1NTNU
Norges teknisk-naturvitenskapelige universitet

Fakultet for informasjonsteknologi, matematikk og elektroteknikk

Institutt for datateknikk og informasjonsvitenskap

Faglig kontakt under eksamen:
Institutt for datateknikk og informasjonsvitenskap
Heri Ramampiaro, 73593440

EKSAMEN I EMNE TDT4117 INFORMASJONSGJENFINNING

Fredag 15. aug. 2011.
Tid: kl 09.00 – 13.00 (4 timer)

BOKMÅL

Hjelpemidler: D – Ingen trykte eller håndskrevne tillatt. **Kun** typegodkjent kalkulator er tillatt.

Sensuren faller: 5. sept. 2011

Svar **kort og konsist** på alle spørsmålene. **Stikkord** foretrekkes fremfor lange forklaringer. Les igjennom hele oppgavesettet før du begynner å lage løsning. Disponer tiden godt! Gjør rimelige antagelser der du mener oppgaveteksten er ufullstendig og skriv kort hva du antar. **Lykke til!**

Oppgave I (25%)

1. Hva er hovedhensiktene med **informasjonsgjenfinningssystem**? Forklar hva som er hovedforskjellene mellom data- og informasjonsgjenfinning.
2. Hva er rollen til indekstermer (index terms) i informasjonsgjenfinningssammenheng? Forklar hvordan "invertert indeks" (inverted index) fungerer. Bruk eksempel i forklaringen din.
3. Forklar hvordan boolske similaritetsmodellen (boolean similarity model) fungerer. Hva er grunnene til at boolske spørringer kan ha begrensede bruksområder?
4. Drøft hovedutfordringene med web (dvs WWW) sett fra informasjonsgjenfinningsperspektiv.
5. Forklar hvordan et IR-system kan bli evaluert og hvordan dette kan gjøres.

Oppgave II (25%)

1. Forklar stegene som er nødvendig fra man har en samling av dokumenter til disse er ferdig indeksert. (Tips: indexing pipeline).
2. Forklar hva begrepet "*edit distance*" er. Hva brukes det til? Hvis vi har en tekststrengen s_i, hvor $len(s_i)$ er lengden på s_i Vis at "*edit distance*" mellom s_1 og s_2 ikke kan være større enn max($len(s_1)$, $len(s_2)$).
3. "Feature" er et sentralt begrep i multimedia gjenfinning. Hva er hensiktene med "features"? Hvilke tre krav bør uthenting av features (*feature extraction*) oppfylle? Gi eksempler på features i forbindelse med videogjenfinning.
4. Forklar begrepet "Jaccard Coefficient". Hvorfor egner ikke denne seg så godt til rangering av søkeresultater?

Oppgave III (30%)

1. Sammenlikn *sannsynlighetsmodellen* – **Okapi BM25** og *vektormodellen* – **TF/IDF**. Hvilken ville du foretrekke dersom du skulle lage en tekstgjenfinningssystem. Grunngi svaret ditt. (Tips: Fokuser på prinsippene, ulempene og fordelene).
2. Testkolleksjoner (Test Collections) brukes ofte i evaluering av informasjonsgjenfinningssystemer. Gi eksempler på eksisterende testkolleksjoner. Hvordan brukes de? Hva er *R-precision*, *F-measure* og *MAP* (*mean average precision*)?
3. Utvidelse av spørringer (query extension).
    a. Forklar hva er hensiktene med utvidelse av spørringer (queries).
    b. Forklar kort prinsippene med **Rocchios** metode for ***User Relevant Feedback*** (URF).
    c. Hva er hovedforskjellen mellom "*automatic local analysis*" og "*automatic global analysis*"?
4. Forklar forskjellige metoder for indeksering av tekstdokumenter i informasjonsgjenfinningssystemer.
5. Det er to måter å oppsummere søkeresultater på. Forklar hva disse er og hvordan de fungerer.

Oppgave IV (20%)

Svar rett/galt på følgende utsagn. Hvert **riktig** og **begrunnet** svar belønnes med **2** poeng. **Feil svar** får **-1,5** poeng. **Ubegrunnet** eller **ingen svar** gir **0** poeng.

1. Piksel-til-piksel sammenlikning av to bilder er godt egnet til å beregne/evaluere bildenes likheter.
   (Rett/Galt)
2. To lydfiler kan sammenlignes ved å bruke frekvensspektrene til lydene.
   (Rett/Galt)
3. R-frame og R-precision er to mål som brukes i evaluering av søkesystemer.
   (Rett/Galt)
4. Statistisk Thesaurus er et thesaurus som brukes til å forbedre søk eller utvidelse av spørringer når man tar i bruk "*local automatic analysis*".
   (Rett/Galt)
5. *Harvest* og *Crowler* har samme funksjon, men brukes i to forskjellige web-søkemotorarkitekturer.
   (Rett/Galt)
6. Video-informasjon kan ikke gjenfinnes ved hjelp av gjenfinningsmetoder som er laget for bilder og lyd.
   (Rett/Galt)
7. Rangering av resultater i vektorbaserte og sannsynlighetsbaserte søkesystemer bruker samme prinsipp.
   (Rett/Galt)
8. Audiogjenfinningssystemer kan bruke teknikker kjent fra tekstgjenfinngsystemer.
   (Rett/Galt)
9. Bevegelsesinformasjon er ganske nyttig som feature til videogjenfinning.
   (Rett/Galt)
10. Bildehistogram kan brukes til å gjenfinne bilder. I tillegg kan det brukes i videogjenfinning.
    (Rett/Galt)