Técnicas para Big Data

Clase 05: TF - IDF

¿Cómo buscar texto?

Cuando buscamos “Chilean Mammal”:

- El sistema encuentra todos los documentos que tienen Chilean y Mammal
- ¿Pero cómo lo hacemos para ordenar los resultados?
- Puede ser problemático en grandes bases de datos

Índices Invertidos

- Para hacer la búsqueda eficiente utilizamos índices invertidos
- Para cada palabra del universo de documentos, guardamos punteros que nos indican dónde están los documentos

Índices Invertidos

	Documento 1	Documento 2	Documento 3	...
Chile	1	0	1	
of	1	1	1	
town	0	0	1	
commune	0	1	0	
...				

TF - IDF

Principio 1:

- El puntaje es proporcional a la cantidad de veces que aparece la palabra en el documento

Principio 2:

- El puntaje es inversamente proporcional a la cantidad de documentos en los que aparece la palabra

TF - IDF

Term Frequency:

- $F_D(t)$ = Número de veces que aparece **t** en **D**

Inverse Document Frequency:

- IDF(t) = log(número de documentos / número de documentos en los que aparece **t**)

TF - IDF

$$TF - IDF = F_D(t) \cdot IDF(t)$$

TF - IDF

Ejemplo

- D1: Ojo por ojo, diente por diente
- D2: Ojo por ojo, y el mundo acabará ciego
- D3: Si luchas contra el mundo, ponte del lado del mundo

Calcular el TF-IDF de “ojo” y “mundo” para cada documento

TF - IDF

- Hay distintas funciones para TF e IDF
- Generalmente se incorporan funciones para Stemming y Stop Words
- Cada compañía tiene su receta, depende además del idioma

TF - IDF

Búsqueda de documentos

- Se genera una matriz en donde las dimensiones son las palabras y los documentos
- Cada “casillero” señala el TF - IDF de la palabra en cada documento
- Cuando un usuario busca una frase, se genera un vector y se retornan los documentos con vectores más similares

Técnicas para Big Data

Clase 05: TF - IDF