

Ejercicio 1 - Modelos de Lenguaje

matriz termino-documento				
Term \ Doc	D1	D2	D3	D4
ademas	1	0	0	0
asi	0	0	1	0
como	0	0	1	0
comunicacion	1	0	1	0
crecimiento	1	0	0	1
cultura	0	1	0	1
de	2	0	3	0
del	0	0	0	1
desarrolladores	1	0	0	0
el	2	0	0	1
en	2	0	0	2
entre	1	0	0	0
es	0	1	2	1
eso	0	1	0	0
esta	0	0	0	1
estado	0	0	0	1
favorecido	1	0	0	0
fundamental	1	0	1	1
ha	2	0	0	0
hace	0	1	0	0
hardware	0	0	1	0
incorpore	0	0	0	1
internet	2	0	0	0
la	1	2	2	1
libre	1	1	0	2
lo	0	1	1	0
los	1	0	0	0
mas	0	0	1	0
mayor	0	1	0	0
nuestro	0	0	1	1
pais	0	1	1	1
papel	1	0	0	0
para	0	0	1	0
produccion	0	0	2	0
que	0	1	0	1
riqueza	0	1	0	0
software	2	0	1	2
tecnologia	0	0	1	0
tenido	1	0	0	0
tiene	0	1	0	0
un	1	1	0	0
y	0	0	1	0

Probabilidad P(tf / N)					
Term \ Doc	D1	D2	D3	D4	C
ademas	0.04	0.00	0.00	0.00	0.01
asi	0.00	0.00	0.05	0.00	0.01
como	0.00	0.00	0.05	0.00	0.01
comunicacion	0.04	0.00	0.05	0.00	0.03
crecimiento	0.04	0.00	0.00	0.05	0.03
cultura	0.00	0.07	0.00	0.05	0.03
de	0.08	0.00	0.14	0.00	0.06
del	0.00	0.00	0.00	0.05	0.01
desarrolladores	0.04	0.00	0.00	0.00	0.01
el	0.08	0.00	0.00	0.05	0.04
en	0.08	0.00	0.00	0.11	0.05
entre	0.04	0.00	0.00	0.00	0.01
es	0.00	0.07	0.09	0.05	0.05
eso	0.00	0.07	0.00	0.00	0.01
esta	0.00	0.00	0.00	0.05	0.01
estado	0.00	0.00	0.00	0.05	0.01
favorecido	0.04	0.00	0.00	0.00	0.01
fundamental	0.04	0.00	0.05	0.05	0.04
ha	0.08	0.00	0.00	0.00	0.03
hace	0.00	0.07	0.00	0.00	0.01
hardware	0.00	0.00	0.05	0.00	0.01
incorpore	0.00	0.00	0.00	0.05	0.01
internet	0.08	0.00	0.00	0.00	0.03
la	0.04	0.14	0.09	0.05	0.08
libre	0.04	0.07	0.00	0.11	0.05
lo	0.00	0.07	0.05	0.00	0.03
los	0.04	0.00	0.00	0.00	0.01
mas	0.00	0.00	0.05	0.00	0.01
mayor	0.00	0.07	0.00	0.00	0.01
nuestro	0.00	0.00	0.05	0.05	0.03
pais	0.00	0.07	0.05	0.05	0.04
papel	0.04	0.00	0.00	0.00	0.01
para	0.00	0.00	0.05	0.00	0.01
produccion	0.00	0.00	0.09	0.00	0.03
que	0.00	0.07	0.00	0.05	0.03
riqueza	0.00	0.07	0.00	0.00	0.01
software	0.08	0.00	0.05	0.11	0.06
tecnologia	0.00	0.00	0.05	0.00	0.01
tenido	0.04	0.00	0.00	0.00	0.01
tiene	0.00	0.00	0.00	0.00	0.01
un	0.04	0.07	0.00	0.00	0.03
y	0.00	0.00	0.05	0.00	0.01

Tamaño Docs	
Doc	Terminos
1	25
2	14
3	22
4	19

Query-likelihood + Sin Smoothing

Query 1 - (“pais cultura”)			
Tabla Documentos - Query 1			Ranking de documentos
Doc \ Term	pais	cultura	Doc Rank
1	0.00	0.00	1 0.0000
2	0.07	0.07	2 0.0051
3	0.05	0.00	3 0.0000
4	0.05	0.05	4 0.0028

Query 2 - (“pais libre cultura”)				
Tabla Documentos - Query 1-1				Ranking de documentos-1
Doc \ Term	pais	libre	cultura	Doc Rank
1	0.00	0.04	0.00	1 0.0000
2	0.07	0.07	0.07	2 0.0004
3	0.05	0.00	0.00	3 0.0000
4	0.05	0.11	0.05	4 0.0003

Query 3 - (“software propietario licencia”)				
Tabla Documentos - Query 1-1-1				Ranking de documentos-1-1
Doc \ Term	software	propietario	licencia	Doc Rank
1	0.08	0	0	1 0.0000
2	0.00	0	0	2 0.0000
3	0.05	0	0	3 0.0000
4	0.11	0	0	4 0.0000

¿Qué problemas encuentra?

El principal problema es que cuando un termino no aparece en el documento este hace que la probabilidad sea cero y por lo tanto termina haciendo cero a todo el producto. Por lo que no importa cuan buenos eran los términos restantes. Para ello se utilizan técnicas de smoothing que evitan que la probabilidad se cancele.

Query-likelihood + Jelinek-Mercer

λ 0.1000

Query 1 - (“pais cultura”)			
Tabla Documentos - Query 1-2			Ranking de documentos-2
Doc \ Term	pais	cultura	Doc Rank
1	0.0038	0.0025	1 0.00001
2	0.0680	0.0688	2 0.00454
3	0.0447	0.0025	3 0.00011
4	0.0511	0.0499	4 0.00255

Query 2 - (“pais libre cultura”)				
Tabla Documentos - Query 1-1-2				Ranking de documentos-1-2
Doc \ Term	pais	libre	cultura	Doc Rank
1	0.0038	0.0410	0.0568	1 0.00001
2	0.0680	0.0693	0.0025	2 0.00001
3	0.0447	0.0050	0.0499	3 0.00001
4	0.0511	0.0997	0.0250	4 0.00013

Query 3 - (“software propietario licencia”)				
Tabla Documentos - Query 1-1-1-1				Ranking de documentos-1-1-1
Doc \ Term	software	propietario	licencia	Doc Rank
1	0.0783	0	0	1 0.0000
2	0.0063	0	0	2 0.0000
3	0.0472	0	0	3 0.0000
4	0.1010	0	0	4 0.0000

Comparando ambos modelos

Podemos observar que además de no cancelar los documentos que tienen términos con frecuencia 0 sobre alguno de los términos de la query observamos que aumenta el ranking de los documentos con valores más chicos.

Aun así podemos observar que si se consulta por un termino que no esta en el corpus, esto resulta en cancelar todos los documentos candidatos. Lo que da como resultado ningún documento.

Observaciones

Podemos observar que la query #3 devolvió documentos. Por lo que ya no es un problema realizar una consulta por términos que no tengamos dentro del corpus. Esto se debe principalmente a que es una sumatoria y no una productoria la que se realiza en el ultimo paso.