Obligatorisk oppgave 3 INFO 221

Espen Kleivane student nummer 154542

Vårsemester: 04.05.12 ved u.i.b

1 a)

Ut fra et enkelt IR system blir SP, FP, FN og SN forklart:

- Det som blir levert tilbake og er relevant. Dette er de sanne positive (SP).
- Det som ikke er relevant, og ikke blir levert tilbake. Dette er de sanne negative (SN).
- Det som er relevant men ikke blir levert tilbake. Dette er de falsk negative (FN).
- Det som ikke er relevant men blir levert tilbake. . Dette er de falsk positive (FP).

Dokument	Positiv	negativ
Med relevans	SP	FN
Uten relevans	FP	SN

(Manning et all s 143)

b)

To dokumenter d1 og d2, for informasjonsbehovet er bare d1 relevant.

To spørringer q1 og q2.

Søker med q1 og får tilbake begge dokumentene, da vil d1 være sann positiv (SP) og d2 vil vare falsk positiv (FP)

c)

Søker med q2 og får ingen dokumenter tilbake, da vil d1 være falsk negativ (FN) og d2 vil sann negativ (SN)

2 a)

Presisjon i henhold til IR systemer er relevante dokumenter returnert delt på alle som ble returnert.

$$\text{Precision} = \frac{\#(\text{relevante dokumenter returnert})}{\#(\text{returnerte dokumenter})} = P(\text{relevant/returnert})$$

Recall i henhold til IR systemer er relevante dokumenter returnert delt på alle relevante dokumenter som er i systemet og eller databasen

Recall = .#(relevant dokumenter returnert) = P(returnert/relevant)

#(relevante dokumenter totalt)

(Manning et all s143)

b)

Test data til utvikling / justering og et annet data sett til å måle hvor bra systemet er, hvorfor?

Mange systemer inneholder ulike vekter (parametre) som kan justeres for å finjustere systemet ytelse. Det er galt å rapportere resultater på en test samling som ble innhentet ved å justere disse parameterne for å maksimere ytelsen på denne samlingen. Det er fordi en slik justering overdriver den forventede ytelsen til systemet, fordi vektene vil bli satt til en maksimal ytelse på ett bestemt sett av spørringer snarere enn for et tilfeldig utvalg av spørringer. I henhold til dette er riktig fremgangsmåte å ha ett eller flere ”development test collections”, for å justere parameterene på denne test - samlingen (DTC). Testeren kjører deretter systemet med de vektene som var på den nye test - samlingen og rapporterer resultatene på at samlingen som et upartisk estimat av ytelse (manning s 141)

c)

Når man søker etter noe for å fylle et informasjonsbehov er det viktig å snakke om resultater og anslå hvor relevant, eller ikke relevant resultatet ble etter utført søk. Ordene ”precision & recall” blir brukt for å beskrive kvaliteten av informasjonsigjenfinning av søket som er utført i henhold til dokumenter. ”Precision is a ratio of the number of relevant documents (references) in the result over the total number of documents returned […..] Recall is defined as the number of relevant documents returned, compared to the number of relevant documents in the database”[Olsen: s 125]. Med dagens webteknologi har universelle søkemotorer som google, yahoo o.l mange datamaskiner som jobber sammen for og indekserer hele internett (eller deler av det) slikt at det skal være raskt å søke i. Hvis man bare har et søkefelt og ikke noe videre innsikt i systemet, vil det i denne oppgaven være litt vanskelig å måle ”precision & recall” i nøyaktighet. Hvis man har søkt med en term(er) og fått tilbake noe som var relevant for informasjonsbehovet eller ikke relevant, kan man til en viss grad anslå nøyaktigheten, som f. eks: nøyaktig, mindre nøyaktig, ikke nøyaktig. Dette blir da vage ord for nøyaktighet. Når det gjelder gjennkalling blir det problematisk i og med at man ikke vet hva som er i databasen.

3a)

Vektor for dokumentene blir

V1 = (1,1,1,1,1,1,1,1,1) for dokument 1

V2 = (1,1,1,1,1,1,1,1,1) for dokument 2, den samme samlingen med ord finnes i begge dokumentene.

b)

Når det gjelder ord – semantikk er d1 og d2 like, men når det gjelder setnings- semantikk er de helt motsatt av hverandre. Modellen som ikke tar hensyn til slike forskjeller som i dokument d1 og d2 heter ”Bag of words modell”(Manning et all s107). Denne modellen tar ikke hensyn til orden på rekkefølgen av ordene men antall forekomster av hvert ord (term) er vesentlig.

c)

Bilde 2*2 segmenter:

VF for bilde a:

VF = (r(93), g(135), b(170)) celle 1.1

(r(185), g(154), b(153)) celle 1.2

(r(185), g(212), b(217)) celle 2.1

(r(185), g(66), b(217)) celle 2.2

R: 648 / 4 = 162

G: 567 / 4 = 141.7

B: 757 / 4 = 189.2

VF = (r(162), g(142), b(189))

VF for bilde b:

VF = (r(185), g(239), b(153) celle 1.1

(r(185), g(66), b(144) celle 1.2

(r(105), g(66), b(118) celle 2.1

(r(35), g(163), b(43)) celle 2.2

R: 510 / 4 = 127.5

G: 534 / 4 = 133.5

B: 458 / 4 = 114.5

VF =(r,(128), g(134), b(115))

Vektor for bilde a vil holde de tre koordinatene VF = (r(162), g(142), b(189)), vektor for bilde b vil holde de tre koordinatene VF =(r,(128), g(134), b(115)).

d)

Ikke gjort denne

4)

Se vedlagt program i java (KO3_INFO221) forklaring til programmet ligger i ReadMe filen.

5 a)

Dokument frekvensen til ”knowledge” er 14, dokument frekvensen til ”truth” er 6.

Idf for ”knowledge” er log(137 dokumenter/14 dokument forekomster) = 0.990592531.

Idf for ”truth” er log(137 dokumenter / 6 dokument forekomster) = 1.35856932.

b)

Her ble det brukt database - tabellen documents2, hvor det ble utført en spørring i sql (SELECT ID FROM documents2 WHERE Doc_Name=” Anaxagoras.txt”) med navnet Anaxagoras.txt og fikk tilbake id nummer 6.

Deretter ble funksjonen (metoden i KO3_INFO221) Finn tf –idf med input ”knowledge” i dokument 6 som resulterte i output: 2.9717775944345064.

Samme prosedyre ble brukt for termen ”truth” som resulterte i output: 1.358569316772763.

Kilder:

Olsen. K.A ”The internet, the web and eBusiness” Formalizing Applications for the Real World. The Scarecrow Press, Inc Lanham, Maryland. Toronto. Oxford 2005

Manning, D, C. Raghavan, P. Schutze, H. “Introduction to information Retrieval”. Cambridge university press. 2008.