Repositorios de Información

Recuperación de información

Daniel Gayo Avello

Ejercicio 1

- Sea la siguiente colección de documentos:
  - `d1 = "Big cats are nice and funny"`
  - `d2 = "Small dogs are better than big dogs"`
  - `d3 = "Small cats are afraid of small dogs"`
  - `d4 = "Big cats are not afraid of small dogs"`
  - `d5 = "Funny cats are not afraid of small dogs"`
- Normalizar los documentos (paso a minúsculas y *stemming*)
- Generar el vocabulario correspondiente a la colección

Ejercicio 2

- Dada la colección de documentos del Ejercicio 1 construir la matriz de documentos-términos (no el índice invertido) necesaria para un modelo booleano (i.e., sin ponderación de términos)

Ejercicio 3

- Dada la matriz del Ejercicio 2, resolver las siguientes consultas (**¡atención!** Lo fundamental no es la lista de resultados sino las operaciones a realizar con la matriz para obtener dicha lista)
  - `q1 = funny AND dog`
  - `q2 = nice OR dog`
  - `q3 = big AND dog AND NOT funny`

Ejercicio 4

- Dada la matriz del Ejercicio 2 suponer que se implementa un modelo basado en conjuntos con la distancia coseno $\frac{|X \cap Y|}{|X| \cdot |Y|}$
- Dada la consulta `q = {big, cat, funny, small, dog}` calcular la relevancia de cada documento para la misma y generar una lista descendente de resultados

Ejercicio 5

- Dada la colección del Ejercicio 1 calcular la ponderación TF (para cada término en cada documento) y la IDF (para cada término en la colección) $\operatorname{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$
- Generar el correspondiente fichero invertido anotando las puntuaciones TF en los documentos y las puntuaciones IDF en los términos

Ejercicio 6

- Dado el fichero invertido del Ejercicio 5 determinar la similitud coseno de la consulta `big cat funny small` con cada uno de los documentos

Ejercicio 7

- Sea una colección de 100 documentos `d1..d100`
- Dada una consulta *q* el conjunto de documentos relevantes para el ususario es `D* = {d2, d13, d43, d65, d89}`
- Para dicha consulta un sistema de RI retorna el siguiente conjunto de resultados `R = {d2, d13, d42, d65, d66, d88, d95}`
- En base a esa información calcular precisión y exhaustividad
- ¿Cómo se podría dibujar una curva precisión/exhaustividad?

Ejercicio 8

- Reflexiona sobre lo siguiente:
  - ¿Por qué no se usa `grep` para hacer recuperación de información?
  - ¿Por qué no se usan bases de datos relacionales para hacer recuperación de información?
  - ¿Cuál crees que es el paso más complejo en el proceso de creación de un índice para un modelo vectorial?
- Googlewhack es un juego que consiste en encontrar un par de términos que retornen exactamente un documento al buscar en Google. Trata de localizar alguna de dichas combinaciones.