NTNU
Det skapande universitetet

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4117 Informasjonsgjenfinning

LØSNINGFORSLAG/Sensurveiledning

Faglig kontakt under eksamen: Heri Ramampiaro
Tlf.: 73591459 / 99027656

Eksamensdato: 18 . desember 2013
Eksamenstid (fra-til): 09.00-13.00
Hjelpemiddelkode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrevne hjelpemiddel tillatt. Bestemt, enkel kalkulator tillatt.

Annen informasjon:

Målform/språk: Bokmål
Antall sider: 5
Antall sider vedlegg: 0

Kontrollert av:

_____

Dato Sign

Det ønskes korte og konsise svar på hver av oppgavene.
Les oppgaveteksten meget nøye og vurder hva det spørres etter i hver enkelt oppgave.
Begrunn svar på alle oppgaver.

Dersom du mener at opplysninger mangler i oppgaveformuleringene, beskriv de antagelsene du gjør.

Merk at hvor mye hvert delspørsmål teller kan justeres ved sensur.

Oppgave I - Dokumentforberedelser og indeksering (30 %)

1. Det er uenighet om hvorvidt stemming er nyttig eller ikke. Forklar når **stemming** er nyttig og når det ikke er det.

   Svar: Stemming er nyttig når man dokumentsamlingen ikke er så stor (som i Web-størrelse) og recall er både viktig og mulig å beregne. Stemming er ikke så nyttig for store dokumenter hvor man har lite fokus på recall, men samtidig mest fokus på precision. Generelle forklaringer som er relatert til språk (feks. Norsk vs. Engelsk) kan godtas men skal ikke gi full pott. (6%)

2. Begrunn hvorfor søk i *Gulesider* eller *Yelp* både kan være datagjenfinning og informasjonsgjenfinning.

   Svar: Gulesider eller Yelp kan ha alle egenskapene med IR: delvis match, rangering av treff, og feiltoleranse. Det er også mulig å søke på eksakt match hvis man fokuserer kun på å finne et navn og adresse. (5%)

3. I dette faget har vi gått gjennom flere forskjellige indekseringsmetoder. Ta for deg fire av disse metodene og fortell hvilke(n) metode(r) egner seg for:

   a. Store dokumentsamlinger.

      Svar: Invertert index - effektiv og lavt krav til lagring, Supra Index utnytter både egensakpene til suffix arrays samt blokkadressering som gir mindre disk akssess en suffix arrays.

   b. Små dokumentsamlinger.

      Svar: Signatur fil - kan gå tom for hashkode for store dokumenter, men er veldig effektiv for små samlinger. Suffix tree - høy overhead men effektiv for små dok.samlinger.

   Begrunnet svar er eneste som gir full pott. (8%)

4. Valg av indekstermer kan gjøres enten automatisk eller manuelt/bruk av en ekspert. Drøft fordelene og ulempene ved å bruke **eksperter** til å velge indekstermer manuelt.

   Svar: Fordeler med eksperter: domenekunnskap som kan være nyttig for å vite mer nøyaktig hvilke ord som bør indekseres. Her er det også mulig for en ekspert å vite hvordan ord kan vektes. Ulemper: Subjektivitet, mangel på effektivitet, og kostnaden ved håndtering av store mengder av informasjon. (5%)

5. Konstruer "**Supra index**" basert på følgende tekst. Gjør de antakelsene du finner nødvendig. "`Mandela was a widely known person. He was an important person for ANC`".

   Svar: Antar at vi velge en k=4 på suffix-tegnlengde og 2 suffix array pr. blokk da får vi:

Lage suffix strengen først for å finne ordene. Anta at vi ikke trenger å ta med adverb og artikler (stoppord)

1: Mandela was a widely known person. He was an important person for ANC

22: known person. He was an important person for ANC

28: person. He was an important person for ANC

56: person for ANC

67: ANC

Vi får da følgende Supra Index-struktur:

(6%)

Oppgave II - Similaritetsmodell og Websøk (30%)

1. Sammenlign similaritetsmodellene **Språkmodellen** *("The language Model")* og **Sannsynlighetmodellen** *("The Probabilistic modell")*. Du skal fokusere på **prinsippene** i forklaringen din.

   Svar: Både Språkmodellen (LM) og sannsynlighetsmodellen (PM) baserer beregning av likhet/ similaritet på sannsynlighet. Hovedforskjellen er at LM baserer dette på en antakelse på at det finnes en modell for hvert dokument. Likhet mellom et dok. og spørring bergnes dermed basert på sannsynligheten for at dokumentetsmodell genererer spørringen. PM baserer seg derimot på intuisjonen om relevans. Dvs. likhet mellom dokument og spørring bergnes ut fra sannsynliget for at dokumentet er relevant. (7%)

2. Vector Space Model (VSM) er en ofte brukt similaritetsmodell. **TF-IDF** er et viktig element i denne modellen. Forklar hva TF-IDF er. Forklar spesielt hvorfor **IDF** er viktig for denne modellen.

   Svar: Hovedforklaringen skal fokusere på definisjonen av TF (Term Frequency) som er basert på hvor ofte en term er nevnt i et dokument og IDF (invers document frequency) som er log(N/ df), som er basert på hvor ofte termen er nevnt i hele samlingen. TF-IDF vil i hovedsak fortelle om hvor viktig en term er samt hvor unik den er. Høy IDF kombinert med høy TF vil dermed bidra til høy ***diskrimineringsgrad***. Det er derfor IDF er viktig for VSM. (9%)

3. Forklar fordelene med det distribuerte websøkesystemet sammenliknet med det sentraliserte crawler-baserte systemet. Fokuser forklaringen din på det som går utover fordelene med generelle distribuerte systemer.

   Svar: Fokuset her er Haverst-systemet som består av Brokers og Gatherers. For å få full pott er det viktig at studenten forklarer hvordan kombinasjon av Brokers og Gatherers kan bidra til

mer effektiv innsamling av informasjon om websider til indeksering osv. Det er også et viktig poeng at studenten får med seg at brokere kan gjøres domenespesifikke. (9%)

4. I likhet med generelle informasjongjenfinningsystemer bruker Websøkemotorer også modeller til å rangere søkeresultater. Hva heter minst tre av disse?

   Svar: Her trenger studentene i utgangspunktet bare nevne navn, men forklaring gir et pluss dersom de ikke husker som mange som 3. Eksempler her er: HITS, PageRank, MostCited, Vector-Spread, Bolean-Spread. OKAPI BM25 eller bare vector space modellen gir ikke poeng. (5%)

Oppgave III - Evaluering av søkeresultater (20%)

1. Drøft hvorfor en god evaluering av et IR-system er viktig.

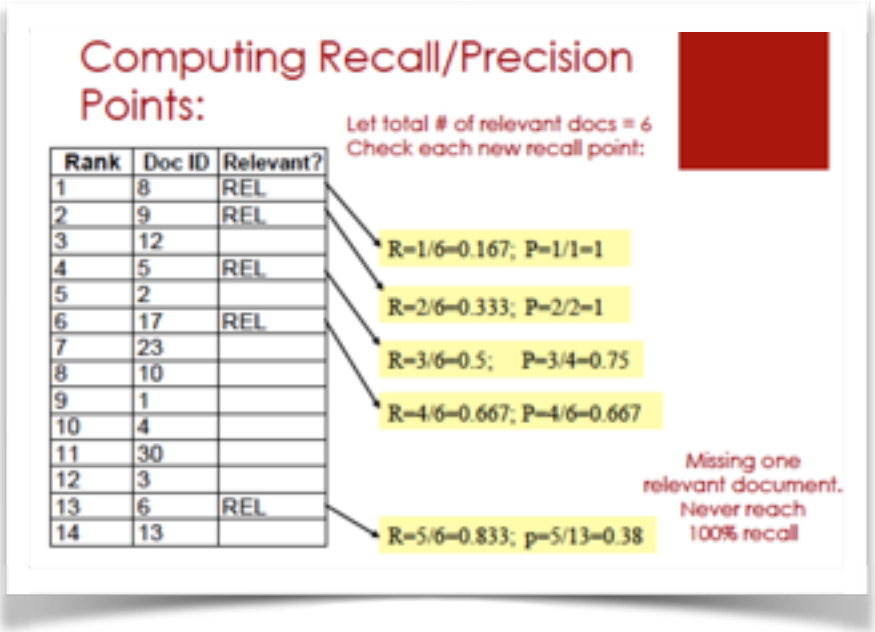
   Svar: Studentene forventes å bygge forklaringen her med hvordan eksisterende mål og/eller metoder for evaluering av IR systemer kan bidra til å få innsikt i hvor godt IR-systemet er sammenliknet med andre systemer og i forhold til å kunne levere på effektivitet (dvs. retrieval performance). For å få full pott behøves det mer enn bare generelle betraktninger om evaluering, dvs. studentene forventes å kunne si noe om precision og recall og andre mål for å illustrere viktighet. (5%)

2. Til å evaluere et IR-system brukes ofte forskjellige mål. Et mål som ofte blir brukt er mean average precision (MAP). Forklar hvordan MAP beregnes.

   Svar: Her kan studentene enten baser seg på MAP-formelen eller forklare den i ren tekst. Viktigst er at de forstår hvordan den fungerer. MAP beregnes ved først å beregne snittpresisjon (Average Presision) for hver spørring og snittet av dette igjen for alle spørringer. Snittpresisjon beregnes her ved å regne ut persisjonsverdiene for hvert relevant treff og finne gjennomsnittsverdien av dette. (7%)

3. Lag et eksempel som illustrerer hvordan man bergner precision-punkter. Tips: Beregn både precision- og recall-verdiene. Gjør de antakelsene du finner nødvendig.

   Svar: Her er et eksempel tatt fra en av forelesningen. For å få full pott må studenten vise at han/hun vet hvordan recall og presisjon bergnes og at punktene bergnes **kun** der relevante dok. er hentet. (8%)



Oppgave IV - Diverse (20%)

Svar ***rett/galt*** med ***begrunnelse*** på følgende utsagn. Hvert **riktig** og **begrunnet** svar får **2** poeng. **Feilsvar** får **-1** poeng. Mens **ubegrunnet** eller **ingen svar** er **0** poeng.

Poengivning: Hvert **riktig** og **begrunnet** svar får **2** poeng. **Feilsvar** får **-1** poeng dersom både svaret og begrunnelsen er helt feil. **Ubegrunnet** eller **ingen svar** er **0** poeng.

1. ”F-measure” eller ”Harmonic Means” er en god måte å kombinere precision or recall på. (RETT/GALT)

   Svar: RETT - Denne gir oss muligheten til å evaluere et IR-system ved å kombinere precision og recall i samme mål.

2. Thesaurus er et verktøy til å utvide spørringer og brukes ofte i forbindelse med “automatic global analysis”. (RETT/GALT)

   Svar: (RETT) - AGA bruker thesaurus som feks. Statistical Thesaurus til å utvide spørringer.

3. “Signature Files” er en metode for å signere et informasjonsgjenfinningsdokument på. (RETT/GALT)

   Svar: (GALT) - Signature Files er en indekseringsmetode.

4. Multimedia informasjonsgjenfinning er ofte enklere enn tekstegjenfinning da man ikke trenger å utføre tekstoperasjoner. (RETT/GALT)

   Svar: (GALT) - Multimedia objekter er mer komplekse enn tekst.

5. Micon er en viktig feature for bilder og brukes i bildegjenfinning, og kan sees på som en analogi av r-frames innen videogjenfinning. (RETT/GALT)

   Svar: (GALT) - Micon (står for motion icon) brukes kun til videogjenfinning.

6. User Relevance Feedback (URF) er ofte brukt til å redefinere spørringer slik at man får økt søkehastighet. (RETT/GALT)

   Svar: (GALT) - URF brukes til å redefinere spørringer for å forbedre spørringene slik at man får flere relevante treff.

7. R-Precision er en forkortelse på Recall-Precision. (RETT/GALT)

   Svar: (GALT) - R-precision er precision på R-te posisjon i resultatlista, hvor R er total antall relevante dokument.

8. MRR (Mean Reciprocal **Rank**) er veldig godt egnet til evaluere systemer der man mest er opptatt av å finne relevante resultatet i en topp-k (feks. topp-10) resultatliste. (RETT/GALT)

   Svar: (RETT) - MRR brukes til å evaluere hvor i top-k lister man finner det første relevante dokumentet. Derfor er det godt egnet til å evaluere systemer der man mest er opptatt av å finne relevante resultater i en topp-k resutltatliste.

9. “Vocabulary Trie” og “Suffix Trie” er to begrep som beskriver samme indekseringsmetode. (RETT/GALT)

   Svar: (GALT) - “Vocabulary Trie” brukes som datastruktur for invertert index mens “Suffix Trie” er en indekseringsmetode.

10. Fjerning av stoppord kan ha negative påvirkninger på Recall. (RETT/GALT)

    Svar: (RETT) - Fjerning av stoppord kan redusere sjansene for å finne spesifikke uttrykk som “to be or not to be” og “the Who” etc, i disse tilfellene er det en ulempe med fjerning av stoppord.