

Notas 2do parcial - Ema

☒ Archive ☐

TEÓRICO

https://prod-files-secure.s3.us-west-2.amazonaws.com/206ee924-3842-491f-9926-b87d27a727f3/1c1e61f5-eeab-40ca-a69d-c500003fbeb0/Captulo_4_-_Almacenamiento_de_tablas_en_archivos_e_ndices.pdf

https://prod-files-secure.s3.us-west-2.amazonaws.com/206ee924-3842-491f-9926-b87d27a727f3/46156097-ad3e-4ed7-afcc-fa23bcd801a7/Captulo_4_-_Procesamiento_de_consultas.pdf

https://prod-files-secure.s3.us-west-2.amazonaws.com/206ee924-3842-491f-9926-b87d27a727f3/330490cb-5898-483c-8f90-f7a6ee963957/Captulo_4_-_Factor_de_selectividad_y_operadores_fsicos.pdf

https://prod-files-secure.s3.us-west-2.amazonaws.com/206ee924-3842-491f-9926-b87d27a727f3/cde7c58b-4432-4288-b906-040fc0c6ec1d/Captulo_5_-_Normalizacin.pdf

Práctico - RETORNO DE LA INFORMACIÓN

Pasado por una compañera:

Hola! Esto es lo que logré hacer del práctico de RI. Las únicas aclaraciones que te hago es que me faltan los ejercicio 8 y 10 (el 9 no hay que hacerlo) pero el 10 se respondió en clases y es una pavada. Y que el ejerció 7 me faltó la partecita esa de dar algoritmos y el ejercicio dos no le des mucha bola porque me ayude con chatGPT.

Práctico 5 - Retorno de la información

Hoja 1

Ejercicio 0.

La recuperación de info. es el resultado de una consulta cuyos datos del resultado son relevantes para la misma.

En el contexto de la web el resultado de la consulta incluye enlaces a páginas web que sean relevantes para la consulta + un pequeño texto que describe el contenido de la página.

Ejercicio 1.

$\text{Func. Find_documents}(\text{index}, \text{query_terms}, k) \{$

$\text{result} = []$

 for term in query_terms

 if term in index

$\text{docs_containing_terms} = \text{index}[\text{term}]$

 if not result:

$\text{results} = \text{docs_containing_terms}$

 else:

$\text{results} = \text{list}(\text{set}(\text{results}).\text{intersection}(\text{docs_containing_terms}))$

$\text{final_results} = [\text{doc} \text{ for } \text{doc} \text{ in } \text{results} \text{ if } \text{results}.\text{count}(\text{doc}) \geq k]$

 return final_result.

Ejercicio 2.

Los componentes principales de un sist. de RI

(1) Obtener colecciones

(2) Construir vocabulario

(3) Resumir docs. (construir vectores).

(4) Construir índice invertido

(5) Procesador de consulta. \rightarrow Se puede subdividir:

 Ordenar resultados por relevancia

 obtener relevancia ej. calcular similitud de vectores.

 Calcular resultados.

Ejercicio 3.

El modelo del espacio vectorial de RI es una técnica que nos ayuda a encontrar documentos relevantes en respuesta a una consulta

Un vector tiene la forma: (termino, peso) donde el termino es alguno de los terminos de la consulta y peso es alguna métrica de relevancia como p.ej: la Frecuencia, el peso en terminos de TF-IDF, etc.

Ejercicio 4.

Un tesoro es un diccionario de sinonimos y antonimos.

Es beneficioso para el RI porque facilita la búsqueda de documentos que están relacionados con los terminos de la búsqueda, no solo con los terminos en si sino con palabras que están relacionadas con esos terminos.

húsaes

Ejercicio 5.

No, en este contexto no tendría sentido usar la Frecuencia como métrica ya que al ser textos pequeños la probabilidad de que aparezcan múltiples veces un mismo término es muy baja (a lo sumo podría aparecer 1 o dos veces).

En cambio la Frecuencia inversa de docs. sí tendría sentido por que en el caso de las tablas podríamos ver en cuantos registros aparece una palabra. Si aparece en pocos registros la relevancia del término será alta, de lo contrario será poco relevante.

Una manera de definir la similitud de dos tuplas usando TF-IDF es con el coseno del ángulo entre los vectores de ambas tuplas.

$$\text{Formula de la Similitud del coseno} = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Ejercicio 6.

Doc 1

Las computadoras pueden aprender de los datos
El aprendizaje automático es un tipo de IA
Los algoritmos de aprendizaje automático predicen resultados

Doc 2

Las computadoras pueden aprender el lenguaje
El procesamiento del lenguaje natural es un campo de IA
Los algoritmos de procesamiento de lenguaje natural analizan texto

Doc 3.

Los ordenadores pueden ver como los humanos
La visión del ordenador es un campo de la IA
Los algoritmos de visión artificial reconocen objetos en imágenes

Doc 4.

Las computadoras pueden tomar decisiones
La inteligencia artificial es un campo de la informática
Los algoritmos de inteligencia artificial resuelven problemas

Doc 5

El deep learning aprende patrones complejos
El deep learning potencia el reconocimiento de imágenes
El deep learning potencia el procesamiento del lenguaje natural

Índice invertido

term	Document, Frecuencia
computadoras	1:1, 2:1, 4:1
IA	1:1, 2:1
lenguaje	2:3, 5:1
algoritmos	1:1, 2:1, 3:1, 4:1

Ejercicio 7.

Para construir el resultado de una consulta:

- crear una lista con los terminos de la consulta.
- Para cada termino, se busca en el indice invertido para encontrar todos los docs en los que el termino aparece
- luego estos docs se ordenan segun algun criterio
- Por ultimo la lista de los docs ordenados se devuelven como el resultado de la consulta.

Si las consultas son basadas en proximidad entonces en el indice invertido no solo se guarda el doc en donde sale el termino sino tambien en que posición el doc se encuentra dicho termino.

near: dos terminos son "ceranos" si estan como mucho a cuatro terminos de distancia.

adjacent: dos terminos están adyacentes si ambos terminos aparecen uno seguido del otro sin ningun termino de por medio

after: un termino está después que otro si ambos terminos aparecen uno seguido del otro con al menos un termino de por medio.

Sintaxis: near(term1,term2), adjacent(term1,term2)
after(term1,term2).

PageRank - EMA

Es algoritmo de ordenamiento de "popularidad" de los sitios web. Básicamente se trata de ordenar los sitios web en función de cuál es más probable de que visites si hacés una caminata aleatoria por la web (estás en una página, entonces decidís irte a otra de forma random o seguir alguno de los links que están puestos en la pág.)

Dado esto, sabemos que cada página tiene un "rango" (lo que abarca, la popularidad, etc). También, sabemos que las págs tienen enlaces hacia afuera y hacia dentro (los 1eros son los links que tengo y los 2dos son los links de otras páginas hacia mí)

Dado esto, entonces, la popularidad de un sitio web es mayor mientras más enlaces hacia dentro tenga. Cada uno de estos tiene un peso, el cual es la "popularidad" del sitio web que tiene ese link. Por ello, un sitio web con más rango que me tiene como link me hace aumentar mucho más mi popularidad que si no abarca mucho.

Además, decimos que mi página tiene un rango alto si la suma de los rangos de las páginas que me apuntan tiene un valor alto

Básicamente:

- Las páginas apuntadas por muchas páginas web es más probable que sean visitadas y van a tener un PageRank más alto
- Las páginas apuntadas por páginas web con alto PageRank van a tener una mayor probabilidad de ser visitadas y entonces van a tener un mayor PageRank

DATAZOS

Retorno de la información

- El **retorno de la información** es el proceso de retornar documentos a partir de una colección de documentos en respuesta a una consulta.
 - Los documentos suelen estar en lenguaje natural no estructurado.

Bases de Datos Relacionales	Sistemas de retorno de la información
Datos estructurados	Datos no estructurados
Dirigidos por esquemas relacionales	No hay esquemas fijos
Modelo de consultas estructurado	Modelos de consulta libres de forma
Operaciones sobre metadatos	Operaciones sobre datos
Las consultas retornan datos	Las búsquedas retornan lista de punteros a documentos.
Los resultados se basan en correspondencia exacta y son siempre correctos.	Los resultados se basan en correspondencia aproximada y medidas de efectividad.
Trabajan con transacciones	No trabajan con transacciones

- La **relevancia** se basa en **factores** como:
 - **Frecuencia de términos:** frecuencia de ocurrencia de término de una consulta en un documento.
 - **Frecuencia inversa de documentos:** ¿en cuántos documentos ocurre la palabra? Si ocurre en menos documentos se le da más importancia a la palabra.
 - **Enlaces a documentos:** Si hay más enlaces a un documento, el documento es más importante.

- **La mayoría de SRI agregan:**

- Las palabras que ocurren en el título, lista de autores, títulos se les da más importancia.
- Las palabras cuya primera ocurrencia es tarde en el documento se les da poca importancia.
- **Proximidad:** si las palabras de una consulta aparecen cerca entre sí en el documento, el documento tiene más importancia que si las palabras ocurren bien lejos unas de otra.

- Otra posibilidad es usar **TF-IDF** (frecuencia de término-frecuencia inversa de documento).
- TF-IDF es usado para evaluar la importancia de una palabra en una colección de documentos.
- **Idea de TF-IDF:** términos que capturan la esencia de un documento aparecen frecuentemente en el mismo; pero para que un término sea bueno para discriminar un documento de los demás, entonces debe ocurrir en unos pocos documentos de la colección.

Índices invertidos

