NTNU
Det skapande universitetet

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4117 Informasjonsgjenfinning

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: 3
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.
2. Begrunn hvorfor søk i *Gulesider* eller *Yelp* både kan være datagjenfinning og informasjonsgjenfinning.
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.
    b. Små dokumentsamlinger.

    NB: Husk å begrunne svarene dine.
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.
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`".

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

Oppgave III - Evaluering av søkeresultater (20%)

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

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

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.

1. ”F-measure” eller ”Harmonic Means” er en god måte å kombinere precision or recall på. (RETT/GALT)
2. Thesaurus er et verktøy til å utvide spørringer og brukes ofte i forbindelse med “automatic global analysis”. (RETT/GALT)
3. “Signature Files” er en metode for å signere et informasjonsgjenfinningsdokument på. (RETT/GALT)
4. Multimedia informasjonsgjenfinning er ofte enklere enn tekstegjenfinning da man ikke trenger å utføre tekstoperasjoner. (RETT/GALT)
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)
6. User Relevance Feedback (URF) er ofte brukt til å redefinere spørringer slik at man får økt søkehastighet. (RETT/GALT)
7. R-Precision er en forkortelse på Recall-Precision. (RETT/GALT)
8. MRR (Mean Reciprocal Mean) er veldig godt egnet til evaluere systemer der man mest er opptatt av å finne relevante resultater i en topp-k (feks. topp-10) resultatliste. (RETT/GALT)
9. “Vocabulary Trie” og “Suffix Trie” er to begrep som beskriver samme indekseringsmetode. (RETT/GALT)
10. Fjerning av stoppord kan ha negative påvirkninger på Recall. (RETT/GALT)