



**EKSAMENSOPPGAVE I FAG TDT4117 – INFORMASJONSGJENFINNING
(SENSURVEILEDNING)**

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.

Svar: Alt 2. er valgt fordi IR støtter delvis matching, rangering av søkeresultater og er feiltolerant overfor feil is spørringer.

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

Svar: Her skal studentene forklare hvordan man gjøre leksikalanalyse, fjerning av stoppord, stemming, thesaurus og valg av indextermer.

- 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.

Svar: Her skal gi en kort oversikt over similaritetsmodeller (som rangerer) som finnes inkludert vektorspacemodellen, tradisjonelle sannsynlighetsmodellen, Okapi BM25, og Språkmodellen for IR.

- 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?

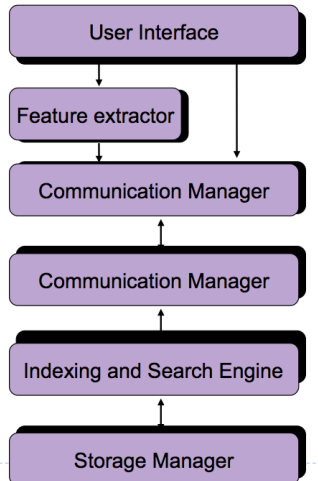
Svar: Her er det viktig at studentene forklarer at (1) hvert dokument samt spørring modelleres som vektor med elementer av termvekter. (2) Hver termvekt fås ved å bruke TF-IDF. Similaritet mellom et dokument D og en spørring Q kan beregnes deretter ved å beregne cosinus-verdien mellom vektorene for disse.

Oppgave II (25%)

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

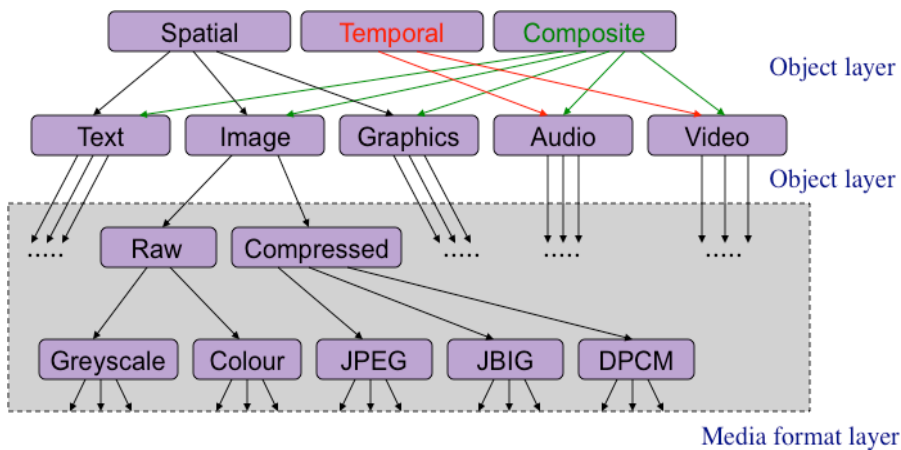
- a) Du skal nå designe et systemarkitektur for bedriftens multimediasøkesystem. Forklar hvordan denne arkitekturen kan se ut. Bruk figur som en del av forklaringen din.

Svar: Studentene må tegne denne figuren eller tilsvarende



- 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.

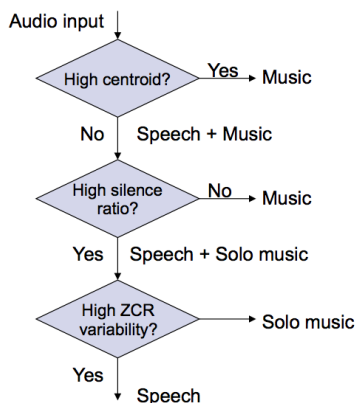
Svar: Studentene skal tegne dette:



Her heter lag 2 egentlig "Data Format Layer" men det er feil i transparenten fra forelesning og dette skal derfor ikke påvirke karakterene til studentene.

- c) Lydfilene består stort sett 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.

Svar: Studentene skal tegne denne:



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?

Svar: $H_1 = (9, 9, 9, 9, 9, 9, 9, 9, 9)$, $H_2 = (7, 7, 11, 11, 11, 7, 7, 11, 9)$, $H_3 = (3, 3, 3, 12, 12, 12, 12, 12, 12)$.

II. Vis hvordan du beregner avstanden mellom de tre bildene.

Svar: $H_1 - H_2 = 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 0 = 16$

$H_1 - H_3 = 6 + 6 + 6 + 3 + 3 + 3 + 3 + 3 + 3 = 36$

$H_2 - H_3 = 4 + 4 + 8 + 1 + 1 + 5 + 5 + 1 + 3 = 32$

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 dokumentssøk 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).
Svar: Precision: andelen av resultatene som er relevante. Recall: andelen av alle relevante dokumenter som er med i resultatsettet. $P = 4/10 = 40\%$ mens $R = 4/8 = 50\%$
- 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.
Svar: For A er precision viktigere enn recall da man er mer interessert i at resultatsettet inneholder mest mulig relevante dokumenter, men det er ikke viktig å finne alle av dem. Dermed er recall viktig for B da man er mest interessert i å finne alt.
- c) Hva er R-precision? Regne ut verdien av R-precision for resultatet over.
Svar: Precision når R resultater er tatt / hentet fram. Dvs. $4/8 = 50\%$.
- d) Hva er MAP (Mean Average Precision)? Forklar.
Svar: MAP – mean average precision er et mål som blir brukt til å måle evnen til et IR-system til å finne relevante dokumenter basert på et sett av queries.

For å finne verdien av MAP må man beregne gjennomsnittsverdien av Precision for hver relevant og gjenfunnet dokument. For eksemplet vårt i tabell 1 betyr dette at vi beregner precision på når vi har hentet 2, 4, 6, og 8 dokumenter. Deretter regnes snittet (AveP) av dette ut.

Dvs. $P(2) = 0.5$, $P(4) = 0.5$, $P(6) = 0.5$, og $P(8) = 0.5$. AveP for denne spørringen er derfor 0.5. Man gjør tilsvarende for alle de andre spørringene og regner en snittverdi for dette og får MAP.

$$\text{MAP} = \frac{\sum_{q=1}^Q \text{AveP}(q)}{Q}$$

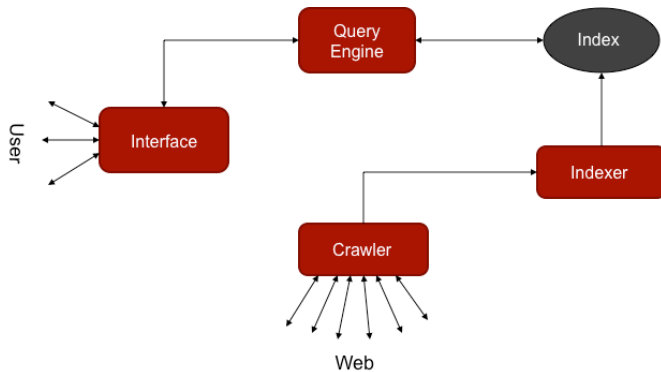
Hvor, Q er antall spørringer man har AveP verdien for.

- e) Hva er *E-measure*?
Svar: E-measure er brukt til å kombinere verdiene av precision og recall. $E = (1+b)PR / (b^2P + R)$, b er parametre brukt til å bestemme om viktigheten av P og/eller R.

Oppgave IV (25%)

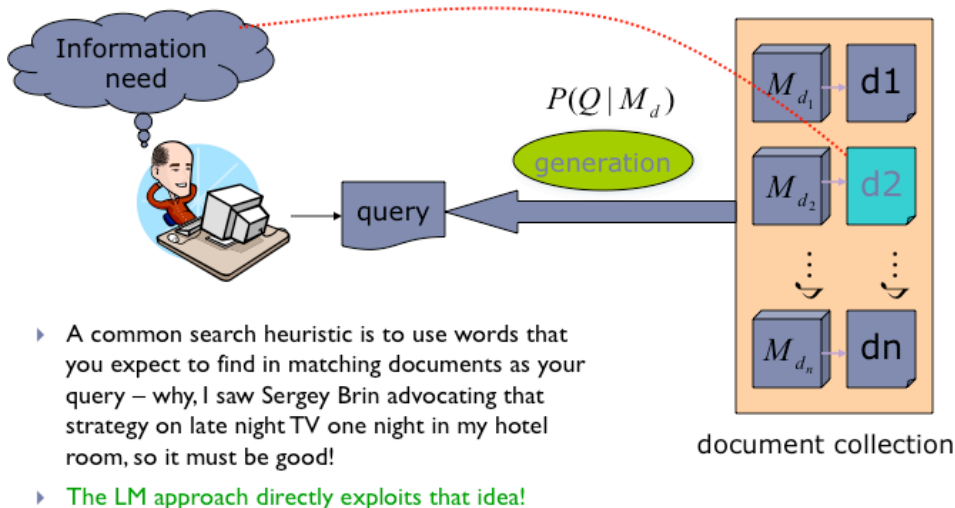
- a) I likhet med andre bedrifter, brukes Web-søk også til finne andre informasjon som er relevant for bedriften. Du skal forklare en medarbeider hvordan websøkesystemene ser ut og vil vise henne spesielt den typen arkitektur som er mest brukt, nemlig den sentraliserte typen. Hvordan ser denne ut? Bruk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

Svar: Her forventes at studentene skal tegne opp arkitekturen som vist nedenfor og forklare kort hvordan den fungerer



- 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.

Svar: Language modellen er bygd basert på lage dokumentmodeller basert på statistikk, og beregne hvor stor sannsynlighet er det for at hver dok.modell skal generere spørringen. Se figur under.



Hovedproblemet med LM er:

- LM antar at det er likhet mellom dokument og informasjon. Dette er i seg selv en svakhet
- LM baserer seg på en veldig forenkling av språkmodellen
- Vanskeligere å bruke URF enn i PM

Vanskeligere å integrere frasesøk, spørsmål og svar spørringer, og boolsøk.

- 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.

Svar: Her spørres det etter URF eller andre typer søkeforbedringer som lokal automatisk analyse og global automatisk analyse.

- 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 tree" basert på denne teksten. Gjør de antakelsene du finner nødvendige.

Svar: Her skal det legges mer vekt på at studentene viser at de har forstått prinsippet enn at de viser detaljekunnskap. Det er nok å vise deler av trie'et og tree't