

ZH megoldások

A csoport

1. Mi a rangsortartás?

Két hasonlósági mérték rangsortartó, ha az általuk számolt találati listák sorrendje megegyezik, bármely keresőkérdés esetén.

Ha két hasonlósági mérték rangsortartó, akkor ekvivalensnek tekintjük, és egymással helyettesíthetjük. Vektortér modellben, általában a hasonlósági mértékekre ez nem jellemző.

2. Mik az információ-visszakeresési folyamat részfolyamatai?

a. Objektumbázis definiálása

- Meghatározzuk a dokumentumok halmazát
- Meghatározzuk a szövegfeldolgozó műveleteket
- Kijelöljük a szövegmodellt

b. Szövegfeldolgozó műveletek alkalmazása

- Dokumentumok átalakítása
- Logikai nézet létrehozása

c. Szöveg indexelése

- Az indexelő modul végzi
- Lehetővé teszi a gyors keresést
- Legtöbbször inverfájlstruktúrában tároljuk

d. Lekérdezés

- A keresőkérdés és az indexek összehasonlítása

e. Rangsorolás

- A kapott adatok alapján
- hasonlósági értékek kiszámítása
- Ezek alapján rangsorolás

f. Találati lista visszaadása

3. Adott egy $Q = (1, 0, 1, 0, 0, 0, 0, 0)$ keresőkérdés és a gyakorisági kifejezés-dokumentum mátrixa (A) egy dokumentum halmazra!

$$A = \begin{pmatrix} 0 & 2 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

a. Adja meg a TD mátriot max-normált, és TF-IDF súlyszámítási sémával is!

Max-normálthoz leosztod az összes oszlopban található értékeket külön külön az oszlopban található legnagyobb értékkel.

A TF-IDF-hez az adott elemet megszorozod azzal, hogy

$$\log\left(\frac{\text{összes dokumentum száma}}{\text{hány dokumentumban fordul elő az adott index}}\right)$$

b. Adja meg a találati listát (relevanciaértékekkel) a dot mértékkel max normált súlyszámítással!

Megvan a maxnormált súlyszámítási sémában elkészült TD mátrix az előző feladatból. Megcsinálod a kérdéssel is ezt, hogy az is valamilyen normával le legyen normálva. Kapsz egy vektort. Mivel dot mértékkel kérdezi, ezért összeszorozod skalárisan a Q-t az összes oszloppal, amelyből értékeket fogsz kapni minden doksira. Ezeket az értékeket csökkenő sorrendbe rendezed és ez a találati lista.

B csoport

1. Mi az index, (definíció, történet), inverz fájlstruktúra?

Az *indexálás* egy olyan eljárás, amely megadja

- egy azonosítónak (szó, kifejezés, név, tárgy, kód)
- egy egységben (könyv, adatbázis) elfoglalt
- pontos helyét (fejezet, oldal, sor)

Az ókori görögök és Rómaiak papirusztekercsre írtak, jegyzeteket készítettek arról, hogy mely fontos rész, hol található a műben. Egy írott dokumentumon könnyebb eligazodni, ha fejezetekre, alfejezetekre tagolódik. Ezt is a római tudósok használták először a módszert.

Legyen az O feldolgozandó objektumok halmaza, és D az objektumoknak megfelelő dokumentumokat.

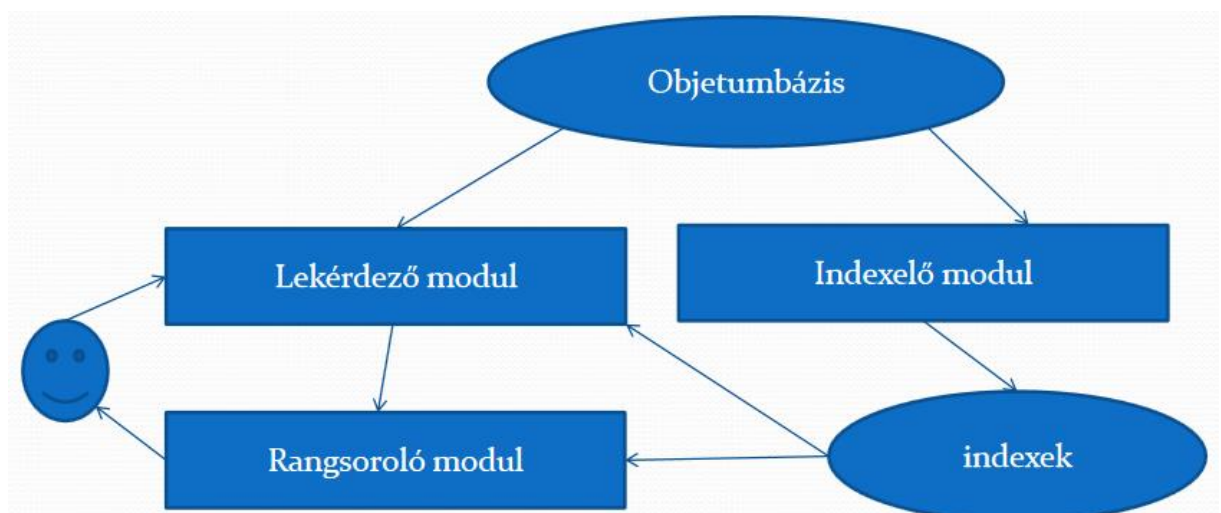
$$O = \{O_1, \dots, O_n\} \quad D = \{D_1, \dots, D_n\}$$

A szövegfeldolgozás első három lépése (lexikai egység azonosítása, stoplista alkalmazása, szótövesítés) után az alábbi index kifejezéseket kapjuk: $T = \{t_1, \dots, t_n\}$ Az indexkifejezések ABC rendezése után megalkotjuk az inverz táblát, amelyben $\forall r_i$ sorban a t_i indexkifejezés van D_j dokumentumokra vonatkozó kódjaival, amely D_j dokumentumokban a t_i indexkifejezés előfordul.

Tehát az indexkifejezések ABC sorrendben vannak, és mellettük azon dokumentumok kódjai, amelyek tartalmazzák az adott kifejezést. Inverz fájlstruktúrának nevezzük az indextáblából, és a hozzá tartozó objektumhalmazból álló struktúrát.

Felhasználva a megfelelő dokumentumok kódjait a megfelelő objektumok visszakereshetők. A struktúrában tárolhatjuk továbbá az előfordulások számát is.

2. Rajzolja le egy általános információ-visszakereső rendszer architektúráját, és írja le röviden a modulok működését.



- **Objektumbázis:**
 - Tárolja a objektumok halmazát, melyen keresünk
 - Objektumok lehetnek: szöveg, kép, hang, multimédia
 - Minden objektum megadható egy szöveges dokumentumként
- **Indexelő modul:**
 - Az objektumbázisban lévő dokumentumokból készít indexet
 - inverz fájl struktúrában tárol
- **Lekérdező modul:**
 - Beolvassa a kérdést, és átalakítja megfelelő formátumra
 - Az indexek segítségével kiválasztja a dokumentumokat, amelyek megfelelő válaszok lehetnek
- **Rangsoroló modul**
 - Kiszámítja a hasonlósági értékeket, majd ezek alapján csökkenő sorrendbe rendezi azokat, és megjeleníti a felhasználónak, mind találati lista.

3. Adott egy Q keresőkérdés az alábbi dokumentumhalmazra az alábbi indexkifejezésekkel!

Terms		Documents	
T1	baby	D1	<u>Infant and Toddler First Aid</u>
T2	child	D2	<u>Baby and Child's Room (For Your Home)</u>
T3	guide	D3	Your <u>Baby's Health</u> and <u>Safety</u> : From <u>Infant</u> to <u>Toddler</u>
T4	health	D4	<u>Baby Proofing</u> Basics
T5	home	D5	Your <u>Guide</u> to Easy Rust <u>Proofing</u>
T6	infant	D6	Beanie <u>Baby</u> <u>Collectors</u> <u>Guide</u>
T7	proofing		
T8	safety		
T9	toddler		

a) Adja meg a TD mátriot hossz-normált, és TF_IDF súlyszámítási technikával is!

Megfogod és ugyanúgy, mint az A csoport ZH-jában, csak neked kell felírni a mátrixot.

Btw itt hossznormáltat számolsz, tehát nem a legnagyobb elemmel osztasz le, hanem a hosszal.