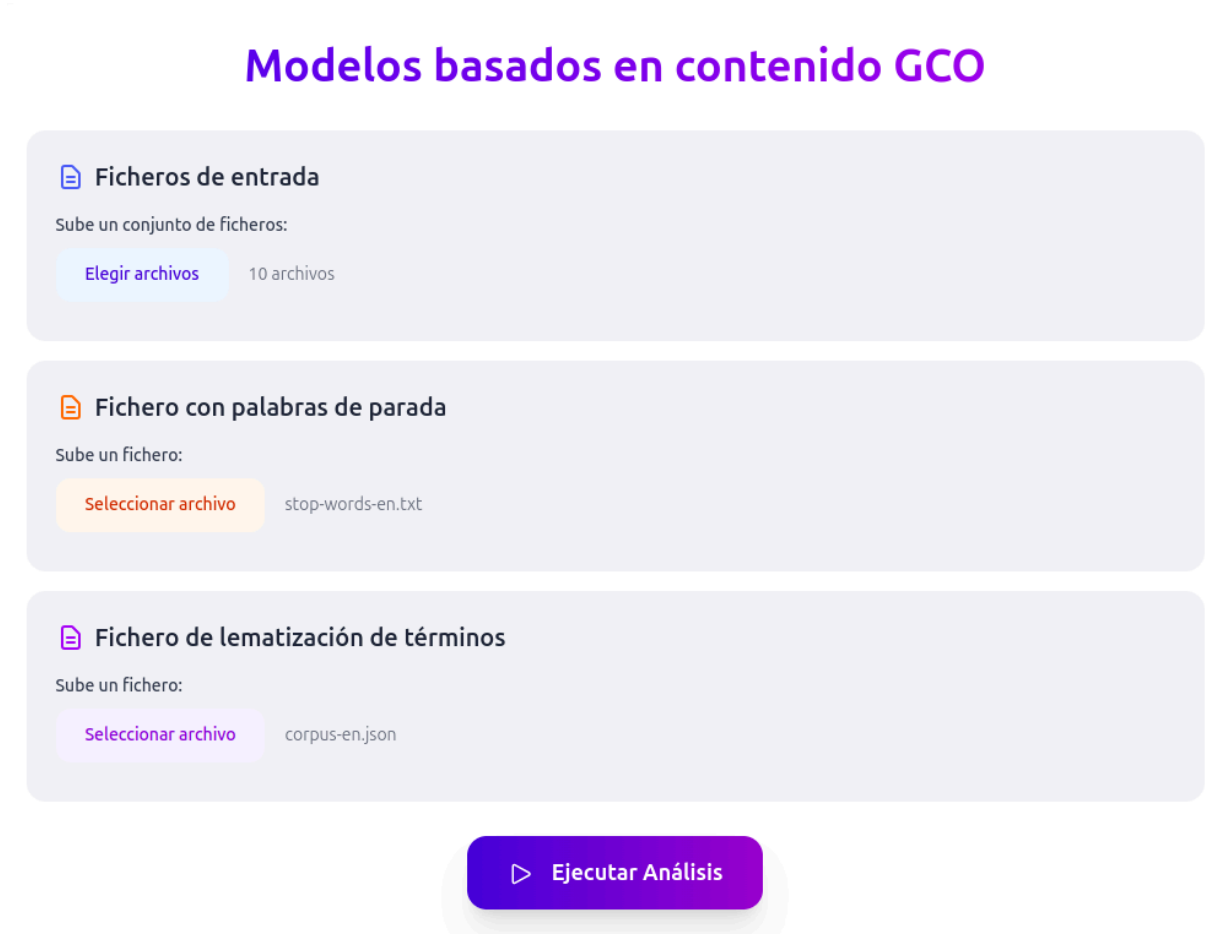
Conclusiones:

Como podemos apreciar el Quijote (DOC1) tiene grados de similitud muy bajos comparados con el resto de textos debido a que usan un vocabulario muy distinto.

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:



Documento 1

Índice	Término	TF	IDF	TF-IDF
20	i	0.099	-0.095	-0.009
27	a	0.042	-0.095	-0.004
25	t	0.031	-0.095	-0.003
23	lake	0.019	-0.095	-0.002
53	feel	0.016	-0.095	-0.002
40	trees	0.014	-0.095	-0.001

Documento 2

Índice	Término	TF	IDF	TF-IDF
20	i	0.093	-0.095	-0.009

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	0.288	0.332	0.306	0.311	0.333	0.306	0.292	0.219	0.252
Doc 2	0.288	1.000	0.328	0.278	0.324	0.316	0.324	0.298	0.207	0.273
Doc 3	0.332	0.328	1.000	0.321	0.341	0.345	0.383	0.306	0.271	0.243
Doc 4	0.306	0.278	0.321	1.000	0.276	0.280	0.294	0.300	0.216	0.249
Doc 5	0.311	0.324	0.341	0.276	1.000	0.374	0.361	0.334	0.301	0.348
Doc 6	0.333	0.316	0.345	0.280	0.374	1.000	0.341	0.336	0.296	0.322
Doc 7	0.306	0.324	0.383	0.294	0.361	0.341	1.000	0.307	0.228	0.290
Doc 8	0.292	0.298	0.306	0.300	0.334	0.336	0.307	1.000	0.248	0.287
Doc 9	0.219	0.207	0.271	0.216	0.301	0.296	0.228	0.248	1.000	0.260

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.