NTNU
Det skapende universitet

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4117 Informasjonsgjenfinning

Faglig kontakt under eksamen:

Svein Erik Bratsberg: 9953 9963

Eksamensdato: 1. desember 2014

Eksamenstid (fra-til): 09:00 - 13:00

Hjelpemiddelkode/Tillatte hjelpemidler:

D – Ingen trykte eller håndskrevne hjelpemidler tillatt. Bestemt, enkel kalkulator tillatt.

Annen informasjon:

Målform/språk: Norsk bokmål

Antall sider: 4

Antall sider vedlegg: 0

Kontrollert av:

Robert Neumayer (sign.)

_____

Dato Sign.

Oppgave 1: Generelle spørsmål (20%)

a Datagjenfinning (data retrieval) vs. informasjonsgjenfinning (IR):

- IR-system: Forklar i korthet de viktigste begrepene.
- Forklar hovedforskjellene mellom datagjenfinning og informasjonsgjenfinning med hensyn til "matching" og "content".

b IR-sytemarkitektur:

- Forklar i korthet input og output fra "query expansion"-steget i et IR-system. Hva er fordelene med dette?
- Hva er hensikten med indekseringen og hvorfor trenges det en indeks i et IR-system?

c Forklar i korthet begrepene "user need" og "query", og forklar sammenhengen mellom de to begrepene i et IR-system.

Oppgave 2: Multimedia IR (20%)

a Forklar i korthet hva "semantic gap" betyr i multimedia-IR.

b Hva er hovedforskjellene mellom tekstbasert IR og multimedia-IR ("content-based IR")?

c Fargehistogrammer:

- Beskriv fordeler og ulemper ved å bruke fargehistogrammer som en egenskap ("feature")
- Forklar stegene involvert ved bruk av fargehistogrammer i et "image retrieval system".

Oppgave 3: IR-evaluering (20%)

Anta du får et rangert ("ranked") resultat fra et eksisterende IR-system. Nedenfor viser vi: Rank, document id, score av topp-10-resultater.

Rank	Document ID	Score
1	24	.8
2	38	.64
3	40	.62
4	28	.4
5	36	.32
6	48	.3
7	22	.28
8	44	.26
9	32	.1
10	60	.05

I tillegg vet vi at dokumentene med de følgende document-id'ene er relevante for queryet: 38, 40, 28, 22, 44, 80, 92.

a Forklar begrepene "recall" og "precision" basert på tabellen over og de oppgitte relevante dokumentene. Beregn "precision" og "recall" for alle nivåer for det gitte eksempelet.

b Forklar rollene til "precision" og "recall" i Web-søk. Hva er problemene med "precison/recall" her? Hvilket mål ("measure") er mest interessant for brukere av Websøkemotorer?

c Forklar og beregn MAP for det gitte eksempelet. Hva er fordelen med MAP framfor "precision"/"recall"?

"Mean average precision" er gitt ved:

$$MAP_i = \frac{1}{|R_i|} \sum_{k=1}^{R_i} P(R_i[k])$$

Oppgave 4: HITS-algoritmen (20%)

a Oppgi minst 3 forskjeller mellom HITS- og PageRank-algoritmene.

b Gitt grafen under og beregn *hub*- og *authority*-scores for websidene merket A, B, C og D ved å bruke HITS-algoritmen. Utfør minst 3 iterasjoner av algoritmen og vis dine utregninger ved å vise formler fyllt ut med verdier for minst en av iterasjonene. For enkelhets skyld, ignorer normaliseringsteget.

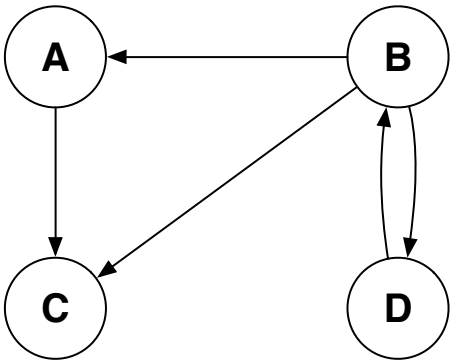

Figure 1: Graf av websider forbundet med lenker.

Oppgave 5: Modellering (20%)

Gitt den følgende dokumentsamlingen (5 dokumenter, ett dokument per linje)

```
Norwegian University of Science and Technology is a university in Trondheim.

The University of Oslo is a university in Norway.

The University of Bergen is in Bergen, Norway.

The University of Tromso is the northernmost university.

The University of Stavanger is a university in Stavanger, Norway.
```

Etter å ha laget små bokstaver av all tekst, fjernet alle "."- og ","-tegn, delt ord ved "space", og fjernet stoppord, så får vi følgende "tokens":

```
'university', 'bergen', 'northernmost', 'norway', 'stavanger',
'technology', 'norwegian', 'trondheim', 'oslo', 'science', 'tromso'
```

a Beregn dokumentfrekvensene for alle gitte "tokens"

b Beregn termfrekvensene for alle tokens i alle dokumenter

Ranger alle dokumenter ved relevans og deres relevansscore for det følgende query: "**science norway**"

c ... ved bruk av "boolean retrieval"

d ... ved bruk av "vector space retrieval", først ved å vise term-vektorene for alle dokumenter og så ved å bruke følgende ligning for både dokument- og query-term-vekting[1]:

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & if\ f_{i,j} > 0 \\ 0 & otherwise \end{cases}$$

Bruk av "cosine similarity for similarity calculations" mellom query- og dokument-vektorer:

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}||\vec{q}|} = \frac{\sum w_{i,j} w_{i,q}}{\sqrt{\sum w_{i,j}^2}\sqrt{\sum w_{i,q}^2}}$$

e Gitt metodene brukt i oppgave c) og d), forklar de følgende begrepene: "document normalisation," "document component," og "term component." Forklar hvordan disse begrepene er integrert i de forskjellige rammeverkene og hvorfor de er viktige for gjenfinningen. Vis hvordan begrepene er realisert i ligningene over.

[1] Valg av base for logaritmen er uten betydning for rangeringen, men vi anbefaler bruk av log2.