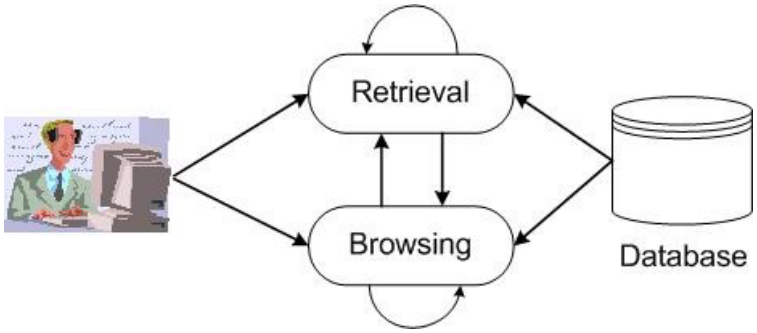
Sensurveiledning Eksamen i TDT4117 Informasjonsgjenfinning Høst 2018

Oppgave 1 – Lett blanding (15%)

John er eier av datakonsulentfirmaet John Wayne Consulting. Firmaet hans har spesialisert seg på å hjelpe andre bedrifter med å etablere interne bedriftssøkesystem (enterprise search systems).

1. Johns ansatte får ofte spørsmål om å forklare bedrifter hvorfor de ikke bare kan bruke en vanlig database for både lagring og søk av dokumenter. Hva mener du selv? (4%)

2. Et annet spørsmål som de ofte får er hvorfor de ikke bare kan lagre dokumentene på fil, strukturert i forskjellige kataloger, i stedet for å bruke et informasjonsgjenfinningssytem (IR-system). John tenker at den beste måten å svare på spørsmålet på er en forklaring på prinsippet bak et IR-system. Hvis du var John, hvordan ville du ha forklart prinsippet bak et IR-system? Bruk gjerne figur til å støtte forklaringen din. (6%)



3. John er irritert fordi en ansatt har brukt masse tid på å kjøre *stemming* som en del av stegene før han gjorde klart dokumentene for søk og gjenfinning. John mener stemming generelt er bare bortkastet tid da det ikke gir noen nytte. Hva mener du? Dvs. hvorfor ville du enten anbefale eller fraråde stemming generelt? (5%)

Oppgave 2 – Indeksering (15%)

Til spørsmålene nedenfor anta at vi har følgende tekst:
«**Britain** and the **European Union agreed** a **draft text setting** out a close **post-Brexit relationship, officials said.**»

1. Forklar hvordan signaturfil (signature file) fungerer. Bruk eksempel til å støtte forklaringen din. (5%)

2. Før selve indeksering trenger vi å gjennomføre opptil fem forskjellige steg for å forberede dokumentene. Disse stegen kalles ofte tekstoperasjoner. Forklar kort hver av disse, og vis hva vi får etter hver operasjon/steg hva vi får. Gjør de antakelsene du finner nødvendig. (5%)

3. Konstruer et «Suffix Tree» basert på teksten vår over.
   Gjør de antakelsene du finner nødvendig. (5%)

Oppgave 3 – Evaluering (10%)

Hva er Mean Average Precision (MAP)?

Anta at etter å ha kjørt to spørringer Q1 og Q2 får vi to rangerte liste av søkeresultater som er gitt i følgende tabell:

Rank	Q1
1	R
2	
3	R
4	
5	
6	R
7	
8	
9	R
10	R

Rank	Q2
1	R
2	R
3	
4	
5	
6	R
7	
8	
9	
10	R
11	
12	
13	
14	
15	R

Vi hvordan MAP blir beregnet basert på søkeresultatlistene i tabellene over. Gjør de antakelsene du finner nødvendige.

Oppgave 4 (15%)

1. Gitt følgende spørring, uttrykt ved hjelp av boolsk uttrykk:

$$q = G7 \wedge (\neg NATO \vee EU)$$

Gitt du har følende dokument:

```
d1:  G7 is not the EU
d2:  G7 and NATO it the West
```

Hva blir disjunctive normal form (DNF) for q? Hva er «conjunctive component» for dokumentene over. Gjør de evt. antakelsene du finner nødvendig. (6%)

2. I arbeid innen informasjonsgjenfinning har Vektormodellen (Vector Space Model) og Okapi BM25 blitt sammenliknet mot hverandre. Hva er felles for disse to modellene? Hvis man fokuserer på fordelen og ulempene, hvilken av disse to er den beste modellen for dokumentsøk? Begrunn svaret ditt. (9%)

Oppgave 5 (15%)

1. Forklar ulempene med det distribuerte websøkesystemet sammenliknet med det sentraliserte crawler-baserte systemet. (5%)

2. Det finnes flere forskjellige rangeringsmetoder for websøkesystem. Forklar kort to av disse. (5%)

3. Forklar hvorfor precision-recall-grafen ikke mulig bruke i evalueringer innen websøk. (5%)

Oppgave 6 (30%)

I følgende deloppgaver skal du krysse av et svar. Selv om du mener det kan være flere enn en påstand som er riktige **skal du ikke krysse av mer enn et svar**. (Alle delspørsmål teller likt, dvs. hvert riktig svar gir 3 poeng)

1.
    a. Precision og recall er like viktige uavhengig av søkeapplikasjoner.
    b. Recall er typisk viktigere enn precision for søk i Gulesider
    c. Precision er typisk viktigere enn recall for søk i rettsdokumenter
    d. Interpolering er nyttig dersom man har for få recall-punkter
2.

a. IDF står for «Intermediate Document Frequency» og brukes som mellomfunksjon for termvekter.
b. IDF står for «Invariant Document Frequency» og brukes til å måle hvor mye variasjoner er det i antall termer per dokument
c. IDF står for «Inverse Document Frequency» og kan brukes til å straffe termer som nevnes ofte i et dokument
d. IDF står for «Inverse Document Frequency» og kan brukes til å straffe termer som nevnes ofte i en samling av dokumenter

3.

a. Språkmodellen (the Language model) og sannsynlighetsmodellen bruker begge sannsynlighet til rangering, men skiller seg mest i hvordan sannsynligheten blir beregnet.
b. Språkmodellen (the Language model) er en variant av Okpi BM25
c. Ifølge forskningen fungere Språkmodellen (the Language model) mye dårligere enn boolsk-modellen.
d. Språkmodellen (the Language model) har ingenting med rangering av søkeresultater å gjøre.

4.

”F-measure” eller ”Harmonic Means” kombinerer precision og recall på same måte som MAP.
”F-measure” eller ”Harmonic Means” har ingenting med precision og recall å gjøre.
”F-measure” eller ”Harmonic Means” er begge mål for hvor god man ektraherer features fra bilder.
”E-measure” er generalisering av ”F-measure” eller ”Harmonic Means”.

5.

a. Multimedia informasjonsgjenfinning er ofte enklere enn tekstegjenfinning da man ikke trenger å utføre tekstoperasjoner.
b. Multimedia informasjonsgjenfinning er ofte vanskeligere enn tekstegjenfinning fordi multimedia objekter ofte er mer komplekse.
c. Multimedia informasjonsgjenfinning er ofte enklere enn tekstegjenfinning fordi de ofte kan lagres i en database.
d. Multimedia informasjonsgjenfinning er like vanskelig som tekstegjenfinning fordi man uansett må bruke tekstlig annoteringer.

6.

a. Fjerning av stoppord har veldig positive påvirkninger på Recall.
b. Fjerning av stoppord har alltid negative påvirkninger på Recall.
c. Fjerning av stoppord har alltid negative påvirkninger på precision.
d. Fjerning av stoppord har positive påvirkninger på precision.

7.

a. User Relevance Feedback (URF) er ofte brukt til å redefinere spørringer slik at man får økt søkehastighet.
b. User Relevance Feedback (URF) er ofte brukt til å redefinere spørringer slik at man får økt recall.
c. User Relevance Feedback (URF) er ofte brukt til å redefinere spørringer men vil aldri bidra til å øke precision.
d. User Relevance Feedback (URF) er ofte brukt til å redefinere spørringer slik at man får bedre spørringer, og dermed bedre recall og precision.

8.

a. Pixel-til-pixel sammenligning av to bilder er en veldig god måte å beregne på likheten mellom de to bildene.
b. Pixel-til-pixel sammenligning av to bilder er en ingen god måte å beregne likheten mellom de to bildene på, histogramsammenligning er enda verre.
c. Pixel-til-pixel sammenligning av to bilder er en ingen god måte å beregne likheten mellom de to bildene på, histogramsammenligning er bedre.
d. Pixel-til-pixel sammenligning av to bilder er en veldig god måte å beregne likheten mellom de to bildene på da pixelverdier er veldig gode features for bilder.

9.

a. «Scalar Cluster» er et begrep for å definere skalering av ordgrupperinger innen automatisk lokal analyse (automatic local analysis).
b. «Scalar Cluster» er et søkeforbedringskonsept innen automatisk global analyse (automatic global analysis).
c. «Scalar Cluster» brukes til søkeforbedring for å finne ord i hele samlingen av dokumenter som er relaterte til hverandre.
d. «Scalar Cluster» er en metode for å bygge thesaurus innen automatisk lokal analyse (automatic local analysis).

10.

a. Standard Rocchio og Ide Regular er to metoder som begge kan brukes til pseudo-relevance feedback
b. Hverken Standard Rocchio og Ide Regular kan brukes til pseudo-relevance feedback
c. Standard Rocchio kan brukes til pseudo-relevance feedback, men Ide Regular kan ikke det.
d. Standard Rocchio og Ide Regular er to av tre metoder som er laget for søkeforbedring men de er *kun* egnet for user relevance feedback.