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 11. aug. 2012.
Tid: kl 09.00 – 13.00 (4 timer)

BOKMÅL

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

Sensuren faller: 31. aug. 2012

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 (30%)

1. Forklar hvorfor dagens bedriftssøk i for eksempel **Gulesider** eller amerikanske **Yelp** kan være både *informasjonsgjenfinning* og *datagjenfinning*. Bruk karakteristikkene på hver av disse gjenfinningskategoriene til begrunne svaret ditt. (6%)
2. Forklar hvordan "*edit distance*" kan brukes til stavekontroll på spørringer. (9%)
3. Forklar prinsippet bak den Boolske modellen. I forklaringen din skal du inkludere et eksempel som forklarer begrepet "*disjunctiv normal form (DNF)*". Gi en formell definisjon av similaritetsfunksjonen, *sim(d, q),* for denne modellen. (9%)
4. Forklar hva *Heap's Law* og Zipf's *Law* er og hva de kan brukes til. (6%)

Oppgave II (40%)

1. Forklar prinsippet bak **Divergence from Randomnes** gjenfinningsmodellen fra kap. 3 i læreboka. Hva er fordelene med denne sammenliknet mot Okapi BM25? Begrunn svaret ditt. (12%)
2. Forklar prinsippet bak "Latent semantic indexing" som er beskrevet i læreboka. (8%)
3. Til å evaluere et informasjonsgjenfinningssystem brukes det ofte forskjellige evalueringsmål (measures). Forklar følgende mål:
    a. Mean average precision (MAP) (3%)
    b. Average precision at *n* (2%)
    c. Mean reciprocal rank (MRR) (4%)
    d. Coverage og novelty i brukerorientert mål (5%)
4. Du skal gjennomføre en evaluering av et gjenfinningssystem, og etablerer en eksperimentsamling på 2000 dokumenter. Evalueringen skjer ved at det utarbeides et sett med spørsmål som kjøres mot eksperimentsamlingen, og at det utarbeides en "fasit" med dokumenter som anses for relevante for hvert spørsmål. I forhold til et gitt spørsmål finner du at 12 av dokumentene i eksperimentsamlingen er relevante. Du finner disse dokumentene på rangnummer I, 3, 6, 8, 9, 12, 15, 24, 36, 42, 45 og 50 i en rangert treffliste. Vis hvordan dette gjenfinningsresultatet kan vises grafisk gjennom en fullstendighet/presisjon (recall/precision)-graf. Hva blir R-presisjon i dette tilfelle? (6%)

Oppgave III (30%)

1. Forklar prinsippet bak invertertindekseringsmetode (inverted indexing method). Du må inkludere et eksempel på hvordan du konstruerer et invertertindeks (inverted index) i forklaringen din. (6%)
2. Forklar prinsippet bak "relevance feedback" for sannsynlighetsmodellen. (10%)
3. Hva er forskjellene og likhetene mellom HITS og Page Rank? Hva brukes de til? (8%)
4. Hva er de viktigste utfordringene ved indeksering og gjenfinning av henholdsvis bilder i et automatisert gjenfinningssystem, og hvilke teknikker er brukt til å møte disse utfordringene? (6%)