Ejercicio 1 - Modelos de Lenguaje

Victorio Scafati - 151038

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	1	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 / Nj)

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	0.07	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

Doc \ Term	pais	cultura
1	0.00	0.00
2	0.07	0.07
3	0.05	0.00
4	0.05	0.05

Ranking de documentos

Doc	Rank
1	0.0000
2	0.0051
3	0.0000
4	0.0028

Query 2 - ("pais libre cultura")

Tabla Documentos - Query 1-1

Doc \ Term	pais	libre	cultura
1	0.00	0.04	0.00
2	0.07	0.07	0.07
3	0.05	0.00	0.00
4	0.05	0.11	0.05

Ranking de documentos-1

Doc	Rank
1	0.0000
2	0.0004
3	0.0000
4	0.0003

Query 3 - ("software propietario licencia")

Tabla Documentos - Query 1-1-1

Doc \ Term	software	propetario	licencia
1	0.08	0	0
2	0.00	0	0
3	0.05	0	0
4	0.11	0	0

Ranking de documentos-1-1

Doc	Rank
1	0.0000
2	0.0000
3	0.0000
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

Doc \ Term	pais	cultura
1	0.0038	0.0225
2	0.0680	0.0668
3	0.0447	0.0225
4	0.0511	0.0499

Ranking de documentos-2

Doc	Rank
1	0.00001
2	0.00454
3	0.00011
4	0.00255

Query 2 - ("pais libre cultura")

Tabla Documentos - Query 1-1-2

Doc \ Term	pais	libre	cultura
1	0.0038	0.0410	0.0668
2	0.0680	0.0693	0.0025
3	0.0447	0.0050	0.0499
4	0.0511	0.0997	0.0250

Ranking de documentos-1-2

Doc	Rank
1	0.00001
2	0.00001
3	0.00001
4	0.00013

Query 3 - ("software propietario licencia")

Tabla Documentos - Query 1-1-1-1

Doc \ Term	software	propetario	licencia
1	0.0783	0	0
2	0.0063	0	0
3	0.0472	0	0
4	0.1010	0	0

Ranking de documentos-1-1-1

Doc	Rank
1	0.0000
2	0.0000
3	0.0000
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.

Ejercicio 2 - Modelos de Lenguaje

Kullback Leible + Dirichlet μ 2000

Query 1 - ("pais cultura")

Tabla Documentos - Query 1-2-1

Doc \ Term	pais	cultura
1	0.0370	0.0247
2	0.0377	0.0253
3	0.0376	0.0247
4	0.0376	0.0248

Tabla Documentos - Query 1-2-1-1

Doc \ Term	pais	cultura
1	0.5000	0.5000

Kullback Leible

Doc \ Term	pais	cultura
1	1.8774	2.1699
2	1.8640	2.1517
3	1.8668	2.1689
4	1.8657	2.1678

Ranking de documentos-2-1

Doc	Rank
1	4.04737
2	4.01567
3	4.03568
4	4.03353

Query 2 - ("pais libre cultura")

Tabla Documentos - Query 1-1-2-1

Doc \ Term	pais	libre	cultura
1	0.0370	0.0499	0.0247
2	0.0377	0.0501	0.0253
3	0.0376	0.0495	0.0248
4	0.0376	0.0505	0.0252

Tabla Documentos - Query 1-1-2-1-1

Doc \ Term	pais	libre	cultura
1	0.3333	0.3333	0.3333

Tabla Documentos - Query 1-1-2-1-2

Doc \ Term	pais	libre	cultura
1	1.0566	0.9135	1.2516
2	1.0477	0.9109	1.2395
3	1.0496	0.9176	1.2490
4	1.0488	0.9073	1.2414

Ranking de documentos-2-1-1

Doc	Rank
1	3.22178
2	3.19803
3	3.21615
4	3.19758

Query 3 - ("software propietario licencia")

Tabla Documentos - Query 1-1-1-1-1

Doc \ Term	software	propetario	licencia
1	0.0627	0	0
2	0.0621	0	0
3	0.0623	0	0
4	0.0629	0	0

Tabla Documentos - Query 1-1-1-1-1-2

Doc \ Term	software	propetario	licencia
1	0.3333	0.3333	0.3333

Tabla Documentos - Query 1-1-1-1-1-1

Doc \ Term	software	propetario	licencia
1	0.8034	0	0
2	0.8084	0	0
3	0.8064	0	0
4	0.8019	0	0

Ranking de documentos-2-1-1-1

Doc	Rank
1	0.80335
2	0.80837
3	0.80644
4	0.80193

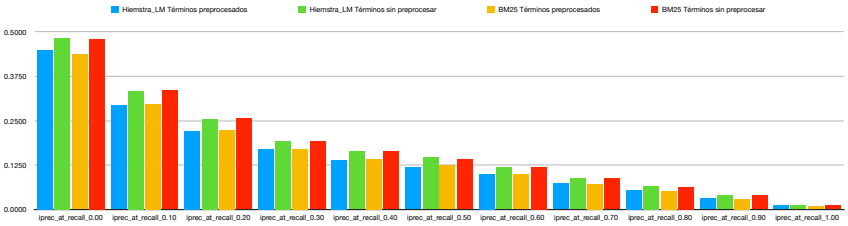
Observaciones

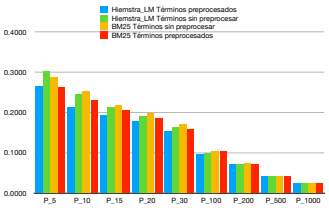
Podemos observar que la query #3 devolvió documentos. Por lo que ya no es un problema realizar una consulta por términos que no tengamos dentro del corpus. Esto se debe principalmente a que es una sumatoria y no una productoria la que se realiza en el ultimo paso.

Ejercicio 3 - TREC EVAL

TREC_EVAL - COMPARACIÓN

Parametros	Hiemstra_LM Términos preprocesados	Hiemstra_LM Términos sin preprocesar	BM25 Términos preprocesados	BM25 Términos sin preprocesar
num_q	112	112	112	112
num_ret	102114	102114	102114	102114
num_rel	3114	3114	3114	3114
num_rel_ret	2731	2740	2731	2740
map	0.1382	0.1574	0.1384	0.1578
gm_map	0.0069	0.0074	0.007	0.0076
Rprec	0.1545	0.1719	0.1571	0.1754
bpref	0.6098	0.611	0.6099	0.6113
recip_rank	0.422	0.4602	0.4002	0.4474
iprec_at_recall_0.00	0.4491	0.4842	0.4378	0.48
iprec_at_recall_0.10	0.2935	0.3326	0.297	0.3352
iprec_at_recall_0.20	0.2227	0.2547	0.2235	0.2558
iprec_at_recall_0.30	0.1721	0.1934	0.1706	0.1925
iprec_at_recall_0.40	0.1394	0.1625	0.1424	0.1635
iprec_at_recall_0.50	0.1192	0.1469	0.1244	0.1421
iprec_at_recall_0.60	0.1002	0.1192	0.0995	0.1203
iprec_at_recall_0.70	0.0745	0.0892	0.0715	0.0888
iprec_at_recall_0.80	0.0551	0.0657	0.053	0.0644
iprec_at_recall_0.90	0.0322	0.0399	0.0294	0.04
iprec_at_recall_1.00	0.012	0.0121	0.0092	0.0107
P_5	0.2661	0.3036	0.2643	0.2875
P_10	0.2125	0.2455	0.2313	0.2527
P_15	0.1917	0.2131	0.206	0.2185
P_20	0.1781	0.1911	0.1866	0.1969
P_30	0.1545	0.1643	0.1583	0.1702
P_100	0.0971	0.0992	0.1018	0.1045
P_200	0.0704	0.072	0.0717	0.0731
P_500	0.0409	0.0415	0.0414	0.0417
P_1000	0.0244	0.0245	0.0244	0.0245





Observaciones

Podemos observar que los valores de precisión a nivel de recall coinciden entre *BM25* y *Hiemstra_LM* (ambos sin preprocesar). Recordamos que durante el preprocesado se le realizo a las queries y lo único que realizamos fue eliminar términos duplicados. Los resultados de ambos modelos son muy parecidos, sobre todo si observamos el MAP que para todos los modelos osciló entre 0.13 para ambos modelos preprocesados y 0.16 para ambos modelos sin preprocesar.