Informe Práctica.
Sistemas de Recomendación.
Modelos Basados en Contenido

Gestión del Conocimiento en las Organizaciones

Alberto Antonio Hernández Hernández

Eduardo Socas Luis

Marcial Álvarez Parejo

Pablo González Martín

PE101

Eduardo Socas Luis
Marcial Álvarez Parejo
Pablo Gónzalez Martín
Alberto Antonio Hernández Hernández

Índice

Eduardo Socas Luis
Marcial Álvarez Parejo
Pablo Gónzalez Martín
Alberto Antonio Hernández Hernández

Introducción

En esta práctica se desarrolla un sistema de recomendación basado en contenido.

El objetivo principal es procesar un conjunto de documentos en texto plano, aplicar técnicas de procesamiento de lenguaje natural (tokenización, eliminación de stop-words, lematización) y generar representaciones vectoriales mediante TF, IDF y TF-IDF.

Análisis realizado

Para el análisis utilizamos los documentos en formato .txt proporcionados en el repositorio de la práctica. A cada documento se le aplicaron los siguientes pasos:

a) Normalización

- Conversión de todo el texto a minúsculas.
- Eliminación de signos de puntuación, números y caracteres no alfabéticos.

b) Tokenización

- Se dividió cada documento en tokens utilizando el separador por espacios y reglas básicas de segmentación.
- Esto permitió obtener listas de palabras con las que trabajar de manera uniforme.

c) Eliminación de stop-words

Eduardo Socas Luis
Marcial Álvarez Parejo
Pablo Gónzalez Martín
Alberto Antonio Hernández Hernández

- Utilizando el fichero de palabras vacías, se eliminaron términos como: de, la, que, en, por, y, un…
- Esta fase redujo el número de términos irrelevantes y permitió aumentar la calidad de la representación vectorial.

d) Lematización

- Se aplicó el fichero de lemas proporcionado, unificando palabras con significados equivalentes.
- Este paso redujo la redundancia y mejoró la representatividad de los términos.

2. Cálculo de TF, IDF y TF-IDF

a) Frecuencia de término (TF)

Esto nos permite observar qué palabras son más representativas de cada texto.

b) Frecuencia inversa de documento (IDF)

Pudimos observar que los términos que aparecen en muchos documentos tienen un IDF bajo, mientras que los términos particulares de un sólo documento tienen un IDF alto.

c) TF-IDF

Se generan vectores que capturan la importancia real de cada término en cada documento.

3. Similitud coseno entre cada par de documentos

La similitud coseno muestra:

- Valores cercanos a 1: documentos con vocabulario muy similar.
- Valores cercanos a 0: documentos casi sin relación semántica.

Eduardo Socas Luis
Marcial Álvarez Parejo
Pablo Gónzalez Martín
Alberto Antonio Hernández Hernández

En general, los resultados mostraron agrupaciones temáticas claras: los documentos que trataban temas similares presentan similitudes altas,

mientras que los documentos con vocabulario técnico o muy específico muestran mayor distancia frente al resto.

Conclusiones extraídas

1. **El preprocesamiento (stop-words + lematización) es esencial.**

   Sin estas fases, los términos repetitivos y las variaciones morfológicas habrían distorsionado los resultados, reduciendo la claridad de los vectores TF-IDF.

2. **TF-IDF proporciona una representación efectiva del contenido.**

   Permite captar el peso real de cada término en el documento y penaliza las palabras comunes que aparecen en todos los textos.

3. **La similitud coseno permite identificar correctamente documentos relacionados.**

   La métrica resultó robusta incluso cuando algunos documentos eran mucho más largos que otros.

4. **Los documentos con vocabulario especializado se muestran claramente diferenciados.**

5. **El modelo basado en contenido funciona especialmente bien cuando los textos tienen temas diferenciables.**

   En cambio, documentos demasiado cortos o extremadamente generales tienden a generar similitudes bajas y menos informativas.

Eduardo Socas Luis
Marcial Álvarez Parejo
Pablo Gónzalez Martín
Alberto Antonio Hernández Hernández

6. **La calidad de la lematización influye directamente en la precisión del sistema.**

   Cuanto mejor esté preparado el fichero de lematización, mayor coherencia semántica se logra en la representación vectorial.

7. **El método permite extenderse a muchos contextos prácticos.**

Resultados reales obtenidos

Matriz de Similaridad Coseno

	document-03.txt	document-04.txt	document-05.txt	document-06.txt	document-07.txt	document-08.txt	document-09.txt	document-10.txt	document-01.txt	document-02.txt
document-03.txt	1.000	0.869	0.876	0.873	0.838	0.864	0.853	0.854	0.846	0.821
document-04.txt	0.869	1.000	0.887	0.887	0.856	0.881	0.866	0.848	0.866	0.821
document-05.txt	0.876	0.887	1.000	0.891	0.851	0.878	0.874	0.866	0.863	0.831
document-06.txt	0.873	0.887	0.891	1.000	0.851	0.876	0.868	0.864	0.864	0.839
document-07.txt	0.838	0.856	0.851	0.851	1.000	0.846	0.852	0.835	0.829	0.793
document-08.txt	0.864	0.881	0.878	0.876	0.846	1.000	0.867	0.861	0.842	0.826
document-09.txt	0.853	0.866	0.874	0.868	0.852	0.867	1.000	0.856	0.846	0.804
document-10.txt	0.854	0.848	0.866	0.864	0.835	0.861	0.856	1.000	0.836	0.803
document-01.txt	0.846	0.866	0.863	0.864	0.829	0.842	0.846	0.836	1.000	0.818
document-02.txt	0.821	0.821	0.831	0.839	0.793	0.826	0.804	0.803	0.818	1.000

Matriz de similaridad coseno (para documentos en inglés)

- Indica qué tan parecidos son entre sí los documentos del corpus
- La similaridad en general es bastante alta, ya que no baja del 0.8
- La diagonal principal tiene 1.000 como resultado porque se trata de la similaridad entre los mismos documentos

Eduardo Socas Luis
Marcial Álvarez Parejo
Pablo Gónzalez Martín
Alberto Antonio Hernández Hernández

Resultados

document-03.txt

Índice	Término	TF	IDF	TF-IDF
1	day	0.007	1.000	0.007
2	quietly	0.002	1.318	0.003
3	sky	0.009	1.000	0.009
4	pale	0.002	1.201	0.003
5	i	0.116	1.000	0.116
6	leave	0.007	1.000	0.007
7	house	0.002	2.299	0.005
8	sun	0.007	1.000	0.007
9	start	0.005	1.606	0.008
10	rise	0.002	1.201	0.003
11	air	0.009	1.000	0.009
12	cool	0.002	1.201	0.003
13	freshness	0.002	1.606	0.004
14	lingers	0.002	2.705	0.006
15	warmth	0.002	2.705	0.006
16	morning	0.005	1.000	0.005
17	set	0.002	2.012	0.005
19	decide	0.002	1.318	0.003
20	walk	0.009	1.000	0.009
21	lake	0.017	1.000	0.017
23	mejust	0.002	2.705	0.006
24	a	0.059	1.000	0.059
25	notebook	0.009	1.000	0.009
27	pen	0.002	1.201	0.003
30	feel	0.014	1.000	0.014
31	write	0.014	1.000	0.014
33	trail	0.002	1.000	0.002
34	water	0.009	1.000	0.009
35	familiar	0.005	1.452	0.007
36	repeat	0.002	2.705	0.006

document-04.txt

Índice	Término	TF	IDF	TF-IDF
1	early	0.003	2.012	0.005
2	morning	0.008	1.000	0.008
3	i	0.111	1.000	0.111
4	decide	0.003	1.318	0.003
5	a	0.061	1.000	0.061
6	walk	0.013	1.000	0.013
7	sky	0.010	1.000	0.010
8	light	0.008	1.000	0.008
9	dawn	0.003	2.705	0.007
10	sun	0.005	1.000	0.005
11	high	0.003	2.705	0.007
12	paint	0.003	2.012	0.005
13	horizon	0.003	1.318	0.003
14	soft	0.005	1.201	0.006
15	gold	0.003	1.318	0.003
16	pale	0.003	1.201	0.003
17	pink	0.003	1.452	0.004
18	air	0.008	1.000	0.008
19	cool	0.003	1.201	0.003
20	chilly	0.003	2.705	0.007
21	carry	0.008	1.000	0.008
22	freshness	0.003	1.606	0.004
23	exist	0.005	1.452	0.007
24	start	0.005	1.606	0.008
25	day	0.005	1.000	0.005
28	notebook	0.010	1.000	0.010
30	pen	0.003	1.201	0.003
32	write	0.018	1.000	0.018
34	reach	0.005	1.000	0.005
35	lake	0.025	1.000	0.025

document-05.txt

Índice	Término	TF	IDF	TF-IDF
1	morning	0.008	1.000	0.008
2	i	0.116	1.000	0.116
3	leave	0.013	1.000	0.013
4	house	0.003	2.299	0.006
5	sun	0.008	1.000	0.008
6	rise	0.003	1.201	0.003
7	send	0.003	1.788	0.005
8	soft	0.003	1.201	0.003
9	light	0.008	1.000	0.008
10	horizon	0.003	1.318	0.003
11	shades	0.003	2.299	0.006
12	gold	0.003	1.318	0.003
13	pink	0.003	1.452	0.004
14	air	0.008	1.000	0.008
15	carry	0.008	1.000	0.008
16	coolness	0.003	2.299	0.006
17	night	0.003	1.606	0.004
18	clean	0.003	2.012	0.005
19	sharp	0.003	1.095	0.003
20	day	0.008	1.000	0.008
21	freshly	0.003	2.705	0.007
22	wash	0.003	2.705	0.007
24	set	0.003	2.012	0.005
25	lake	0.021	1.000	0.021
27	a	0.051	1.000	0.051
28	pen	0.003	1.201	0.003
30	small	0.010	1.000	0.010
31	notebook	0.010	1.000	0.010
34	write	0.015	1.000	0.015
36	path	0.005	1.000	0.005

document-06.txt

Índice	Término	TF	IDF	TF-IDF
1	morning	0.005	1.000	0.005
2	silence	0.008	1.318	0.010
3	sun	0.005	1.000	0.005
4	climb	0.003	1.788	0.005
5	horizon	0.003	1.318	0.003
6	spread	0.005	1.201	0.006
7	faint	0.003	2.299	0.006
8	streaks	0.003	2.299	0.006
9	gold	0.003	1.318	0.003
10	pale	0.003	1.201	0.003
11	pink	0.003	1.452	0.004
12	sky	0.008	1.000	0.008
13	air	0.010	1.000	0.010
14	feel	0.023	1.000	0.023
15	cool	0.003	1.201	0.003
16	light	0.010	1.000	0.010
17	touch	0.003	1.318	0.003
18	breath	0.003	2.299	0.006
19	night	0.003	1.606	0.004
20	i	0.111	1.000	0.111
21	decide	0.003	1.318	0.003
22	walk	0.008	1.000	0.008
23	lake	0.023	1.000	0.023
25	carry	0.005	1.000	0.005
26	a	0.043	1.000	0.043
27	notebook	0.010	1.000	0.010
29	pen	0.003	1.201	0.003
33	write	0.018	1.000	0.018
35	reach	0.008	1.000	0.008
36	water	0.013	1.000	0.013

document-07.txt

Índice	Término	TF	IDF	TF-IDF
1	day	0.003	1.000	0.003
2	softly	0.003	1.606	0.004
3	a	0.050	1.000	0.050
4	pale	0.003	1.201	0.003
5	light	0.008	1.000	0.008
6	stretch	0.008	1.000	0.008
7	horizon	0.003	1.318	0.003
8	sun	0.008	1.000	0.008
9	lift	0.005	1.000	0.005
10	earth	0.005	1.201	0.006
11	sky	0.008	1.000	0.008
12	carry	0.011	1.000	0.011
13	hints	0.003	2.705	0.007
14	gold	0.003	1.318	0.003
15	rise	0.003	1.201	0.003
16	air	0.016	1.000	0.016
17	sharp	0.003	1.095	0.003
18	cool	0.003	1.201	0.003
19	hold	0.008	1.000	0.008
20	night's	0.003	2.705	0.007
21	chill	0.003	2.705	0.007
22	i	0.090	1.000	0.090
23	decide	0.003	1.318	0.003
24	head	0.003	2.299	0.006
25	lake	0.021	1.000	0.021
28	meonly	0.003	2.299	0.006
30	small	0.011	1.000	0.011
31	notebook	0.011	1.000	0.011
33	pen	0.003	1.201	0.003
34	write	0.019	1.000	0.019

Una vez ejecutamos, obtenemos una columna por documento, en las que tenemos:

- Índice (1º ocurrencia del término en el documento)
- Término
- TF
- IDF
- TF-IDF

Todo esto tras evidentemente haber seleccionado también ficheros de tokenización y stopwords, para que todo se ejecute de la forma que buscamos.

Eduardo Socas Luis
Marcial Álvarez Parejo
Pablo Gónzalez Martín
Alberto Antonio Hernández Hernández

Matriz de Similaridad Coseno

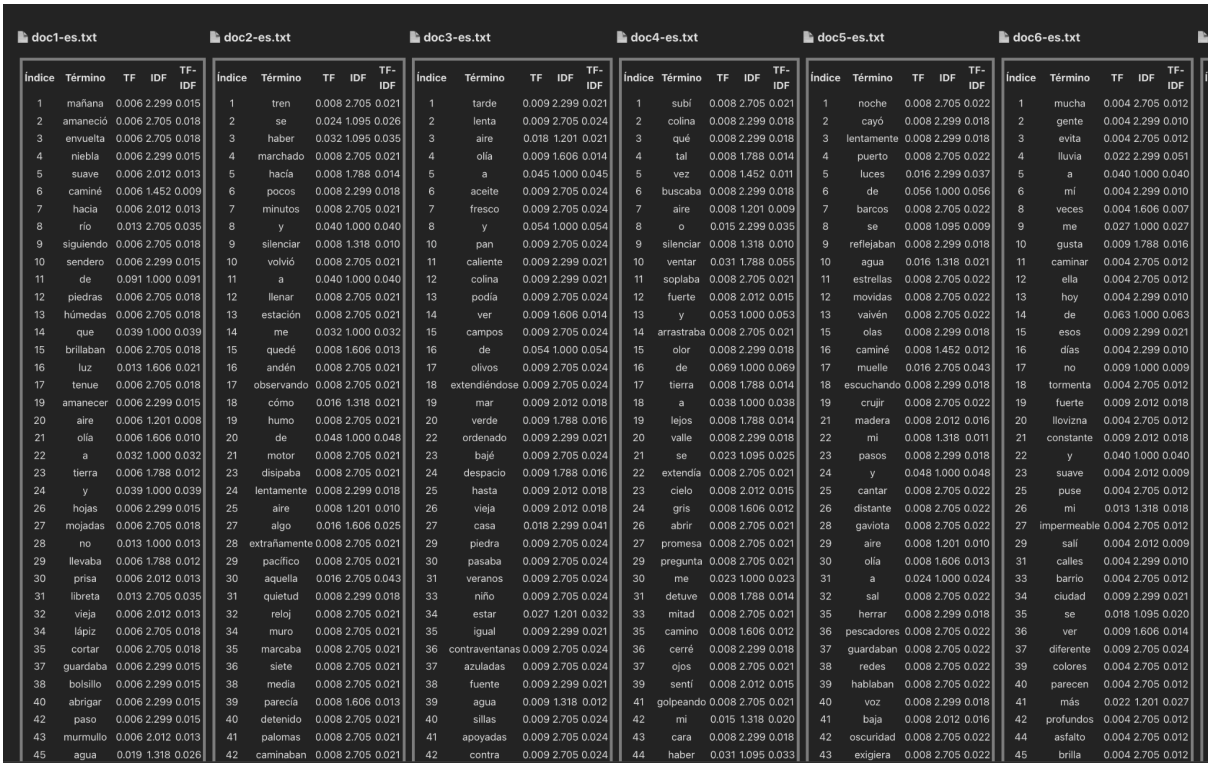
	doc1-es.txt	doc2-es.txt	doc3-es.txt	doc4-es.txt	doc5-es.txt	doc6-es.txt	doc7-es.txt	doc8-es.txt	doc9-es.txt	doc10-es.txt
doc1-es.txt	1.000	0.265	0.282	0.331	0.301	0.352	0.465	0.438	0.355	0.398
doc2-es.txt	0.265	1.000	0.193	0.285	0.243	0.234	0.296	0.300	0.254	0.289
doc3-es.txt	0.282	0.193	1.000	0.260	0.242	0.266	0.294	0.310	0.269	0.310
doc4-es.txt	0.331	0.285	0.260	1.000	0.273	0.330	0.426	0.414	0.383	0.372
doc5-es.txt	0.301	0.243	0.242	0.273	1.000	0.285	0.310	0.298	0.297	0.285
doc6-es.txt	0.352	0.234	0.266	0.330	0.285	1.000	0.374	0.406	0.362	0.395
doc7-es.txt	0.465	0.296	0.294	0.426	0.310	0.374	1.000	0.467	0.434	0.445
doc8-es.txt	0.438	0.300	0.310	0.414	0.298	0.406	0.467	1.000	0.387	0.456
doc9-es.txt	0.355	0.254	0.269	0.383	0.297	0.362	0.434	0.387	1.000	0.378
doc10-es.txt	0.398	0.289	0.310	0.372	0.285	0.395	0.445	0.456	0.378	1.000

Matriz de similaridad coseno (para nuevos documentos en español)

- La similaridad general, al contrario que en el caso de la matriz para los documentos en inglés, es bastante baja, ya que queríamos comprobar el comportamiento del modelo en otro contexto.

Eduardo Socas Luis
Marcial Álvarez Parejo
Pablo Gónzalez Martín
Alberto Antonio Hernández Hernández

- Al igual que en el modelo anterior, hace referencia a la similaridad entre documentos, y la diagonal principal vuelve a ser 1.000 porque es la similaridad entre los mismos documentos.



doc1-es.txt

Índice	Término	TF	IDF	TF-IDF
1	mañana	0.006	2.299	0.015
2	amaneció	0.006	2.705	0.018
3	envuelta	0.006	2.705	0.018
4	niebla	0.006	2.299	0.015
5	suave	0.006	2.012	0.013
6	caminé	0.006	1.452	0.009
7	hacia	0.006	2.012	0.013
8	río	0.013	2.705	0.035
9	siguiendo	0.006	2.705	0.018
10	sendero	0.006	2.299	0.015
11	de	0.091	1.000	0.091
12	piedras	0.006	2.705	0.018
13	húmedas	0.006	2.705	0.018
14	que	0.039	1.000	0.039
15	brillaban	0.006	2.705	0.018
16	luz	0.013	1.606	0.021
17	tenue	0.006	2.705	0.018
19	amanecer	0.006	2.299	0.015
20	aire	0.006	1.201	0.008
21	olía	0.006	1.606	0.010
22	a	0.032	1.000	0.032
23	tierra	0.006	1.788	0.012
24	y	0.039	1.000	0.039
26	hojas	0.006	2.299	0.015
27	mojadas	0.006	2.705	0.018
28	no	0.013	1.000	0.013
29	llevaba	0.006	1.788	0.012
30	prisa	0.006	2.012	0.013
31	libreta	0.013	2.705	0.035
32	vieja	0.006	2.012	0.013
34	lápiz	0.006	2.705	0.018
35	cortar	0.006	2.705	0.018
37	guardaba	0.006	2.299	0.015
38	bolsillo	0.006	2.299	0.015
40	abrigar	0.006	2.299	0.015
42	paso	0.006	2.299	0.015
43	murmullo	0.006	2.012	0.013
45	agua	0.019	1.318	0.026

doc2-es.txt

Índice	Término	TF	IDF	TF-IDF
1	tren	0.008	2.705	0.021
2	se	0.024	1.095	0.026
3	haber	0.032	1.095	0.035
4	marchado	0.008	2.705	0.021
5	hacía	0.008	1.788	0.014
6	pocos	0.008	2.299	0.018
7	minutos	0.008	2.705	0.021
8	y	0.040	1.000	0.040
9	silenciar	0.008	1.318	0.010
10	volvió	0.008	2.705	0.021
11	a	0.040	1.000	0.040
12	llenar	0.008	2.705	0.021
13	estación	0.008	2.705	0.021
14	me	0.032	1.000	0.032
15	quedé	0.008	1.606	0.013
16	andén	0.008	2.705	0.021
17	observando	0.008	2.705	0.021
18	cómo	0.016	1.318	0.021
19	humo	0.008	2.705	0.021
20	de	0.048	1.000	0.048
21	motor	0.008	2.705	0.021
23	disipaba	0.008	2.705	0.021
24	lentamente	0.008	2.299	0.018
25	aire	0.008	1.201	0.010
27	algo	0.016	1.606	0.025
28	extrañamente	0.008	2.705	0.021
29	pacífico	0.008	2.705	0.021
30	aquella	0.016	2.705	0.043
31	quietud	0.008	2.299	0.018
32	reloj	0.008	2.705	0.021
34	muro	0.008	2.705	0.021
35	marcaba	0.008	2.705	0.021
36	siete	0.008	2.705	0.021
38	media	0.008	2.705	0.021
39	parecía	0.008	1.606	0.013
40	detenido	0.008	2.705	0.021
41	palomas	0.008	2.705	0.021
42	caminaban	0.008	2.705	0.021

doc3-es.txt

Índice	Término	TF	IDF	TF-IDF
1	tarde	0.009	2.299	0.021
2	lenta	0.009	2.705	0.024
3	aire	0.018	1.201	0.021
4	olía	0.009	1.606	0.014
5	a	0.045	1.000	0.045
6	aceite	0.009	2.705	0.024
7	fresco	0.009	2.705	0.024
8	y	0.054	1.000	0.054
10	pan	0.009	2.705	0.024
11	caliente	0.009	2.299	0.021
12	colina	0.009	2.299	0.021
13	podía	0.009	2.705	0.024
14	ver	0.009	1.606	0.014
15	campos	0.009	2.705	0.024
16	de	0.054	1.000	0.054
17	olivos	0.009	2.705	0.024
18	extendiéndose	0.009	2.705	0.024
19	mar	0.009	2.012	0.018
20	verde	0.009	1.788	0.016
22	ordenado	0.009	2.299	0.021
23	bajé	0.009	2.705	0.024
24	despacio	0.009	1.788	0.016
25	hasta	0.009	2.012	0.018
26	vieja	0.009	2.012	0.018
27	casa	0.018	2.299	0.041
29	piedra	0.009	2.705	0.024
30	pasaba	0.009	2.705	0.024
31	veranos	0.009	2.705	0.024
33	niño	0.009	2.705	0.024
34	estar	0.027	1.201	0.032
35	igual	0.009	2.299	0.021
36	contraventanas	0.009	2.705	0.024
37	azuladas	0.009	2.705	0.024
38	fuente	0.009	2.299	0.021
39	agua	0.009	1.318	0.012
40	sillas	0.009	2.705	0.024
41	apoyadas	0.009	2.705	0.024
42	contra	0.009	2.705	0.024

doc4-es.txt

Índice	Término	TF	IDF	TF-IDF
1	subí	0.008	2.705	0.021
2	colina	0.008	2.299	0.018
3	qué	0.008	2.299	0.018
4	tal	0.008	1.788	0.014
5	vez	0.008	1.452	0.011
6	buscaba	0.008	2.299	0.018
7	aire	0.008	1.201	0.009
8	o	0.015	2.299	0.035
9	silenciar	0.008	1.318	0.010
10	ventar	0.031	1.788	0.055
11	soplaba	0.008	2.705	0.021
12	fuerte	0.008	2.012	0.015
13	y	0.053	1.000	0.053
14	arrastraba	0.008	2.705	0.021
15	olor	0.008	2.299	0.018
16	de	0.069	1.000	0.069
17	tierra	0.008	1.788	0.014
18	a	0.038	1.000	0.038
19	lejos	0.008	1.788	0.014
20	valle	0.008	2.299	0.018
21	se	0.023	1.095	0.025
22	extendía	0.008	2.705	0.021
23	cielo	0.008	2.012	0.015
24	gris	0.008	1.606	0.012
26	abrir	0.008	2.705	0.021
27	promesa	0.008	2.705	0.021
29	pregunta	0.008	2.705	0.021
30	me	0.023	1.000	0.023
31	detuve	0.008	1.788	0.014
33	mitad	0.008	2.705	0.021
35	camino	0.008	1.606	0.012
36	cerré	0.008	2.299	0.018
37	ojos	0.008	2.705	0.021
39	sentí	0.008	2.012	0.015
41	golpeando	0.008	2.705	0.021
42	mi	0.015	1.318	0.020
43	cara	0.008	2.299	0.018
44	haber	0.031	1.095	0.033

doc5-es.txt

Índice	Término	TF	IDF	TF-IDF
1	noche	0.008	2.705	0.022
2	cayó	0.008	2.299	0.018
3	lentamente	0.008	2.299	0.018
4	puerto	0.008	2.705	0.022
5	luces	0.016	2.299	0.037
6	de	0.056	1.000	0.056
7	barcos	0.008	2.705	0.022
8	se	0.008	1.095	0.009
9	reflejaban	0.008	2.299	0.018
10	agua	0.016	1.318	0.021
11	estrellas	0.008	2.705	0.022
12	movidas	0.008	2.705	0.022
13	vaivén	0.008	2.705	0.022
15	olas	0.008	2.299	0.018
16	caminé	0.008	1.452	0.012
17	muelle	0.016	2.705	0.043
18	escuchando	0.008	2.299	0.018
19	crujir	0.008	2.705	0.022
21	madera	0.008	2.012	0.016
22	mi	0.008	1.318	0.011
23	pasos	0.008	2.299	0.018
24	y	0.048	1.000	0.048
25	cantar	0.008	2.705	0.022
26	distante	0.008	2.705	0.022
28	gaviota	0.008	2.705	0.022
29	aire	0.008	1.201	0.010
30	olía	0.008	1.606	0.013
31	a	0.024	1.000	0.024
32	sal	0.008	2.705	0.022
35	herrar	0.008	2.299	0.018
36	pescadores	0.008	2.705	0.022
37	guardaban	0.008	2.705	0.022
38	redes	0.008	2.705	0.022
39	hablaban	0.008	2.705	0.022
40	voz	0.008	2.299	0.018
41	baja	0.008	2.012	0.016
42	oscuridad	0.008	2.705	0.022
43	exigiera	0.008	2.705	0.022

doc6-es.txt

Índice	Término	TF	IDF	TF-IDF
1	mucha	0.004	2.705	0.012
2	gente	0.004	2.299	0.010
3	evita	0.004	2.705	0.012
4	lluvia	0.022	2.299	0.051
5	a	0.040	1.000	0.040
6	mí	0.004	2.299	0.010
8	veces	0.004	1.606	0.007
9	me	0.027	1.000	0.027
10	gusta	0.009	1.788	0.016
11	caminar	0.004	2.705	0.012
12	ella	0.004	2.705	0.012
13	hoy	0.004	2.299	0.010
14	de	0.063	1.000	0.063
15	esos	0.009	2.299	0.021
16	días	0.004	2.299	0.010
17	no	0.009	1.000	0.009
18	tormenta	0.004	2.705	0.012
19	fuerte	0.009	2.012	0.018
20	llovizna	0.004	2.705	0.012
21	constante	0.009	2.012	0.018
22	y	0.040	1.000	0.040
23	suave	0.004	2.012	0.009
25	puse	0.004	2.705	0.012
26	mi	0.013	1.318	0.018
27	impermeable	0.004	2.705	0.012
29	salí	0.004	2.012	0.009
31	calles	0.004	2.299	0.010
33	barrio	0.004	2.705	0.012
34	ciudad	0.009	2.299	0.021
35	se	0.018	1.095	0.020
36	ver	0.009	1.606	0.014
37	diferente	0.009	2.705	0.024
39	colores	0.004	2.705	0.012
40	parecen	0.004	2.705	0.012
41	más	0.022	1.201	0.027
42	profundos	0.004	2.705	0.012
44	asfalto	0.004	2.705	0.012
45	brilla	0.004	2.705	0.012

Una vez ejecutamos, al igual que para los documentos en inglés, obtenemos una columna por cada documento que seleccionamos, y cada columna contiene:

- Índice (1º ocurrencia del término en el documento)
- Término
- TF
- IDF
- TF-IDF