Norges teknisk-naturvitenskapelige universitet
Institutt for datateknikk og informasjonsvitenskap

EKSAMENSOPPGAVE I FAG TDT4117 – INFORMASJONSGJENFINNING (BOKMÅL)

Faglig kontakt under eksamen: Heri Ramampiaro

Tlf.: 93440

Eksamensdato: ***12. desember 2011***

Eksamenstid / varighet: *15.00-19.00 / 4 timer*

Tillatte hjelpemiddel: *D: Ingen trykte eller håndskrevne hjelpemiddel tillatt. Bestemt, enkel kalkulator tillatt.*

Språkform: Bokmål

Sensurdato: 3. januar 2012

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.

NorSoft AS er en bedrift som har et stort lager av tekstdokumenter, samt multimedia dokumenter bestående av bilder og lydfiler som ligger i en server. Hittil har de ansatte bare gjort manuelle søk basert på filnavn for å finne filer de trenger. Dette mener bedriften er for tidskrevende og bedriften må effektivisere generelt for bl.a. å spare penger. Du er nyansatt søkesjef på NorSoft og har fått i oppdrag å gjøre denne søkejobben enklere for de ansatte.

Oppgave I (25%)

I denne oppgaven skal du fokusere bare på tekstdokumentene.

a) Du har to valg: (1) å strukturere dokumentene og legge dem i en relasjonsdatabase, og (2) å hente ut innhold fra dokumentene og legge dem i et nytt informasjonsgjenfinningssystem. Dersom vi skal fokusere på de tre viktigste karakteristikkene som skille disse, forklar hvorfor valget ditt ville være alternativ 2.

b) Ved å fokusere på viktige tekstoperasjoner forklar hvordan du vil gå fram for å forberede dokumentene til indeksering.

c) Du er en bevisst medarbeider og ønsker å gi et godt inntrykk faglig overfor ledelsen. For å få til bedre treff på dokumentsøket bestemmer du deg for å undersøke nærmere hvilke metoder som kan benyttes til å rangere resultatene fra et søk. Hvilke metoder er mulig her? Forklar.

d) Anta at du til slutt landet på vektor-basert likhetsmodellen (”Vector Space Similarity Model”). Forklar hvordan dette fungerer – d.v.s. hva er prinsippet bak denne metoden?

Oppgave II (25%)

I denne oppgaven skal du fokusere på multimedia delen av bedriftens dokumentlager.

a) Du skal nå designe et systemarkitektur for bedriftens multimediasøkesystem. Forklar hvordan denne arkitekturen kan se ut. Bruk figur som en del av forklaringen din.

b) Du skal forklare sjefen din om hvordan multimedia datamodellen ser ut i virkeligheten slik at han forstår løsningene dine bedre. Tegn opp en figur av denne modellen (dvs datamodelltaxonomy). Hvilke tre lag (layers) består denne modellen av? Gi minst et eksempel på noe som hører til hvert lag.

c) Lydfilene består stortsett av taler men også noen musikkfiler. Forklar hvordan du går fram for å klassifisere disse to. Du kan her bruke en flytkartdiagram (flowchart diagram) i forklaringen din.

d) Anta at vi har tre logobilder i bildedatabasen, alle med størrelse på 9x9 piksler. Anta videre at pikslene kan ha en av disse 9 fargene C_1 til C_9 og er fordelt på følgende måte:
Bilde 1: 9 piksler i hver av 9 fargene
Bilde 2: 7 piksler i hver av fargene C_1, C_2, C_6, C_7, 11 piksler av fargene C_3 til C_5 og C_8, 9 piksler av C_9.
Bilde 3: 3 piksler i hver av fargene C_1 til C_3 og 12 piksler i hver av fargen C_4 til C_9
   I. Hva blir histogrammene til bildene over?
   II. Vis hvordan du beregner avstanden mellom de tre bildene.

Oppgave III (25%)

Anta nå at tekstsøkesystemet er opp og kjører og du er kommet til evalueringen av systemet med hensyn til kvalitet på søkeresultatene.

Tabell 1 viser resultat av et dokumentsøk som en av dine medarbeider har utført.

Rang	Dok ID	Score	Relevans
1	232	1534	
2	111	1240	REL
3	2343	1111	
4	332	1022	REL
5	441	1001	
6	112	0999	REL
7	22	0555	
8	1	0233	REL
9	23	0220	
10	334	0100	

Tabell 1 Søkeresultater

Anta at vi i alt har i alt 8 relevante dokumenter i bedriftens samling for dette søket.

a) For at sjefen din skal forstå hvor gode resultatene er forteller du ham at ”*precision*” og ”*recall*” kan brukes til å gjøre dette. Hvordan vil du forklare disse begrepene? Hvar er deres verdi, for eksempel, basert på informasjonen over (jfr. Tabell 1).

b) Det er to type grupper som utfører søkene i bedriften. Den ene gruppen - A - er de som er mest interessert i å finne mest mulig relevante treff, mens den andre - B - er de som er interessert i å finne alle relevante dokumenter dersom dette er mulig. Forklar hvilken gruppe du kan anbefale ”precision” for som hovedmål og hvilken du kan anbefale ”recall” for.

c) Hva er R-precision? Regne ut verdien av R-precision for resultatet over.

d) Hva er MAP (Mean Average Precision)? Forklar.

e) Hva er *E-measure*?

Oppgave IV (25%)

a) I likhet med andre bedrifter, brukes Web-søk også til finne andre informasjon som er relevant for bedriften. Du skal forklare en medarbeider hvordan websøkesystemene ser ut og vil vise henne spesielt den typen arkitektur som er mest brukt, nemlig den sentraliserte typen. Hvordan ser denne ut? Bruk figur i forklaringen din.

b) Etter å ha studert søkeresultatene som er gjort en stund klarer du ikke å bli enig med deg selv om den vektorbasert likhetsmodellen faktisk er den beste. Du vil teste noe annet og du lander på den språkmodellen (”Language Model for information retrieval”). Sammenlign denne modellen med vektorbaserte modellen. Fokuser på prinsipp, fordelene og ulempene.

c) Du fikk en ide om at du kan forbedre søkeresultatene ved å gjøre spørringene som sendes til systemet bedre. Forklar hva du kan bruke til dette formålet. Tips: Fokuser på søkeforbedringer.

d) Anta at i et av dokumentene du finner i bedriftens base inneholder følgende tekst:
”Merkel and Sarkozy both said last week that a fiscal pact should be written into the EU treaty”.
Konstruer et ”*suffix trie*” og ”*suffix tre*” basert på denne tekseten. Gjør de antakelsene du finner nødvendige.