Trabajo Práctico
Modelos de Recuperación de Información

Feha de Entrega: 30/09/2011

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

1) Utilizando la colección provista por el equipo docente[1] (vocabulary.txt y documentVectors.txt) y las 5 consultas (queries.txt) calcule los conjuntos de respuestas usando el modelo booleano y el modelo vectorial y compare los resultados con los relevantes (relevants.txt). Trate de explicar las diferencias. A continuación, usando las necesidades de información (informationNeeds.txt) 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.

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

Repita la operación utilizando la métrica del coseno.

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?

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/)

5) Este es un ejercicio clásico extraido de la bibliografía (traducción). Suponga que tiene una colección de 6 documentos, con términos que aparecen más de una vez.

Doc A: cat, care, persian
Doc B: cat, care, persian, cat, care, persian, cat, care, persian
Doc C: cat, cat, cat, cat, cat, cat, cat, cat, cat
Doc D: cat, care, persian, dog, dog, dog, dog, dog, dog
Doc E: cat, care, dog
Doc F: care

También, se cuenta con un conjunto de *queries*:

Query 1: cat
Query 2: cat, care
Query 3: cat, care, persian
Query 4: cat, cat, care

a) Indique los rankings para cada consulta utilizando el modelo vectorial con pesos mediante TF*IDF.
b) ¿Cambia la respuesta al query 3 si se pondera mediante TF los términos de los documentos? En caso afirmativo, explique el motivo.
c) ¿Cambian las respuestas si se utiliza el coeficiente de DICE en vez del coseno? Indique qué ocurre.

6) Utilizando la herramienta Lemur Tookit (http://www.lemurproject.org/) 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 vectorial y BM25. ¿Cómo se comportan los rankings? Calcule el coeficiente de correlación para los primeros 10, 25 y 50 resultados. ¿Qué conclusiones obtiene?

7) A partir de la lectura del artículo "*Pivoted Document Length Normalization*" por by Amit Singhal, Chris Buckley, Mandar Mitra y Ar Mitra explique cómo funciona la normalización mediante pivot y qué problemas resuelve.

8) 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], páginas 28, 29 y 30.

9) 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.

10) Indexe una pequeña colección con Lemur y luego con su software (del ejercicio 8). Ejecute las consultas y compare los resultados. ¿Son consistentes?

11) Genere una nueva versión de su programa del ejercicio 8 para implementar el modelo probabilístico. El mismo debe recibir el *query* y un parámetro de iteracción para la reatroalimentación del modelo. Haga pruebas y compare los resultados con su programa anterior.

12) Calcule el modelo de lenguaje (unigramas) para los documentos del ejercicio 2. Utilizando el modelo de Query Likelihood calule los rankings para las siguiente consultas:

a) país cultura
b) país libre cultura
c) software propietario licencia

¿Qué problemas encontró? Luego, calcule las probabilidades de los términos utilizando una combinación con el ML de la colección (suavizado Jelinek-Mercer). Compare con las probabilidades anteriores y explique las diferencias. Repita las consultas con los nuevos valores. Explique los resultados.

13) Utilizando modelos de lenguaje en Lemur, repita los experimentos del ejercicio 6 y compare los resultados con los anteriores. ¿Son consistentes?