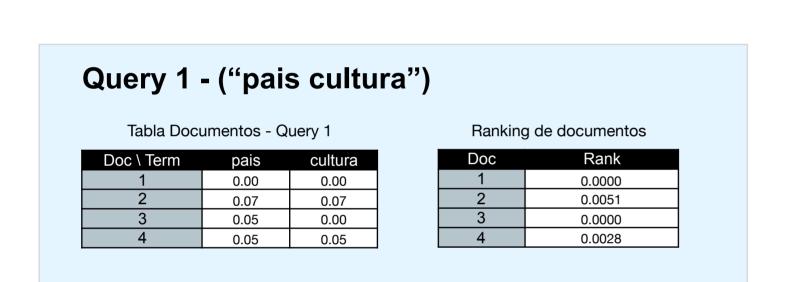
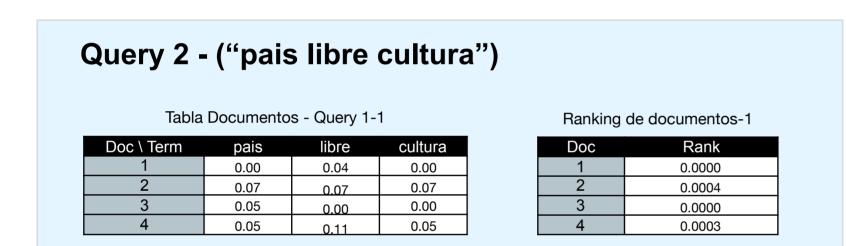
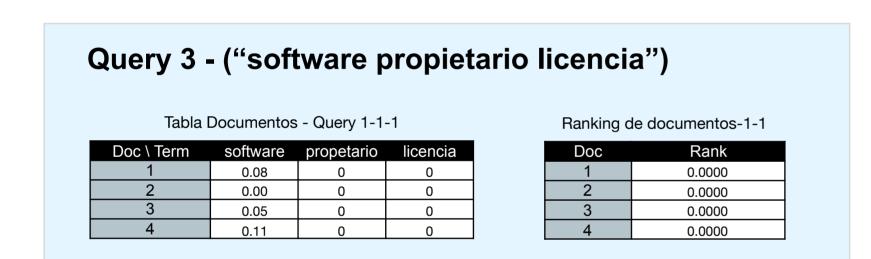
Ejercicio 1 - Modelos de Lenguaje

	matri	z termino-docum	ento ———		Probabilidad P(tf / N)							
Term \ Doc	D1	D2	D3	D4	Term \ Doc	D1	D2	D3	D4	С		
ademas	1	0	0	0	ademas	0.04	0.00	0.00	0.00	0.01		
asi	0	0	1	0	asi	0.00	0.00	0.05	0.00	0.01		
como	0	0	1	0	como	0.00	0.00	0.05	0.00	0.01		
municacion	1	0	1	0	comunicacion	0.04	0.00	0.05	0.00	0.03		
recimiento	1	0	0	1	crecimiento	0.04	0.00	0.00	0.05	0.03		
cultura	0	1	0	1	cultura	0.00	0.07	0.00	0.05	0.03		
de	2	0	3	0	de	0.08	0.00	0.14	0.00	0.06		
del	0	0	0	1	del	0.00	0.00	0.00	0.05	0.01		
sarrolladores	1	0	0	0	desarrolladores	0.04	0.00	0.00	0.00	0.01		
el	2	0	0	1	el	0.08	0.00	0.00	0.05	0.04		
en	2	0	0	2	en	0.08	0.00	0.00	0.11	0.05		
entre	1	0	0	0	entre	0.04	0.00	0.00	0.00	0.01		
es	0	1	2	1	es	0.00	0.07	0.09	0.05	0.05		
eso	0	1	0	0	eso	0.00	0.07	0.00	0.00	0.01		
esta	0	0	0	1	esta	0.00	0.00	0.00	0.05	0.01		
estado	0	0	0	1	estado	0.00	0.00	0.00	0.05	0.01		
avorecido	1	0	0	0	favorecido	0.04	0.00	0.00	0.00	0.01		
undamental	1	0	1	1	fundamental	0.04	0.00	0.05	0.05	0.04		
ha	2	0	0	0	ha	0.08	0.00	0.00	0.00	0.03		
hace	0	1	0	0	hace	0.00	0.07	0.00	0.00	0.01		
hardware	0	0	1	0	hardware	0.00	0.00	0.05	0.00	0.01		
incorpore	0	0	0	1	incorpore	0.00	0.00	0.00	0.05	0.01		
internet	2	0	0	0	internet	0.08	0.00	0.00	0.00	0.03		
la	1	2	2	1	la	0.04	0.14	0.09	0.05	0.08		
libre	1	1	0	2	libre	0.04	0.07	0.00	0.11	0.05		
lo	0	1	1	0	lo	0.00	0.07	0.05	0.00	0.03		
los	1	0	0	0	los	0.04	0.00	0.00	0.00	0.01		
mas	0	0	1	0	mas	0.00	0.00	0.05	0.00	0.01		
mayor	0	1	0	0	mayor	0.00	0.07	0.00	0.00	0.01		
nuestro	0	0	1	1	nuestro	0.00	0.00	0.05	0.05	0.03		
pais	0	1	1	1	pais	0.00	0.07	0.05	0.05	0.04		
papel	1	0	0	0	papel	0.04	0.00	0.00	0.00	0.01		
para	0	0	1	0	para	0.00	0.00	0.05	0.00	0.01		
produccion	0	0	2	0	produccion	0.00	0.00	0.09	0.00	0.03		
que	0	1	0	1	que	0.00	0.07	0.00	0.05	0.03		
riqueza	0	1	0	0	riqueza	0.00	0.07	0.00	0.00	0.01		
software	2	0	1	2	software	0.08	0.00	0.05	0.11	0.06		
ecnologia	0	0	1	0	tecnologia	0.00	0.00	0.05	0.00	0.01		
tenido	1	0	0	0	tenido	0.04	0.00	0.00	0.00	0.01		
tiene	0	1	0	0	tiene	0.00	0.07	0.00	0.00	0.01		
un	1	1	0	0	un	0.04	0.07	0.00	0.00	0.03		
	0	0	1	0	y	0.00	0.00	0.05	0.00	0.01		

Query-likelihood + Sin Smoothing







¿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 Docu	mentos - Qເ	iery 1-2		Ranking	de documentos-2					
Doc \ Term	pais	cultura		Doc	Rank					
1	0.0038	0.0025		1	0.00001					
2	0.0680	0.0668		2	0.00454					
3	0.0447	0.0025		3	0.00011					
1	0.0511	0.0499		4	0.00255					

Query 2	· ("pais	ibre (cultura	ı")		
Tabla	Documentos	- Query 1-1		Ranking	de documentos-1-2	
Doc \ Term	pais	libre	cultura		Doc	Rank
1	0.0038	0.0410	0.0668		1	0.00001
					2	0.00001
2	0.0680	0.0693	0.0025		_	0.00001
2 3	0.0680 0.0447	0.0693 0.0050	0.0025 0.0499		3	0.00001

Query 3	- ("soft	ware p	ropiet	ario I	icencia	a")
Tabla D	ocumentos	- Query 1-1-1	1-1		Ranking de	e documentos-1-1-1
Doo \ Torm	software	propetario	licencia		Doc	Rank
Doc \ Term	Software	p. op otoo				
Doc Vierni 1	0.0783	0	0		1	0.0000
1 2		0	0		1 2	
1 2 3	0.0783	0 0 0	0 0 0		1 2 3	0.0000

Ejercicio 2 - Modelos de Lenguaje

Kullback Leible + Dirichlet

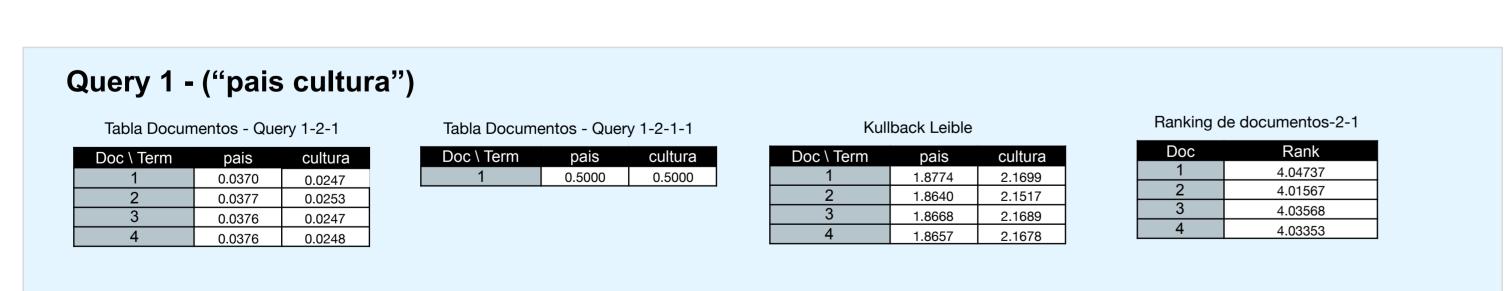


Tabla D	Tabla Documentos - Query 1-1-2-1			Tabla Do	ocumentos -	Query 1-1-2	?-1-1	Tabla Do	ocumentos -	Ranking de	e documentos-2-1-		
		libro	cultura	Doc \ Term	pais	libre	cultura	Doc \ Term	pais	libre	cultura	Doc	Rank
Doc \ Term	pais	libre	Guitura	200 (101111									
Doc \ Term	0.0370	0.0499	0.0247	1	0.3333	0.3333	0.3333	1	1.0566	0.9135	1.2516	1	3.22178
Doc \ Term 1 2		0.0499	0.0247	1		0.3333	0.3333	1 2	1.0566 1.0477	0.9135 0.9109	1.2516 1.2395	1 2	3.22178 3.19803
Doc \ Term 1 2 3	0.0370			1		0.3333	0.3333	1 2 3				1 2 3	

uery 3 -	("soft	ware p	ropiet	ario licenci	a")									
Tabla Documentos - Query 1-1-1-1			Tabla Doo	Tabla Documentos - Query 1-1-1-1-2				Tabla Documentos - Query 1-1-1-1-1					documentos-2-1-1-	
Doc \ Term	software	propetario	licencia	Doc \ Term	software	propetario	licencia	Doc \ Term	software	propetario	licencia		Doc	Rank
1	0.0627	0	0	1	0.3333	0.3333	0.3333	1	0.8034	0	0		1	0.80335
2	0.0621	0	0		•	•		2	0.8084	0	0		2	0.80837
3	0.0623	0	0					3	0.8064	0	0		3	0.80644
1	0.0629	0	0					4	0.8019	0	0		4	0.80193

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.