Modelos de Recuperación de Información (Parte 2)

Alumno: Gallozo, Luis Angel 156308

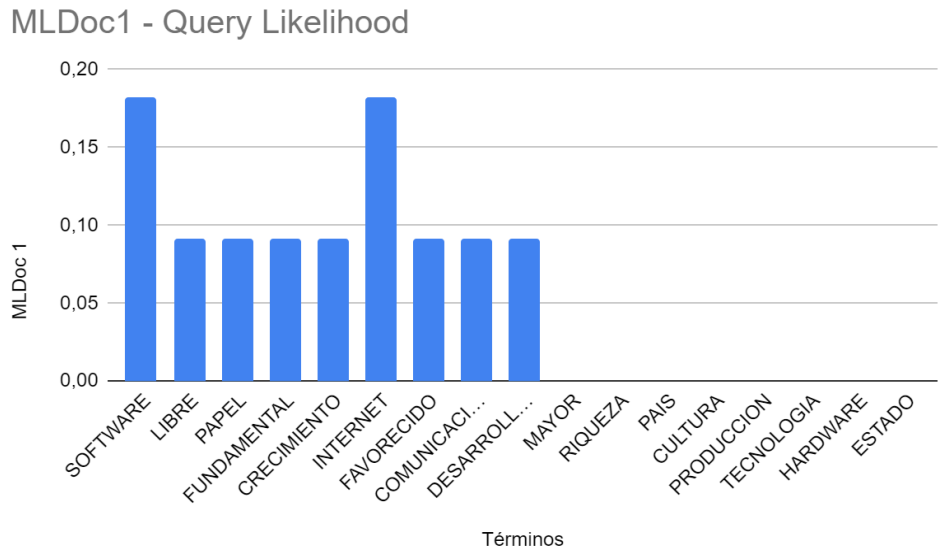
1. Retome el TP de ”Modelos de RI” y calcule el modelo de lenguaje (unigramas) para los documentos del ejercicio 2. Utilizando el modelo de Query Likelihood calcule los rankings para las siguiente consultas:

a) país cultura; b) país libre cultura; c) software propietario licencia
¿Qué problemas encuentra?

Utilizando el modelo de Query Likelihood los rankings son los siguientes:

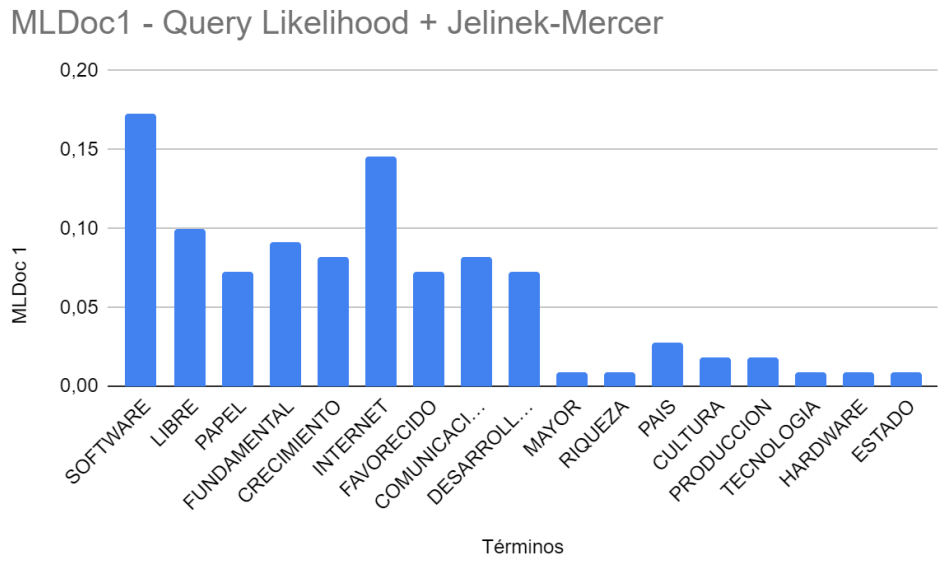
	Docs			
Querys	Doc1	Doc2	Doc3	Doc4
Q1(país cultura)	0	0,04	0	0,01234567901
Q2(país libre cultura)	0	0,008	0	0,002743484225
Q3(software propietario licencia)	0,1818181818	0	0,125	0,2222222222

Se observa el problema de frecuencia 0, donde tenemos que existen términos de la consulta que no existen en el documento, afectando al ranking. Esto ocurre debido a la productoria de probabilidades que se utiliza para calcular el score, si algún término tiene probabilidad 0 da como resultado un score de 0 sin tener en cuenta los demás términos de la query.
Si graficamos el ML del Doc 1, se aprecia mejor esa frecuencia 0 de los términos:



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.

Luego se realiza una combinación entre Query Likelihood y el método de suavizado Jelinek-Mercer con λ = 0,7, donde ocurre el siguiente cambio en el ML del Doc 1:



Con este método se soluciona la frecuencia 0 de los términos que no existen en documento, ya que se les asigna una porción de la probabilidad del término en la colección. Cabe mencionar que todas las probabilidades recibieron una disminución esto causado porque el método trata la sobreestimación en documentos cortos.

Para finalizar, obtenemos los siguientes resultados utilizando Query Likelihood + JM en los rankings:

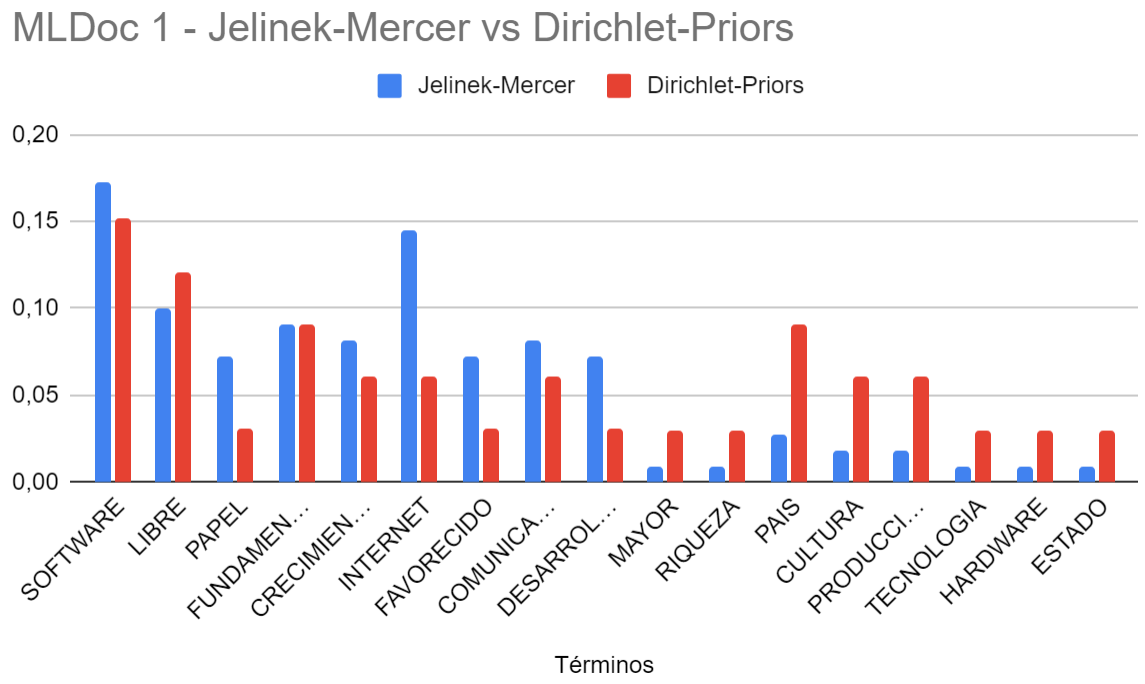
	Docs			
Querys	Doc1	Doc2	Doc3	Doc4
Q1(país cultura)	0,0004958677686	0,02645950413	0,00208677686	0,01008060402
Q2(país libre cultura)	0,00004958677686	0,004666494365	0,00007588279489	0,001934661378
Q3(software propietario licencia)	0,1727272727	0,04545454545	0,1329545455	0,201010101

La frecuencia 0 dejó de afectar el score de los documentos. Tomando como ejemplo la Query 1, con solo utilizar QL no teníamos forma de distinguir la relevancia entre el Doc 1 y Doc 3, ya que tenían score 0. Con estos nuevos resultados es más simple determinar el ranking para ambos documentos, tiene sentido que el Doc 3 sea mejor rankeado,

debido a que para la query 1 el cuenta con el término **país**, mientras que el Doc 1, no tienen ninguno de los términos.

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

Se aplica el método Dirichlet-Priors con un μ=2000. A continuación se realiza una comparación con JM, obteniendo lo siguiente para el ML del Doc 1:



Se observa una gran diferencia entre los métodos de suavizado, casos a remarcar son por ejemplo los términos: internet, país o producción.

Al calcular los rankings utilizando la divergencia de Kullbak-Leibler, obtenemos lo siguiente:

	Docs			
Querys	Doc1	Doc2	Doc3	Doc4
Q1(país cultura)	0,0007321508002	0,00009908210245	0,0004158387606	0,0002309612062
Q2(país libre cultura)	0,0009091407868	0,0001186796663	0,0007314599197	0,000139306395
Q3(software propietario licencia)	0,0001122630216	0,00034917152	0,0002304339773	0,0000466049582

Los scores en este caso hace referencia a la divergencia, osea a que tan diferentes son los modelos de lenguaje de la documentos sobre el modelo de lenguaje de la query. Si el valor es 0 o cercano, quiere decir que son idénticos o similares.

Comparación de Rankings de QL+JM vs KL+DP:

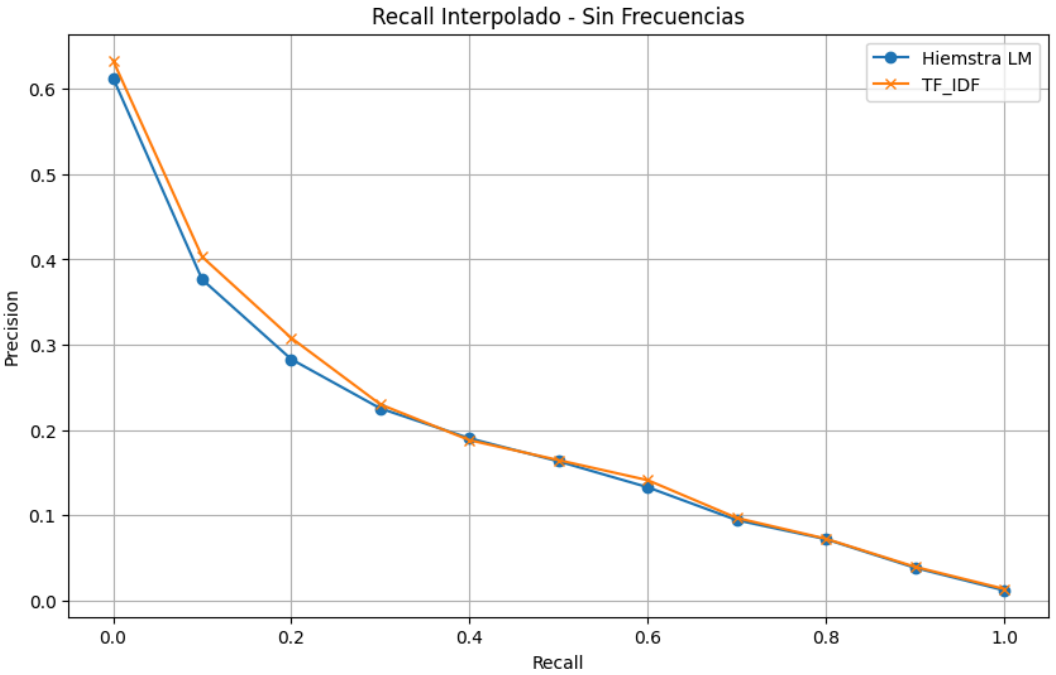
	QL+JM			KL+DP
	Q1			Q1
Doc 2	0,02645950413		Doc 2	0,00009908210245
Doc 4	0,01008060402		Doc 3	0,0001186796663
Doc 3	0,00208677686		Doc 4	0,00034917152
Doc 1	0,0004958677686		Doc 1	0,0007321508002
	QL+JM			**KL+DP**
	Q2			Q2
Doc 2	0,004666494365		Doc 2	0,0001186796663
Doc 4	0,001934661378		Doc 4	0,000139306395
Doc 3	0,00007588279489		Doc 3	0,0007314599197
Doc 1	0,00004958677686		Doc 1	0,0009091407868
	QL+JM			**KL+DP**
	Q3			Q3
Doc 4	0,201010101		Doc 4	0,0000466049582
Doc 1	0,1727272727		Doc 1	0,0001122630216
Doc 3	0,1329545455		Doc 3	0,0002304339773
Doc 2	0,04545454545		Doc 2	0,00034917152

Se observa que los scores no son los mismos, pero el orden de los documentos del ranking es casi idéntico, solo hay una diferencia en el rank de la Q1 donde el 2° y 3° difieren. Aún así los dos modelos son consistentes en cuanto a la estimación de relevancia que le asigna a los documentos.

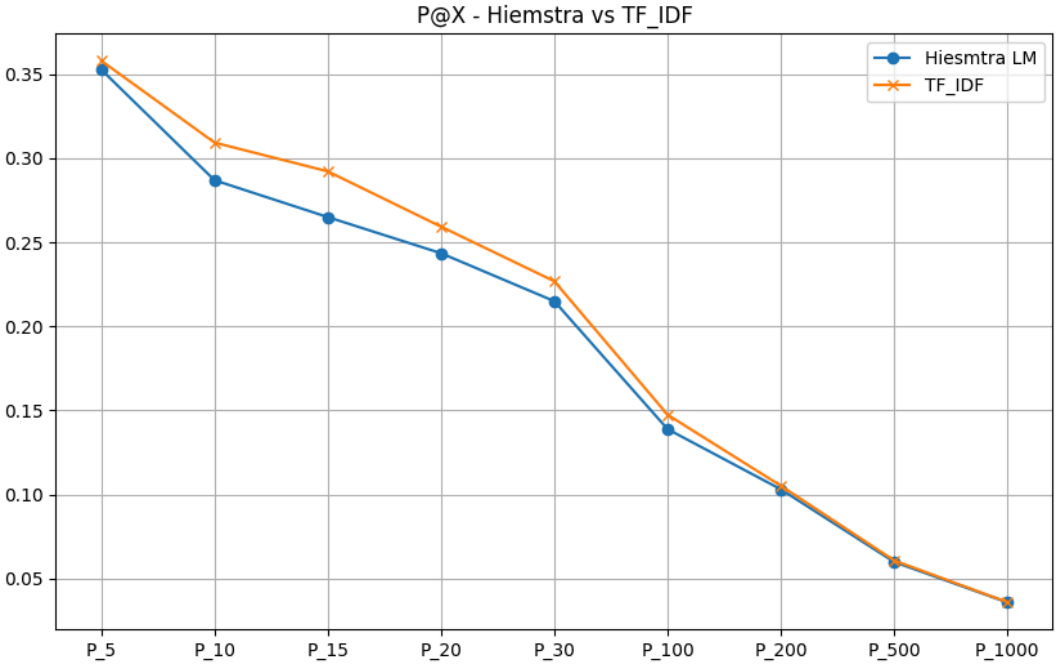
3. Utilizando modelos de lenguaje en Terrier (use Hiemstra LM), repita los experimentos del ejercicio 9 del TP de ”modelos” y compare los resultados con los anteriores. ¿Son consistentes? Calcule las métricas apropiadas para comparar los diferentes sistemas y configuraciones.

Tal como pedía el punto 9 se realizan comparaciones entre dos versiones de querys: una que manejaba los términos unívocos y otro que mantenía la frecuencia de los términos.

Sin Frecuencias:

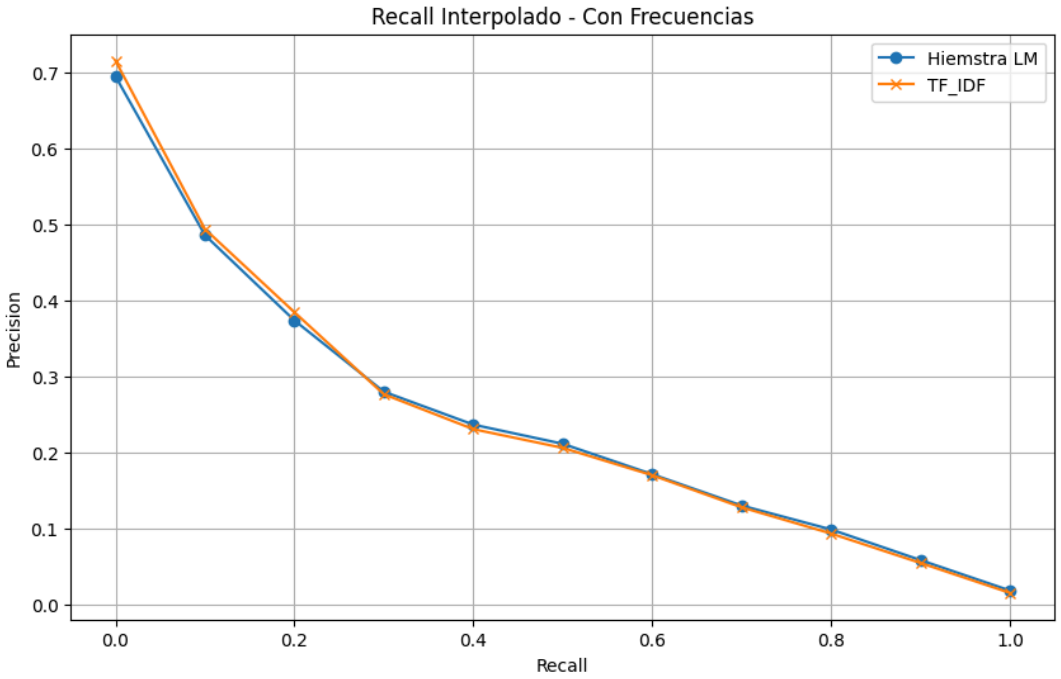


A lo largo del gráfico se ve que el modelo TF_IDF, tiene valores de precisión similares o superiores que el modelo de Hiemstra LM, se denota más la mayor precisión entre los intervalos [0 ; 0.3] y [0.5 , 0.7] de Recall.

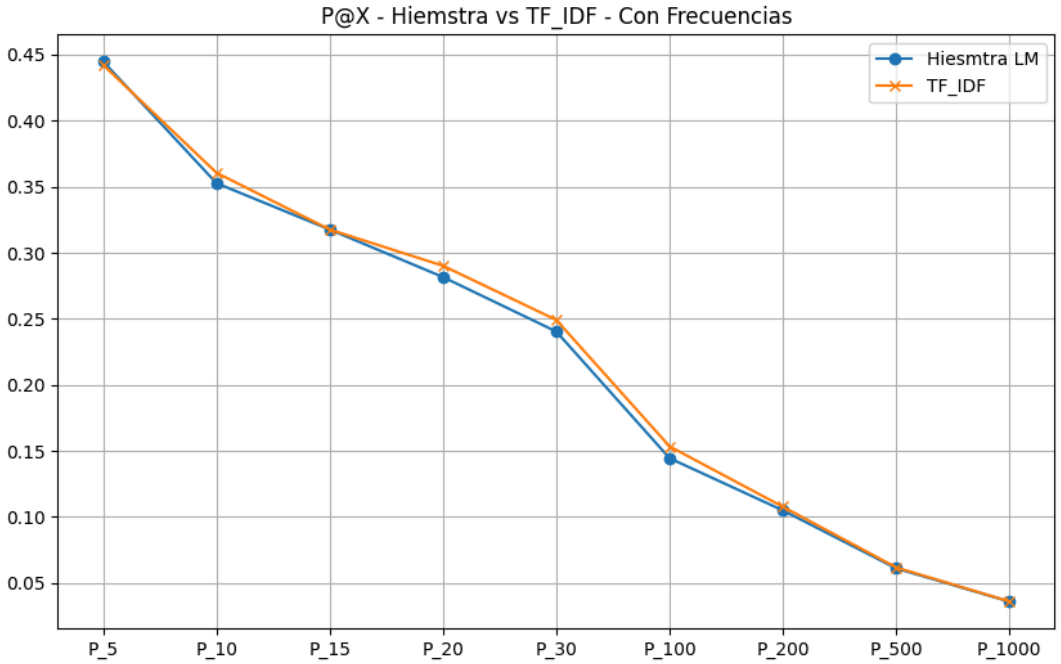


En comparación con el modelo Hiemstra LM, el modelo TF-IDF parece tener una precisión ligeramente mejor en los puntos de corte más altos (P@5 y P@10). Además, la precisión generalmente se mantiene más alta en los puntos de corte más bajos en comparación con el modelo Hiemstra LM.

Con Frecuencias:



Se ve una gran similitud en la precisión de ambos modelos habiendo incorporado las frecuencias de los términos en la query. Esto sucede porque el modelo de Hiemstra, también conocido como modelo de divergencia de Hiemstra, es un enfoque en la recuperación de información que se basa en la teoría de la divergencia de Kullback-Leibler (KL). Donde el cálculo del puntaje del documento se basa en la probabilidad del término en la colección, que puede estimarse utilizando las frecuencias de los términos en la colección.



Hiemstra LM ha mejorado significativamente en comparación con la prueba donde no se tiene en cuenta la frecuencia de los términos. La precisión en los primeros puntos de corte (P@5 y P@10) ha aumentado considerablemente, y la tendencia de disminución en la precisión a medida que aumenta el punto de corte aún se mantiene, aunque en menor medida que antes.

MAP Y NDCG:

Sin Frecuencias		
Modelo	map	ndcg
BR(TF_IDF)	187.836	547.212
BR(Hiemstra_LM)	180.204	539.450
Con Frecuencias		
Modelo	map	ndcg
BR(TF_IDF)	230.075	584.667
BR(Hiemstra_LM)	228.591	580.327

Para finalizar vemos que ambos modelos tienen un rendimiento bastante similar, con TF-IDF ligeramente por encima de Hiemstra ML en términos de MAP y NDCG, aunque la diferencia es mínima.