Trabajo Práctico

Modelos de Recuperación de Información (y evaluación)

Bibliografía: [MIR] Capítulo 2 y 3, [MAN] Capítulos 1,7,8,12.

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

vocabulary.txt	→ [id_termino, idf, término]
documentVectors.txt	→ [id_doc, lista(id_terminos)]
queries.txt	→ [id_query, lista(id_terminos)]
relevants.txt	→ [id_query, $lista_{relevantes}$ (id_doc)]
informationNeeds.txt	→ [id_in, texto_libre]

calcule los conjuntos de respuestas usando el modelo booleano y el modelo vectorial (asuma en todos los casos TF = 1) y compare los resultados contra los relevantes. Trate de explicar las diferencias. A continuación, usando las necesidades de información reescriba los 5 queries y repita la operación. Indique si pudo mejorar la eficiencia a partir de las nuevas consultas.

2) Dados los siguientes documentos, arme la matriz término-documento (TD). Nota: No tenga en cuenta los artículos, preposiciones y conectores.

Doc 1: *"El software libre ha tenido un papel fundamental en el crecimiento de Internet. Además, Internet ha favorecido la comunicación entre los desarrolladores de software."*

Doc 2: *"La mayor riqueza que tiene un país es la cultura, eso lo hace más libre. "*

Doc 3: *"La producción de software es fundamental para nuestro país, como así también lo es la producción de tecnología de hardware y comunicación."*

Doc 4: *"La cultura del software libre está en crecimiento. Es fundamental que nuestro país incopore software libre en el estado."*

¿Que documentos se recuperan en cada caso para las siguientes consultas booleanas?

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

Muestre mediante operaciones con conjuntos cómo se resuelven las consultas.

3) Utilizando los documentos del ejercicio anterior arme la matriz TD pero calculando w_{ij} como la frecuencia del i-ésimo término 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

1 Esta colección corresponde a un subconjunto de la "Cystic Fibrosis Collection". Los ejercicios fueron adaptados del curso del Prof. Berthier Ribeiro-Neto (http://sunsite.dcc.uchile.cl/irbook/)

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?

5) Utilizando Terrier[2] indexe la colección provista por el equipo docente. 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. ¿Cómo se comportan los rankings? Calcule el coeficiente de correlación para los primeros 10, 25 y 50 resultados. ¿Qué conclusiones obtiene?

7) Escriba un pequeño 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].

8) Modifique su programa del ejercicio anterior para soportar consultas mediante el modelo booleano extendido (con p-norms). Ejecute las mismas consultas del ejercicio 8 usando ambos operadores booleanos (AND y OR), con p = 3 y 4 y compare los resultados. Indique cuáles son documentos relevantes y – bajo algún criterio propio – cuál resulta mejor.

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

10) Se requiere evaluar la performance en la recuperación de un sistema. Para una consulta q_1, dicho sistema entregó 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 q_1. Suponga – además – que existen en el corpus otros 6 documentos relevantes a q_1 que el sistema no recuperó. A partir de esta salida calcule las siguientes medidas:

a) Recall y Precision para cada posición j
b) Recall y Precision promedio
c) Precisión al 50% de Recall
d) Precisión interpolada al 50% de Recall
e) Precisión-R

Finalmente, realice las gráficas interpolada y sin interpolar. Luego, interprete brevemente los resultados y brinde una explicación.

11) Utilizando la colección de prueba CISI[3] 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. Los resultados deben ser comparados contra los juicios de relevancia de la colección utilizando el software trec_eval[4]. Realizar el análisis y escribir un reporte indicando los resultados obtenidos, junto con la gráfica de R–P en los 11 puntos standard.

2 http://www.terrier.org/
3 ftp://ftp.cs.cornell.edu/pub/smart/cisi/
4 http://trec.nist.gov/trec_eval/

12) A continuación se prentan las salidas de tres sistemas de recuperación de información para 3 consultas cualquiera y los juicios de relevancia creados por asesores humanos.

Query 1

SRI A	SRI B	SRI C
5	7	11
2	16	2
1	9	10
15	18	21
9	4	1
19	17	5
3	8	22
6	5	9
18	11	20
4	15	3
12	12	23
8	10	4
11	6	19
17	2	6
7	19	15
20	20	25
10	21	18
21	25	16
16	22	7
23	1	8
13	23	17
22	13	12
24	24	13
25	3	14
14	14	24

Query 2

SRI A	SRI B	SRI C
12	16	7
14	23	23
8	3	3
23	13	13
9	4	4
5	14	8
20	17	17
21	11	25
4	7	16
19	21	21
3	18	18
10	10	22
7	19	19
18	22	10
11	20	20
17	1	5
24	2	2
13	12	12
2	15	15
16	5	1
6	8	14
22	6	24
15	9	9
25	25	11
1	24	6

Query 3

SRI A	SRI B	SRI C
1	5	23
18	24	9
11	13	25
16	19	18
19	17	1
10	12	2
23	22	11
21	4	10
3	1	19
6	8	16
22	11	12
12	3	22
17	16	15
2	23	8
13	25	24
24	18	14
25	2	7
14	20	13
5	14	17
4	10	20
7	6	21
9	21	4
15	15	5
20	7	6
8	9	3

Juicios de Relevancia

D	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Q1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Q2	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1
Q3	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	1	0

Para cada sistema calcule:

- La precisión media
- La precisión media a intervalos de Recall de 20%.
- P@5, P@10, P@20

Exponga un escenario posible y medidas complementarias para decidir qué sistema utilizar.