Trabajo Práctico

Modelos de Recuperación de Información (y evaluación)

Fecha de Entrega: 08/04/2020

Victorio Scafati
victorioskfati@gmail.com

1. Utilizando la colección provista por el equipo docente[1], cuya estructura es la siguiente:

a) Calcule los conjuntos de respuestas usando el modelo booleano y el modelo vectorial (asuma en todos los casos TF = 1).

Modelo Booleano

Query (Id Términos)	Resultados AND (Doc Id)	Resultados OR (Doc Id)
(72, 117, 191)	Ninguno	**36,** 17, **5, 15, 25, 29, 32**
(147, 195, 196)	Ninguno	**8,10,18,32**, 1, 4, 6, 9, 19, 21, 29, 37, 38, **4, 9, 14, 19, 21, 36, 37, 38**
(55, 56, 141, 142, 147)	Ninguno	**14, 19, 32,** 22, **2, 11, 20, 22, 33, 34,** -, **8,10,18,32**
(147, 179, 180)	Ninguno	**8,10,18,32**, 7, 13, 27, 30, 35, **-**
(147, 182, 184)	8,10,18,32	**8,10,18,32**, 8,10,18,32, **8,10,18,32**,

b) Compare los resultados contra los relevantes y trate de explicar las diferencias.
c) Usando las necesidades de información reescriba los 5 queries y repita la operación.
d) Indique si pudo mejorar la eficiencia a partir de las nuevas consultas.

Los resultados y observaciones de los puntos b), c) y d) se pueden encontrar en el archivo **RI-TP03.01.pdf**

2. Dados los siguientes documentos, arme la matriz termino-documento (TD)[2]

> **Doc 1** = {El software libre ha tenido un papel fundamental en el crecimiento de Internet. Ademas, Internet ha favorecido la comunicación entre los desarrolladores de software.}
> **Doc 2** = {La mayor riqueza que tiene un pas es la cultura, eso lo hace más libre.}
> **Doc 3** = {La producción de software es fundamental para nuestro pas, como as también lo es la producción de tecnología de hardware y comunicación}
> **Doc 4** = {La cultura del software libre esta en crecimiento. Es fundamental que nuestro pas incorpore software libre en el estado.}

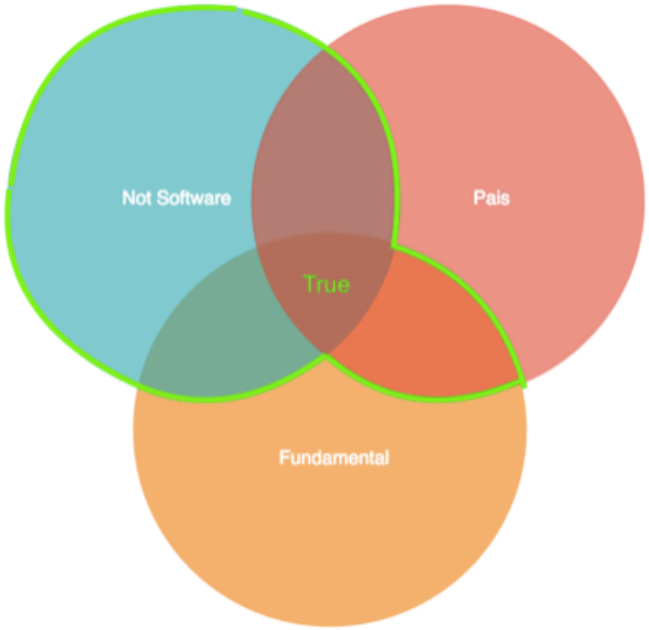
¿Qué documentos se recuperan en cada caso para las siguientes consultas booleanas? (Muestre mediante operaciones con conjuntos como se resuelven las consultas)

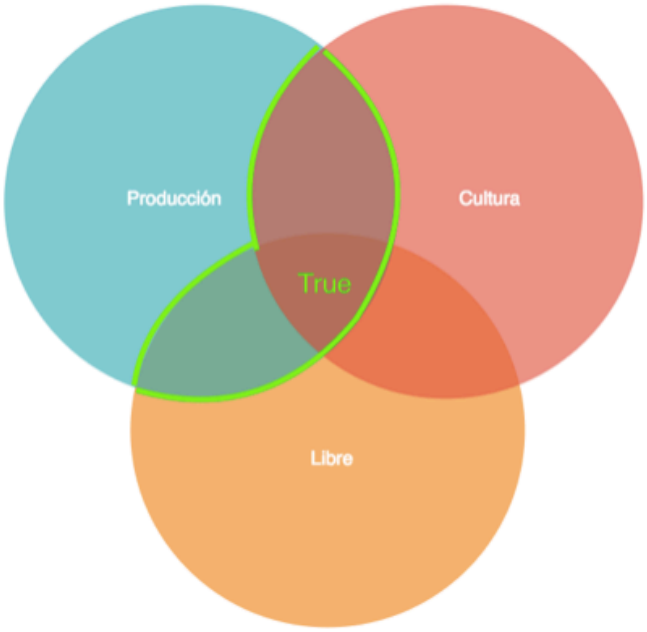
a) (not software) or (pais and fundamental)
b) produccion and (cultura or libre)
c) fundamental or libre or pais

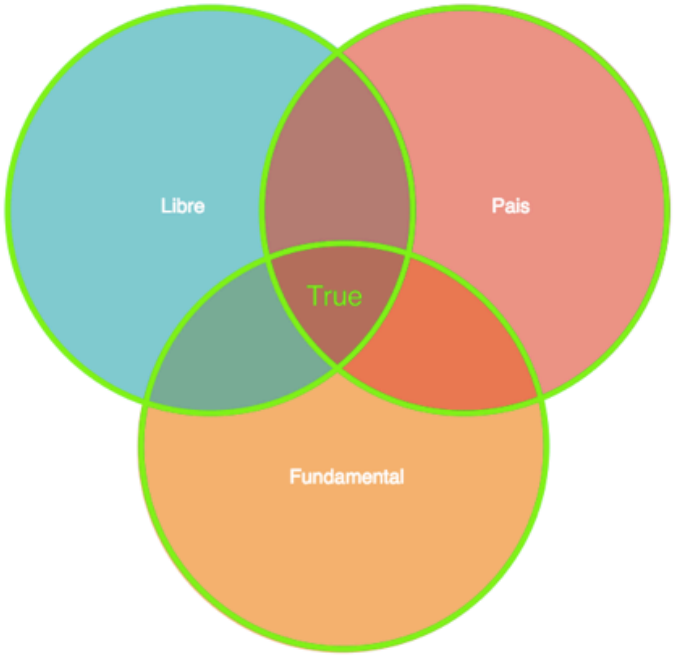
Term \ Doc	D1	D2	D3	D4
comunicacion	1	0	1	0
crecimiento	1	0	0	1
cultura	0	1	0	1
desarrolladores	1	0	0	0
estado	0	0	0	1
fundamental	1	0	1	1
hardware	0	0	1	0
internet	1	0	0	0
libre	1	1	0	1
mayor	0	1	0	0
pais	0	1	1	1
papel	1	0	0	0
produccion	0	0	1	0

#	Query	Docs
1	Not software OR (pais and fundamental)	2, 3, 4
2	Produccion and (cultura or libre)	-
3	Fundamental or libre or pais	1, 2, 3, 4

A continuación realizamos el análisis de la queries con diagramas de Venn:







Destacamos que las **lineas verdes** indican los items que cumplen con la query solicitada.

3. Utilizando los documentos del ejercicio anterior arme la matriz TD pero calculando w_{ij} como la frecuencia del i-ésimo termino en el j-ésimo documento. Calcule el ranking para la siguientes consultas utilizando como métrica el producto escalar y luego repita con la métrica del coseno.

a) software
b) país libre
c) producción software país

Para ver el desarrollo de este punto remitirse al archivo **RI-TP03.03-04.pdf,** las conclusiones de este punto se encuentran en el siguiente donde comparamos puntos 3 y 4.

4. Rearme la matriz del ejercicio anterior pero calcule los pesos de acuerdo a TF_IDF. Repita todas las consultas (por ambas métricas). ¿Puede obtener alguna conclusión?

Conclusión

A continuación comparamos los rankings obtenidos en los puntos 3 y 4**.**

Query 1

Doc	TF		TF_IDF	
	Rank (Prod Escalar)	Rank (Coseno)	Rank (Prod Escalar)	Rank (Coseno)
1	0.2	1	0.6	1
2	0	0	0	0
3	0.13	1	0.42	1
4	0.22	1	0.6	1

Query 2

Doc	TF		TF_IDF	
	Rank (Prod Escalar)	Rank (Coseno)	Rank (Prod Escalar)	Rank (Coseno)
1	0.05	0.71	0.42	0.71
2	0.2	1	0.84	1
3	0.07	0.71	0.42	0.71
4	0.17	0.95	1.02	0.99

Query 3

Doc	TF		TF_IDF	
	Rank (Prod Escalar)	Rank (Coseno)	Rank (Prod Escalar)	Rank (Coseno)
1	0.07	0.58	1.02	0.28
2	0.07	0.58	0.42	0.2
3	0.17	0.95	6.42	0.97
4	0.11	0.77	1.19	0.28

Como primer comentario podemos ver que TF e TF_IDF aplicado sobre métrica del coseno no se vio muy afectada salvo en pocos casos donde discrepo, por otro lado podemos ver que la utilización de TF vs TF_IDF si comprometió el resultado de las métricas con producto escalar en donde en breves ocasiones coincidieron.

Para ver el desarrollo del punto remitirse al archivo **RI-TP03.03-04.pdf**

5. Utilizando Terrier[3] indexe la colección Wiki-Small[4]. Tome 5 necesidades de información y { de forma manual { derive una consulta (query). Para cada una, pruebe la recuperación por los modelos basados en TF IDF y BM25. ¿Como se comportan los rankings? Calcule el coeficiente de correlación para los primeros 10, 25 y 50 resultados. ¿Qué conclusiones obtiene?

Queries propuestas

Id	Query
1	pink floyd top album
2	football
3	tenis grandslam 2000
4	olympic games
5	lead singer acdc

Conclusión

Podemos concluir que es casi indiferente aplicar TF_IDF y BM25, las 5 queries seleccionadas obtuvieron una correlación de entre 0.99 y 1. Lo que nos indica que ambos algoritmos rankean de manera similar. A continuación se puede ver la correlación de ambos modelos a través del coeficiente de spearman.

Id	Query	Coeficiente de Spearman
1	pink floyd top album	0.999998237502901
2	football	1
3	tenis grandslam 2000	1
4	olympic games	0.9999921094
5	lead singer acdc	0.9997587398

Para ver el desarrollo de este punto abra el archivo **RI-TP03.05.xslx**

6.Escriba un peque~no programa que lea un directorio con documentos de texto y arme una estructura de datos en memoria para soportar la recuperación. Luego, debe permitir ingresar un query y devolver un ranking de los documentos relevantes utilizando el modelo vectorial. Se debe soportar la ponderación de los términos de la consulta. Implemente las versiones sugeridas en MIR.

Comentarios

En este punto se reutilizo la mayoría del código del TP anterior. Por lo que se realizan operaciones de normalización, eliminación de palabras vacías, stemming, etc.

Para ver el **código** de este punto abra el archivo **6/RI-TP03.06.py.** Además dentro de la carpeta 6/ podrá encontrar los archivos:

queries.json donde especificamos las queries.
stopword-list.txt que es el archivo de palabras varias (tomé el mismo archivo de terrier)

Ejecución de código

Para ejecutar el código ejecutar corremos en una terminal, indicando el <*CORPUS*>:

```
python3 RI-TP03.06.py -t porter -d <CORPUS> -q queries.json -s stopword-list.txt
```

Tenga en cuenta que el programa tiene un **help** el cual indica la forma de ejecución:

```
python3 RI-TP03.06.py -h
```

6. Indexe la colección del ejercicio 5 con su software. Ejecute las consultas y compare los resultados con los obtenidos con Terrier. ¿Son consistentes?

No, es mas de un caso la correlación de terrier (usando TF_IDF) es muy diferente a la de mi software (**RI-TP03.06.py**), en el mejor de los casos mi software obtuvo un 0.6 de correlación. A continuación la correlación entre terrier (usando TF_IDF) y mi software

Id	Query	Coeficiente de Spearman
1	pink floyd top album	-3.29061224489796
2	football	-0.179831932773109
3	tenis grandslam 2000	0.634765906362545
4	olympic games	-4.58223289315726
5	lead singer acdc	-0.666650660264106

8. Se requiere evaluar la performance en la recuperación de un sistema. Para una consulta q1, dicho sistema entrego la siguiente salida.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R	N	N	R	R	N	N	N	N	R	N	N	N	R	N

Los documentos identificados como R son los relevantes, mientras que las N's corresponden a documentos no relevantes a q1. Suponga que existen en el corpus otros 6 documentos relevantes a q1 que el sistema no recupero. A partir de esta salida calcule las siguientes medidas:

a) Recall y Precision para cada posición j

Docs	W	Recall (E)	Precision (P)
1	1	0.09 (1/11)	1.00 (1/1)
2			0.50 (1/2)
3			0.33 (1/3)
4	2	0.18 (2/11)	0.50 (2/4)
5	3	0.27 (3/11)	0.60 (3/5)
6			0.50 (3/6)
7			0.43 (3/7)
8			0.37 (3/8)
9			0.33 (3/9)
10	4	0.36 (4/11)	0.40 (4/10)
11			0.36 (4/11)
12			0.33 (4/12)
13			0.31 (4/13)
14	5	0.45 (5/11)	0.36 (5/14)
15			0.33 (5/15)

$$E = \frac{w}{X}$$

$$P = \frac{w}{Y}$$

w: cantidad de docs relevantes recuperados.
X: cantidad de docs relevantes.
Y: cantidad de docs recuperados.

b) Precision promedio

$$AVG_q = 0.44$$

c) Precision al 50% de Recall

$$E = 0.5 \rightarrow P = 0.0$$

d) Precision interpolada al 50% de Recall

Docs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R	x			x	x					x				x	
	1	0.5	0.33	0.5	0.6	0.5	0.43	0.38	0.33	0.4	0.36	0.33	0.31	0.36	0.33

$$E = 0.5 \rightarrow P = 0.0$$

e) Precision-R

$$Precision_R = \frac{4}{11} = 0.36$$

9. Utilizando la colección de prueba CISI[5] y Terrier se debe realizar la evaluación del sistema. Para ello, es necesario construir un índice con los documentos de la colección y luego ejecutar las consultas, las cuales se deben armar a partir de los términos que considere de las necesidades de información. Los resultados deben ser comparados contra los juicios de relevancia de la colección utilizando el software trec_eval[6]. Realizar el análisis y escribir un reporte indicando los resultados obtenidos, junto con la gráfica de R{P en los 11 puntos standard. Realice dos experimentos: en el primero, no considere la frecuencia de los términos en el query mientras que en el segundo lo debe tener en cuenta.

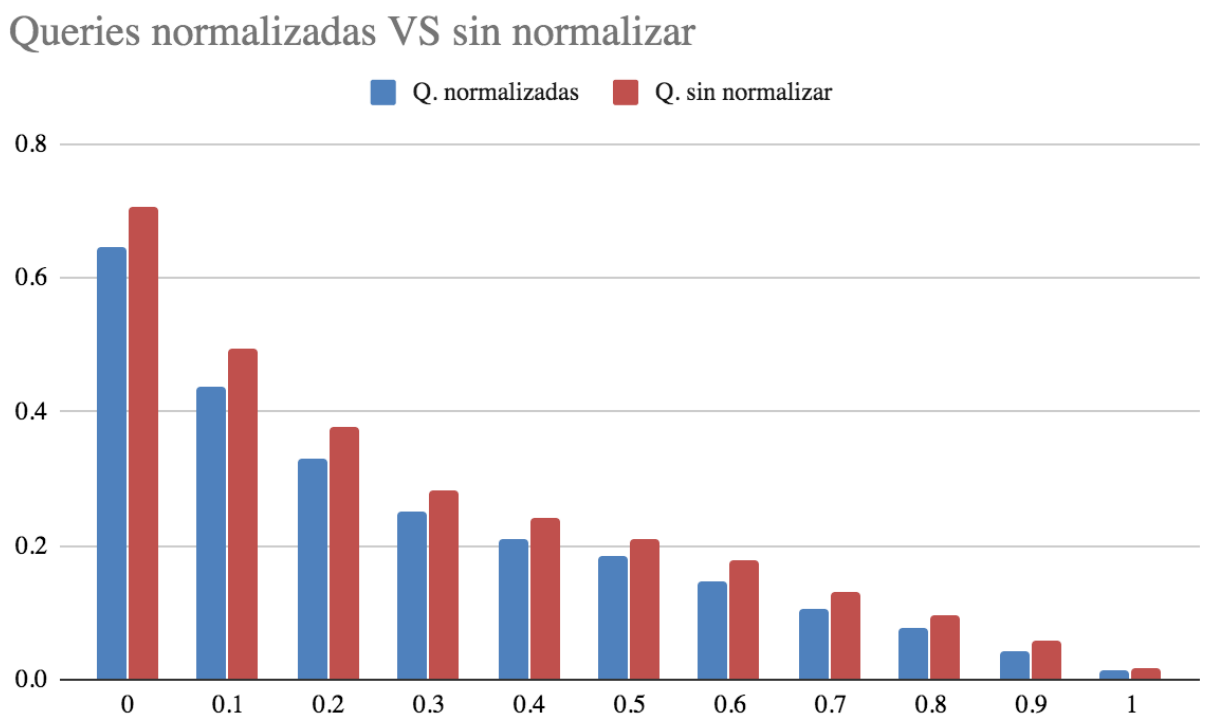
Análisis de Resultados

	Sin Procesar	Procesados*
MAP	0.2325	0.2040
Num. Rel recuperados	2740	2731
P@5	0.4237	0.4237
P@10	0.3724	0.3724
P@15	0.2901	0.3219

*Terminos Normalizados y sacamos palabras repetidas.

Conclusiones

Luego de realizar los experimentos con ambas queries podemos observar en el siguiente gráfico que la diferencia no es tal. Es más podemos advertir que los resultados sin procesar obtuvieron mejor MAP y recuperaron hasta 9 documentos relevantes más que las queries procesadas.



10. Dadas las salidas de tres sistemas de recuperación de información para 3 consultas cualquiera[7] y los juicios de relevancia creados por asesores humanos[8], calcule para cada sistema:

a) La precision media
b) La precision media a intervalos de Recall de 20%
c) P@5, P@10, P@20

Luego, exponga un escenario posible y medidas complementarias para decidir que sistema utilizar.

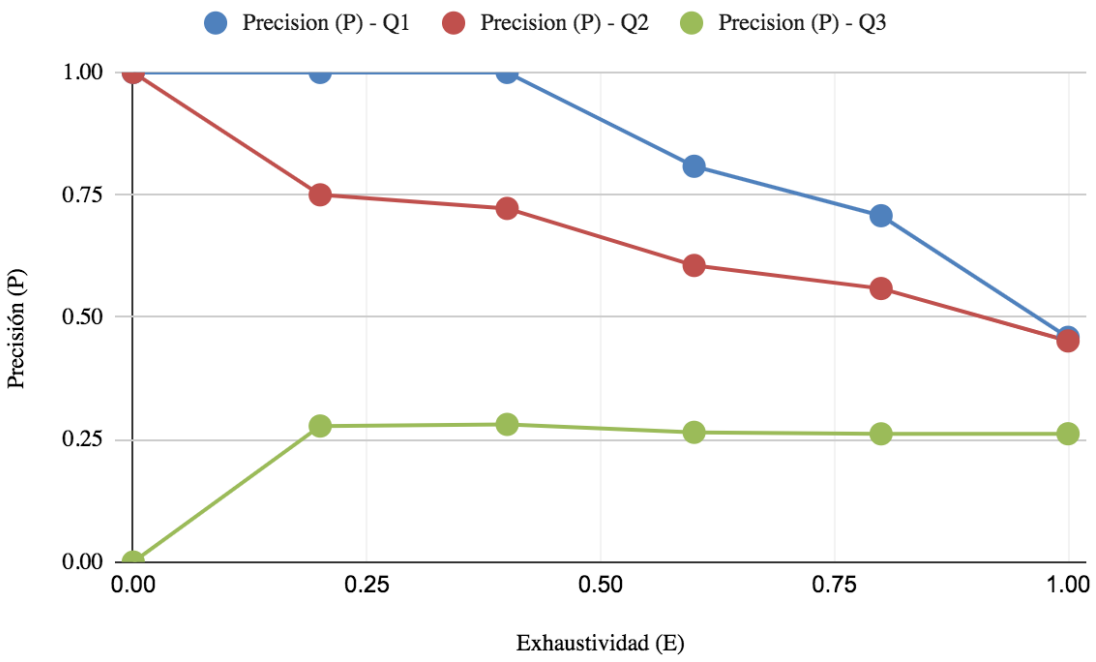
El desarrollo de este punto se puede encontrar en **/10/TP03.10.xslx,** aquí realizamos algunas conclusiones.

	MAP
SRI - A	0.5567903996
SRI - B	0.2884347937
SRI - C	0.3730602951

Sistema a utilizar - SRI A

Luego de comparar las recuperación de los tres sistemas podemos ver una mayor precisión para el sistema SRI A, el cual tuvo una precisión media (MAP) de **0.56**, además esta precisión se puede observar a continuación donde presentamos precisión media junto con los intervalos de Recall.

Precision (P) - Q1, Precision (P) - Q2 y Precision (P) - Q3



Precision Promedio (Queries)