Modelos de Recuperación de Información (y evaluación)

Fecha entrega: 10/04/2024

Para la entrega del TP resuelto arme un único archivo comprimido (.tar.gz) y envíelo a través del siguiente formulario: https://forms.gle/fcTfBySrfNNSordz9 el cual se encontrará habilitado hasta la fecha de entrega establecida.
Bibliografía sugerida: MIR [1] Capítulos 2 y 3, MAN [2] Capítulos 1,7,8,12.

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

   - vocabulary.txt → [id_término, idf, término]
   - documentVectors.txt → [id doc, lista(id_términos)]
   - queries.txt → [id query, lista(id_términos)]
   - relevants.txt → [id query, listarelevantes (id_doc)]
   - informationNeeds.txt → [id_in, texto_libre]

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

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

   - Doc1 = {El software libre ha tenido un papel fundamental en el crecimiento de Internet. Además, Internet ha favorecido la comunicación entre los desarrolladores de software.}
   - Doc2 = {La mayor riqueza que tiene un país es la cultura, eso lo hace más libre.}
   - Doc3 = {La producción de software es fundamental para nuestro país, como así también lo es la producción de tecnología de hardware y comunicación}
   - Doc4 = {La cultura del software libre está en crecimiento. Es fundamental que nuestro país incorpore software libre en el estado.}

[1] Esta colección (https://bit.ly/3cqhqAg) corresponde a un subconjunto de la “Cystic Fibrosis Collection”. Los ejercicios fueron adaptados del curso del Prof. Berthier Ribeiro-Neto
[2]Nota: no tenga en cuenta los artículos, preposiciones y conectores

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

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

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 ḿetrica el producto escalar y luego repita con la métrica del coseno.

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

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[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. ¿Cómo se comportan los rankings? Calcule el coeficiente de correlación para los primeros 10, 25 y 50 resultados. ¿Qué conclusiones obtiene?

6) 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 [1].

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

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

3 http://www.terrier.org/

4 http://dg3rtljvitrle.cloudfront.net/wiki-small.tar.gz

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, además, que existen en el corpus otros 6 documentos relevantes a *q1* que el sistema no recuperó. A partir de esta salida calcule las siguientes medidas:

a) Recall y Precisión para cada posición j
b) Precisión 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.

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 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

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

[5] http://ir.dcs.gla.ac.uk/resources/test_collections/cisi/
[6] https://trec.nist.gov/trec_eval/
[7] http://www.labredes.unlu.edu.ar/sites/www.labredes.unlu.edu.ar/files/site/data/ri/Ejer-Evaluacion-OUTPUT.ods
[8] http://www.labredes.unlu.edu.ar/sites/www.labredes.unlu.edu.ar/files/site/data/ri/Ejer-Evaluacion-JR.ods

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

Bibliografía

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval: The Concepts and Technology Behind Search. Addison-Wesley Publishing Company, USA, 2nd edition, 2008.

[2] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Sch ¨utze. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA, 2008.