Recuperación de Información

Modelos de Recuperación de Información (Parte 2)

Normand Agustín

1. Retome el TP de ”Modelos de RI” y calcule el modelo de lenguaje (unigramas) para los documentos del
ejercicio 2. Utilizando el modelo de Query Likelihood calcule los rankings para las siguiente consultas:
a) país cultura
b) país libre cultura
c) software propietario licencia

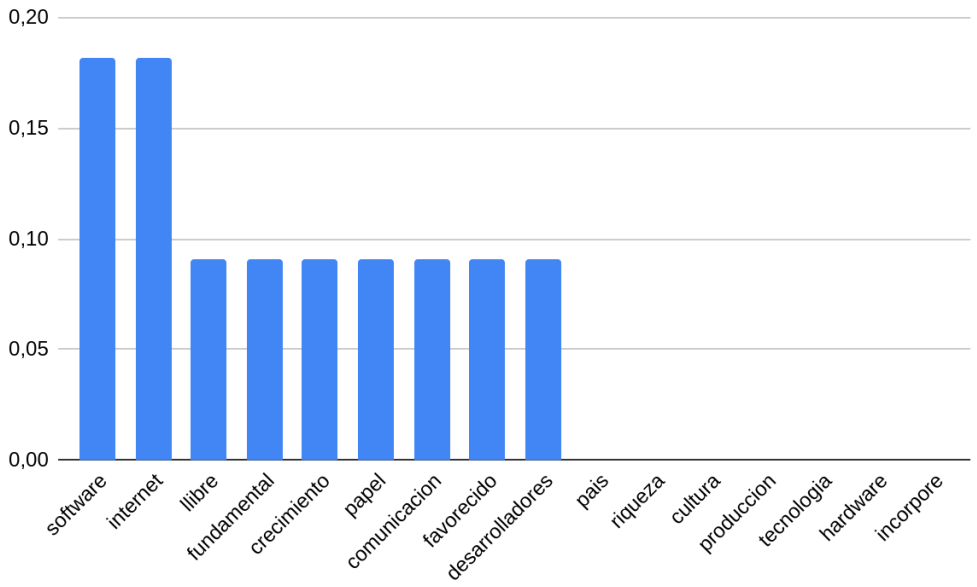
¿Qué problemas encuentra? Luego, calcule las probabilidades de los términos utilizando una combinación con el ML de la colección (suavizado Jelinek-Mercer, λ = 0,7). Compare con las probabilidades anteriores y explique las diferencias. Repita las consultas con los nuevos valores. Explique los resultados.

Utilizando el modelo de Query Likelihood los rankings para las consultas son los siguientes:

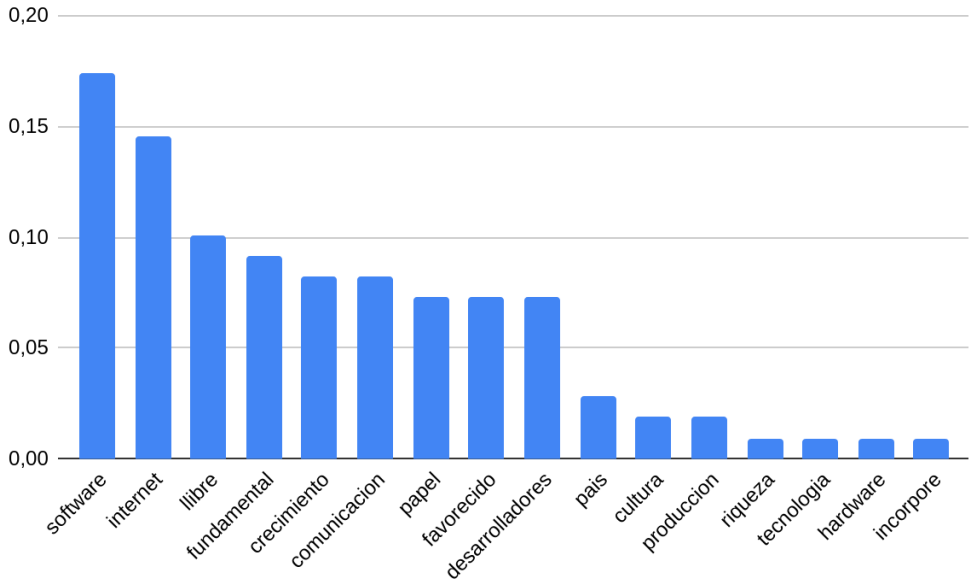
	D1	D2	D3	D4
Q1	0	0,0625	0	0,01234567901
Q2	0	0,015625	0	0,002743484225
Q3	0,1818181818	0	0,125	0,2222222222

El problema que encuentro es que hay documentos que tienen score 0. Es decir, documentos que no tienen ningún término de la query, son igual de parecidos, que documentos que tienen algunos, pero no todos, los términos de la consulta.
Esto se debe a que al ser una productoria, con que uno de los términos de la query, en un documento tenga frecuencia relativa 0, el score del mismo va a ser 0.

Luego de combinar los modelos de lenguaje de los documentos, con el de la colección, utilizando un λ = 0,7, vemos que por ejemplo, el modelo de lenguaje para el documento 1, pasó de ser:



A ser el siguiente:



Vemos que sucedieron por lo menos 2 cosas:

- Ya no hay términos con frecuencia relativa de 0, ahora estos tienen una probabilidad residual, muy chica.
- Se disminuyeron las probabilidades de los términos, esto es debido a que el suavizado trata la sobreestimación en documentos cortos.

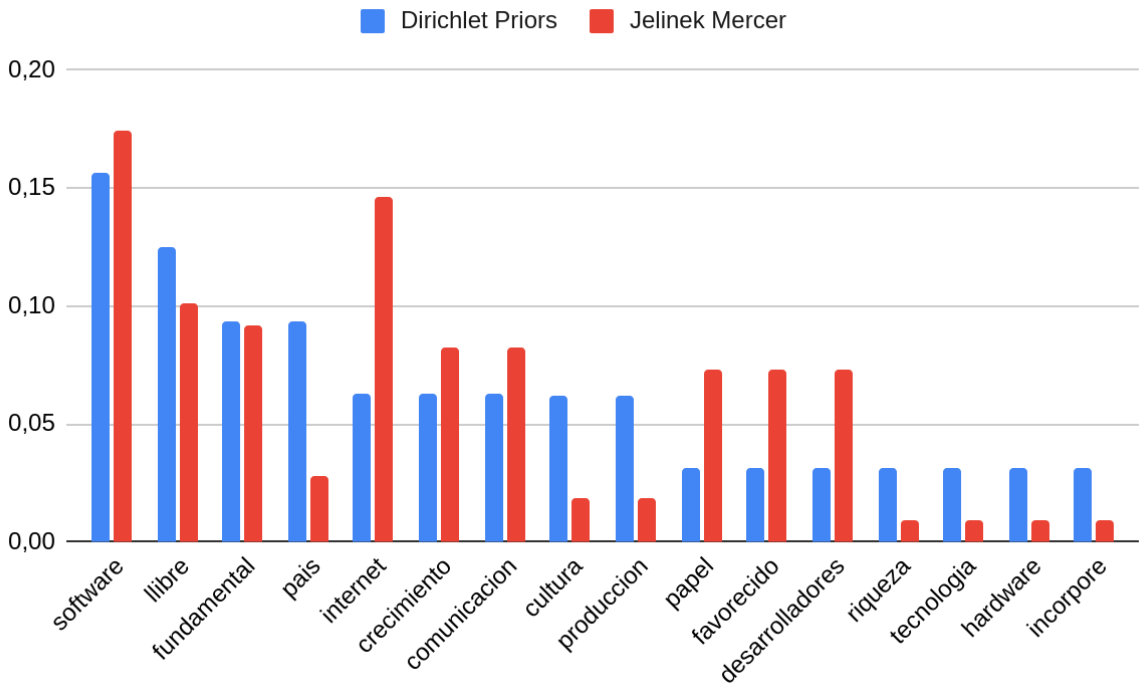
Utilizando Query Likelihood para resolver las consultas, obtengo los rankings:

	D1	D2	D3	D4
Q1	0,00052734375	0,03935546875	0,00216796875	0,0102225598
Q2	0,00005333362926	0,008363037109	0,00008129882813	0,001973521961
Q3	0,1741477273	0,046875	0,134375	0,2024305556

Vemos que ahora no tenemos documentos con score 0.
Utilizando Query Likelihood pero sin suavizar, para la consulta {país, cultura} no era posible determinar qué documento tenía mayor relevancia entre el 1 y el 3, dado que ambos eran 0. Sin embargo, ahora sí es posible, el documento 3 tiene mayor score que el 1. Y esto tiene sentido dado que el documento 1 no posee ninguno de los términos de la consulta, y el documento 3, posee "país" 1 vez, o con una frecuencia relativa de 0,125.

2. Repita el ejercicio pero esta vez utilice la divergencia de Kullback-Leibler y un suavizado por Dirichlet-Priors utilizando para los parámetros los valores sugeridos en la literatura.

En primer lugar, calculé los modelos de lenguaje para los documentos, utilizando el suavizado Dirichlet Priors. Observé que difieren bastante los modelos comparando entre un suavizado y otro, por ejemplo, para Internet, o País.



Apliqué un suavizado en las consultas, utilizando la longitud de la query en el denominador de la fórmula de Dirichlet Priors.

Luego de calcular los rankings para cada uno de los documentos, obtuve la siguiente tabla:

	D1	D2	D3	D4
Q1	0,0007413200759	0,00006812240476	0,0004219592193	0,0002381314205
Q2	0,0009238276035	0,0000611993252	0,0007406336879	0,0001503473596
Q3	0,000122498025	0,0003193206728	0,0002376113537	0,00005479489578

Dichos valores se tratan de la divergencia, es decir, cuán diferentes son los modelos de lenguaje que estoy evaluando, y el 0 significa que son muy similares o idénticos, por este motivo, cuando ordene los rankings, lo hago de menor a mayor, y no como en el punto anterior.

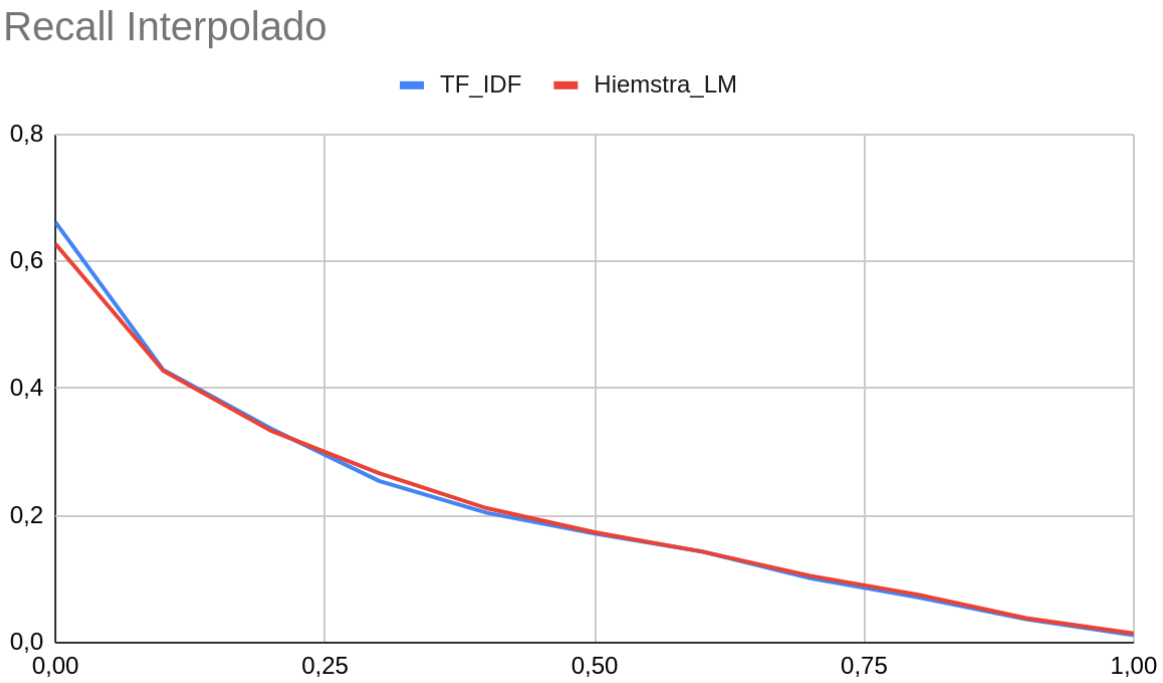
	Q1			Q1
D2	0,03935546875		D2	0,00006812240476
D4	0,0102225598		D4	0,0002381314205
D3	0,00216796875		D3	0,0004219592193
D1	0,00052734375		D1	0,0007413200759
	Q2			Q2
D2	0,008363037109		D2	0,0000611993252

D4	0,001973521961		D4	0,0001503473596
D3	0,00008129882813		D3	0,0007406336879
D1	0,00005333362926		D1	0,0009238276035
	Q3			Q3
D4	0,2024305556		D4	0,00005479489578
D1	0,1741477273		D1	0,000122498025
D3	0,134375		D3	0,0002376113537
D2	0,046875		D2	0,0003193206728

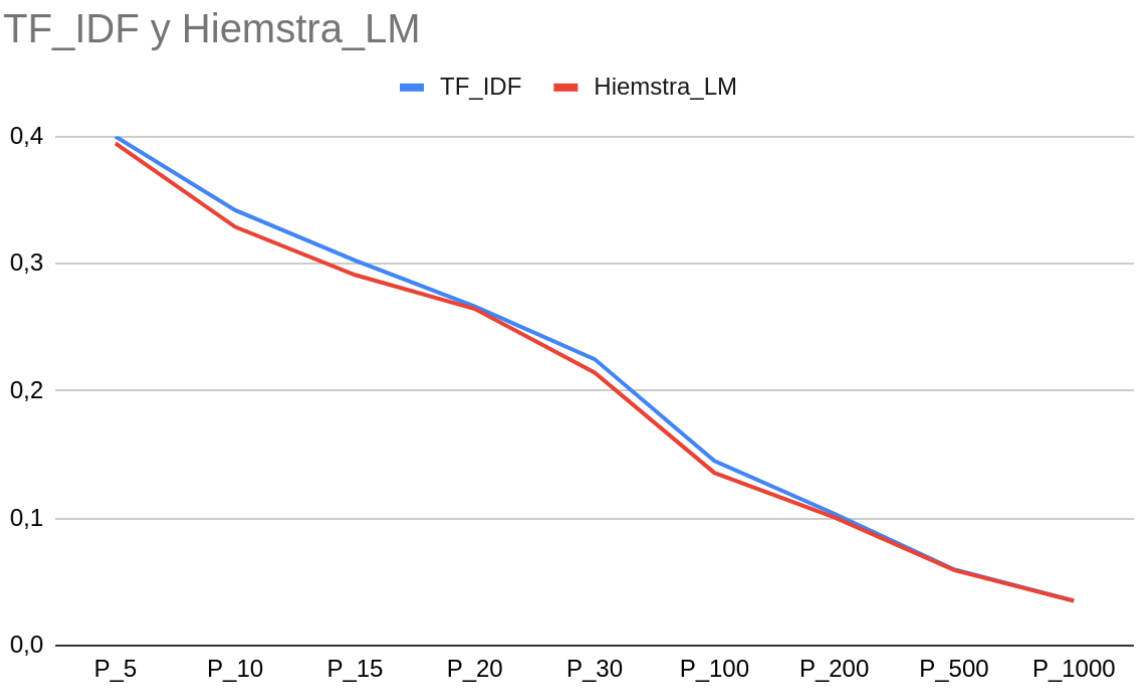
Comparando los rankings de Query Likelihood suavizado con Jelinek Mercer, contra Kulback Leiber suavizado con Dirichlet Priors, vemos obviamente los scores no son los mismos, pero el orden de los documentos del ranking es idéntico, ambos modelos son consistentes en cuanto a la relevancia estimada que le otorgan a los documentos, para esta colección, estas consultas y estos valores de μ y λ.

3. Utilizando modelos de lenguaje en Terrier (use Hiemstra LM), repita los experimentos del ejercicio 9 del TP de ”modelos” y compare los resultados con los anteriores. ¿Son consistentes? Calcule las métricas apropiadas para comparar los diferentes sistemas y configuraciones.

Realicé la comparación utilizando el modelo TF_IDF. Además, utilicé el archivo de querys que agregaba los términos tantas veces como aparecían en la necesidad de información, dado que, en el trabajo práctico anterior, vi que tenía mejor desempeño que solamente agregar los términos una vez.



Vemos que la gráfica para cada uno de los modelos es casi idéntica en la mayoría de sus recorridos.
Podemos observar que TF_IDF tiene levemente más precisión en el primer intervalo de recall, mientras que Hiemstra LM tiene mayor precisión en el intervalo 0,25;0,40, equilibrando de esta forma el desempeño de ambos modelos.



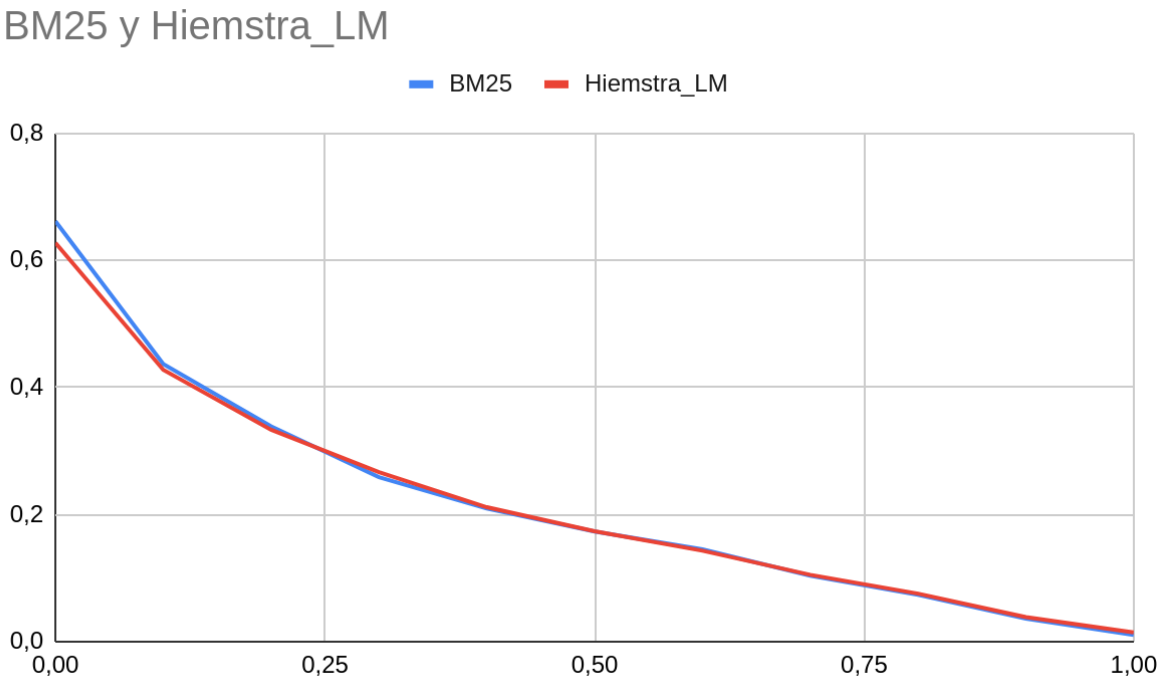
Si vemos como varía la precisión conforme avanzamos en el ranking, vemos que ambos modelos son muy similares.

Por otro lado, calculé el coeficiente de spearman para los conjuntos de respuesta de las 112 consultas, aumentando los rankings en los casos que sea necesario, de la forma que indica Baeza.

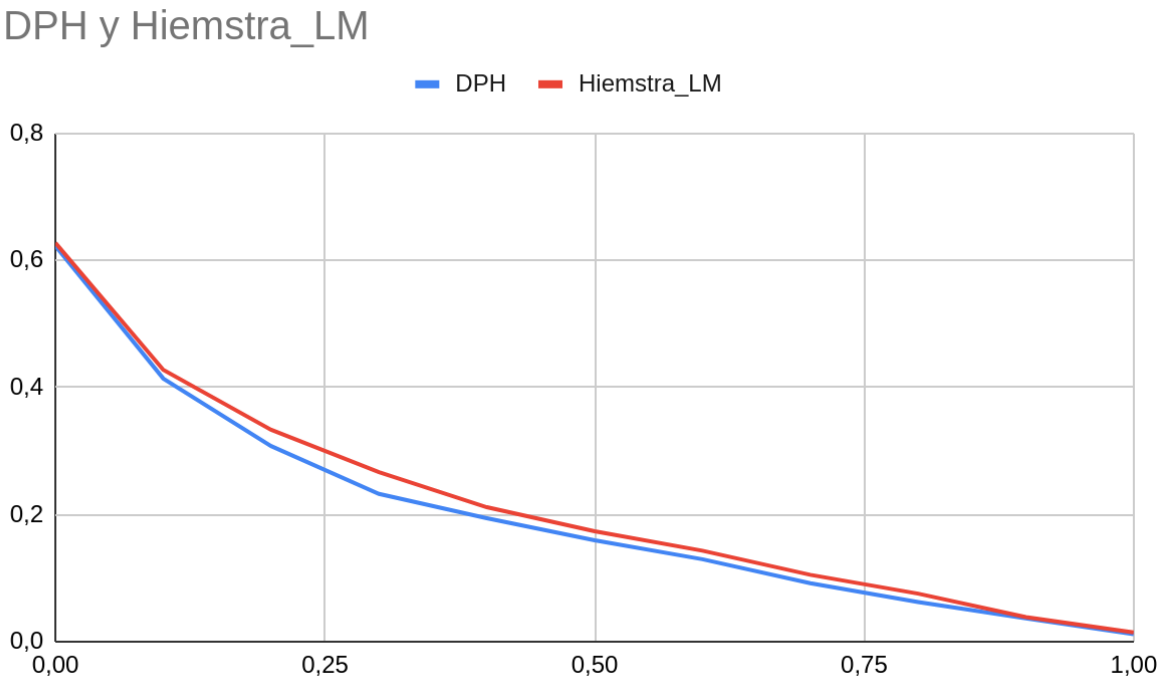
Teniendo en cuenta que el largo promedio de los rankings es de 915 documentos, el coeficiente de Spearman promedio a lo largo de las 112 querys es de 0.951551361.

Por lo que se puede afirmar que ambos modelos son consistentes.

Por otro lado para Hiemstra y BM25:



Anteriormente TF_IDF tenía levemente más precisión en el primer intervalo de recall, también logra mayor precisión en este intervalo BM25. Pero, a diferencia que en el caso anterior, en el intervalo 0,25;0,40 Hiemstra tenía más precisión que TF_IDF, compensando la baja de precisión del primer intervalo, pero esto no ocurre significativamente comparando Hiemstra con BM25



Y para la comparación con DPH, en todo momento se puede observar una mayor precisión por parte de Hiemstra a lo largo de los intervalos interpolados de recall.