T-61.5020 Luonnollisen kielen tilastollinen käsittely

Harjoitus 4, ke 14.2.2007, 12:15–14:00 — Tiedonhaku
Versio 1.0

1. Tietokannassa on 10 000 dokumenttia. Käyttäjä suorittaa haun, johon kannasta oikesti löytyy 6 relevanttia vastausta. Kaksi kilpailevaa hakukonetta pulauttavat kymmenen dokumentin listan, jossa relevantit vastaukset ovat sijoittuneet taulukon mukaan.

   Tarkastele hakukoneiden hyvyyttä allaolevan listan mittojen avulla. Listan suomennoksia ei pidä ottaa vakavasti.

   - tarkkuus *(precision)*
   - saanti *(recall)*
   - hajoama *(fallout)*
   - täsmäävyys *(accuracy)*
   - virhe *(error)*
   - F-mitta *(F-measure)*
   - Interpoloimaton keskitarkkuus *(Uninterpolated average precision)*.

Dokumentti	relevanssi	
	kone 1	kone 2
d1	+	+
d2	+	+
d3	-	+
d4	+	-
d5	-	+
d6	-	-
d7	-	+
d8	-	-
d9	-	+
d10	+	-

2. Sana w_1 on esiintynyt 21 dokumentissa ja sana w_2 500 dokumentissa 10 000 dokumentin kokoelmassa. Sana w_1 esiintyy yhteensä 101 kertaa näissä dokumenteissa ja sana w_2 700 kertaa. Kuinka paljon näitä sanoja kannattaa painottaa hakukoneessa? Kokeile painotuksena käänteistä dokumenttifrekvenssiä *(IDF, Inverse Document Frequency)* ja residuaalista käänteistä dokumenttifrekvenssiä *(RIDF, Residual Inverse Document Frequency)*.

3. Tehtävässä käytetään aineistona seuraavia tekaistuja uutisotsikoita:

   d_1: Moskovan radan tulevaisuus vaakalaudalla — formulat jatkavat Itävallassa.

   d_2: Schumacher itsevarma: Sepangin rata sopii Ferrarille.

   d_3: Tähtien kolari — Hungaroringin radalle jäi formuloista 35 miljoonan euron arvosta romua.

   d_4: 5 formulaa romuna, Schumacher syyttää kolarista Coulthardia.

   d_5: Meteoriittien jäljillä galaxin alkuperää etsimässä.

   d_6: Galaksissamme suistuu vuosittain kymmeniä planeettoja tähtiä kiertäviltä radoiltaan.

   d_7: Tähti olikin planeetta.

Tee dokumentti-sanamatriisi ainoastaan sanoille Schumacher, rata, formula, kolari, galaksi, tähti, planeetta ja meteori. Käytössäsi on sanat perusmuotoistava laite. Miten dokumenttien samankaltaisuudet muuttuvat, kun SVD-hajotelman avulla tiputetaan dokumenttivektorien dimensio kahteen? Tarkastele erityisesti dokumenttien d_5 ja d_7 välistä korrelaatiota.