

TDT4117 - Øving 1

Ole-Christer Selvik, Håkon Løvdal og Kristoffer Andreas Dalby

Today

1 Oppgave 1

1.1 Inverted index

Inverted index er en måte å strukturere indeksert data ved å lage en oversikt som peker fra for eksempel et ord til en lokasjon i en database eller et dokument. Dette er ganske konkret motsatt praksis av indeksering, hvor det er svært vanlig å heller peke fra en nøkkel til en verdi, derav invertert. Det er to hovedvarianter av Inverted index, record level og word level. Hovedforskjellen på disse er at record level kun holder lister for hvilket dokument et ord befinner seg i. Mens word level holder styr på både hvilke dokument ordet befinner seg i og hvilken lokasjon det har i dokumentet. Word level har på grunn av dette mer funksjonalitet, men krever mer tid og plass for å lages. En funksjon som tilbys er blant annet muligheten til å søke på grunnlag av fraser eller sammensetninger av verdier.

1.2 Data Retrieval

Data retrieval eller data gjenfinning er prosessen for å hente ut data fra en database. Det involverer å hente ut ønsket og spesifikk data. Man vet at dataen er der og gjør dette med en spesifikk spørring som har et sett med kriterier. Som regel er dette gjort mot en DBMS(Database Management System) eller et database kontroll verktøy som returnerer data basert på din spørring. Spørringen gjøres gjerne med et Query språk som SQL som er et mye brukt standardisert språk til formålet.

1.3 Structured data

Strukturert data er enkelt og greit data som er strukturert. En annen måte å forklare det på er at dataen er satt i et system som gjør at det er mulig å identifisere og hente frem dataen igjen enkelt. Den vanligste måten å strukturere data på er ved å bruke en database.

2 Oppgave 2

2.1 Term Frequency

Term frequency er et konsept hvor man rangerer dokumenter etter en numerisk verdi man regner ut basert på hvor mange ganger en frase dukker opp i dokumentet. Det er vanlig å først eliminere alle dokumenter som ikke inneholder alle ordene i frasen man søker etter, for så deretter å telle antall ganger ord fra frasen dukker opp i dokumentet.

2.2 Inverse Document Frequency

Inverse Document Frequency er en slags addisjon til TF som ordner opp i problemet med at noen ord brukes veldig mye men ikke nødvendigvis gir noen verdi til søket og kan også virke negativt. Dette er tynnsik ord som for eksempel en, et, og, i og å, da disse ordene har en tendens til å forekomme mange ganger i dokumenter. Derfor i IDF blir det lagt inn faktorer som rydder unna ord som forekommer veldig ofte.

2.3 Meningen med TF-IDF

Formålet med TF-IDF som kombinasjon er å ha en veldig simpel rangeringsfunksjon som kan enkelt regne ut hvilket dokument som kan være mest aktuelt basert på en frase, uten å bli påvirket av ord som forekommer for mange ganger. Man kan også bruke TF-IDF veldig simpelt og veldig avansert, hvor en av de simple måtene er å enkelt og greit kun summere TF-IDF verdien for hver term i spørringen/frasen. De fleste avanserte metodene er basert på den enkle.

3 Oppgave 4

Dokumentsett D:

doc1 = NTNU
 doc2 = NTNU, Computer
 doc3 = Trondheim
 doc4 = Trondheim, NTNU, Computer, NTNU, Trondheim
 doc5 = NTNU, NTNU, NTNU, Computer
 doc6 = Computer
 doc7 = Trondheim, NTNU, NTNU, Trondheim, Trondheim

tabellen under viser term frekvensen $tf_{(t,d \in D)}$ for hvert av dokumentene i dokumentsettet D, hvor $t \in T$.

term-frequencies for document list a terms:							
Term	$tf_{(t,1)}$	$tf_{(t,2)}$	$tf_{(t,3)}$	$tf_{(t,4)}$	$tf_{(t,5)}$	$tf_{(t,6)}$	$tf_{(t,7)}$
Trondheim	-	-	1	2	-	-	3
NTNU	1	1	-	2	3	-	2
Computer	-	1	-	1	1	1	-

Inverse document frequency: $idf(t, D) = \lg(\frac{|D|}{t \in D})$

$$idf(Trondheim, D) = \lg(\frac{7}{3}) = 0,368$$

$$idf(NTNU, D) = \lg(\frac{7}{5}) = 0,146$$

$$idf(Computer, D) = \lg(\frac{7}{4}) = 0,243$$

$$tf - idf_{(t,d,D)} = tf_{(t,d)} \times idf_{(t,D)}$$

term frequency-inverse document frequency:

Term	d_1	d_2	d_3	d_4	d_5	d_6	d_7
Trondheim	-	-	0,367	0,735	-	-	1,104
NTNU	0,146	0,146	-	0,292	0,438	-	0,292
Computer	-	0,243	-	0,243	0,243	0,243	-

dokumenter som skal sammenlignes med $\vec{d_2} = \{0, 1, 1\}$:

$$\vec{d_1} = \{0, 1, 0\}$$

$$\vec{d_4} = \{2, 2, 1\}$$

$$\vec{d_5} = \{0, 3, 1\}$$

$$\vec{d_6} = \{0, 0, 1\}$$

$$\vec{d_7} = \{3, 2, 1\}$$

Euclidean distance:

$$D_{(\vec{d_1}, \vec{d_2})} = 1 - \frac{0 \times 0 + 1 \times 1 + 0 \times 1}{\sqrt{1^2} \times \sqrt{1^2 + 1^2}} = 1 - \frac{1}{\sqrt{1} \times \sqrt{2}} = 0,2928$$

$$D_{(\vec{d_4}, \vec{d_2})} = 1 - \frac{2 \times 0 + 2 \times 1 + 1 \times 1}{\sqrt{2^2 + 2^2 + 1^2} \times \sqrt{1^2 + 1^2}} = 1 - \frac{3}{\sqrt{5} \times \sqrt{2}} = 0,0513$$

$$D_{(\vec{d_5}, \vec{d_2})} = 1 - \frac{0 \times 0 + 3 \times 1 + 1 \times 1}{\sqrt{3^2 + 1^2} \times \sqrt{1^2 + 1^2}} = 1 - \frac{4}{\sqrt{10} \times \sqrt{2}} = 0,1055$$

$$D_{(\vec{d_6}, \vec{d_2})} = 1 - \frac{0 \times 0 + 0 \times 1 + 1 \times 1}{\sqrt{1^2} \times \sqrt{1^2 + 1^2}} = 1 - \frac{1}{\sqrt{1} \times \sqrt{2}} = 0,2928$$

$$D_{(\vec{d_7}, \vec{d_2})} = 1 - \frac{3 \times 0 + 2 \times 1 + 1 \times 1}{\sqrt{3^2 + 2^2 + 1^2} \times \sqrt{1^2 + 1^2}} = 1 - \frac{3}{\sqrt{14} \times \sqrt{2}} = 0,4330$$

Dokument rangeringen fra størst til minst likhet:

$$d_4, d_5, d_1, d_6, d_7$$

merk at d_1, d_6 rangeres helt likt.