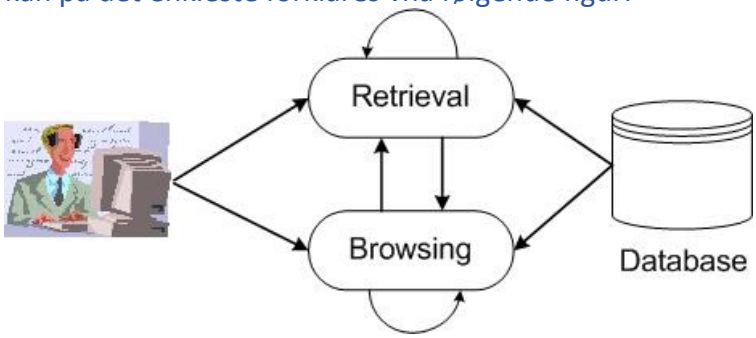
# Sensurveiledning Eksamen i TDT4117 Informasjonsgjenfinning Høst 2018

## 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%)
   Svar: Dette er et spørsmål hvor studentene skal vise at de har forstått forskjellene mellom Datagjenfinning og informasjongjenfinning. Det er derfor et poeng i seg selv at de tar med følgende aspekt som begrunnelse for å få full pott: mangel på rangering i forhold til fulltekst søk, relasjonsdatabaser krever strukturert lagring og fullstendig match.

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%)
   Svar: Studentene forventes her å kunne svare på hovedprinsippet ved å dekke brukerens informasjonbehov (information need) som står sentralt innen IR. Dette kan på det enkleste forklares vha følgende figur:




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%)
   Svar: Nytten av stemming er sterkt debatert innen IR. Fordelene med stemming er at ved å redusere ord til ordstammen kan dette redusere størrelsen på indeksen og øke antall treff ved at man kan finne alle dokumenter som inneholder ord med samme ordstamme(r). Hvis man her da er ute etter å finne flest mulig relevante dokumenter (og dermed økt recall) er dette nyttig. Er man derimot mest opptatt av at de fleste av returnerte dokumenter er relevante (og dermed høyst mulig presisjon) kan dette være unyttig.

## 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%)
   Svar: For å forklare la oss anta at vi har teksten som skal indekseres: ” Britain and the European Union agreed a draft text setting out a close post-Brexit relationship, officials said ”. Til dette trenger vi hashfunksjoner for hvert av ordene som skal indekseres.
   f(britain)= 0100 0001, f(european)= 0101 0010, f(union)= 0110 0011, f(agreed)= 1101 0100, f(draft)= 1100 0101, f(text)= 0110 0110, f(setting)= 0100 0111, etc…

   Antar her at disse ordene har følgende hashfunksjoner og at vi har fjernet stoppordene først, samt gjort leksikalanalyse av teksten. Antar videre at vi deler teksten i 3 blokker. Ved å bruke signturefil får vi basert på bitvis OR-operasjoner av hash-funksjonene i hver blokk følgende:

   Blokk 1: britain and the european union agreed
   Signartur for blokk 1: 0100 0001 OR 0101 0010 OR 0110 0011 OR 1101 0100 = 1111 0111

   Blokk 2: a draft text setting out a
   Signatur for blokk 2: 1100 0101 OR 0110 0110 OR 1100 0101 = 1101 0111

   Etc...

   Vi kan nå bruke dette til å søke om et ord med en gitt hash-funksjon finnes i en av blokkene (dvs. ved å utføre bitvis AND-operasjoner).

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%)
   Svar: Her er det som studentene skal vise er prinsippene bak tekstoperasjoner. Det forventes da at man først forklarer hva disse tekstoperasjonene er og vise hva man får etter hvert steg. Etter å ha utført tekstoperasjonene sitter vi igjen med: britain, europ, union, agree, draft, text, set, close, post, brexit, relat (eller relation), official, say.

3. Konstruer et «Suffix Tree» basert på teksten vår over.
   Gjør de antakelsene du finner nødvendig. (5%)
   Svar: Antar at vi har utført leksikalanalyse og stoppordfjerning før indekseringen. Vi må deretter sortere etter alfabetisk rekkefølge.

For å få til treet må vi først konstruere suffix-strengene og deretter finne posisjonene til ordene som skal indeksere. (Mangle på suffix-streng gir trekk)

1 9 10 17 26 … 84 98 108
Britain and the European Union … relationship, officials said

17 26 … 84 98 108
European Union … relationship, officials said

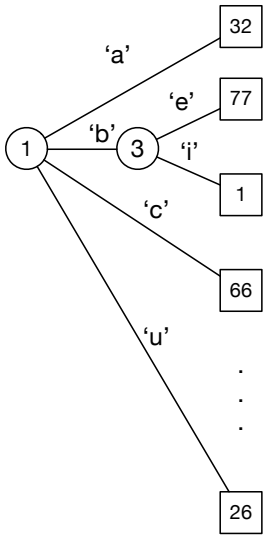
26 … 84 98 108
Union … relationship, officials said

…

108
Said

Sortert i alfabetisk rekkefølge får vi følgende ord og posisjon:
agreed 32
brexit 77
britain 1
close 66
draft 41
eurpean 9
official 98
post 72
relationship 84
said 108
setting 52
text 47
union 26

Da får vi følgende suffix-tre (vocabulary tree i stedet for suffix tree gir 0 poeng fordi dette er feil).




## Oppgave 3 – Evaluering (10%)

Hva er Mean Average Precision (MAP)?

Svar: MAP – Mean Average Precision er et evalueringsmål for å finne snitt av snittet av precisions for hvert relevant treff. Det vil representere et estimert verdi av arealet under precision-recall-kurven uten interpolering.

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.

Svar: MAP – Mean Average Precision ***kan*** beregnes ved å først regne ut gjennomsnitts-precision for hver spørring og deretter snitt av disse igjen. Precisions regnes ut der man finner relevante treff.

| Rank | Q1 | Precision |
|---|---|---|
| 1 | R | 1 |
| 2 | | |
| 3 | R | 2/3 |
| 4 | | |
| 5 | | |
| 6 | R | 1/2 |
| 7 | | |
| 8 | | |
| 9 | R | 4/9 |
| 10 | R | 1/2 |

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

$AVG\text{-}P_{Q1} = 0.62$, $AVG\text{-}P_{Q2} = 0.65$ => MAP = (0.62 + 0.65)/2 = **0.634**

**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%)

Svar: Q-DNF= (1, 0, 0) ∨ (1, 0, 1) ∨ (1, 1, 1). Conjunctive component for d1 = (1, 0, 1), og for d2 = (1, 1, 0).

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%)
   Svar: Felles for disse er: de tillater delvis match og gjør det mulig å rangere resultater av søk. Videre bruker de begge tf og idf i beregningen av likhetsmålet (similarity measure).

| **Vektormodellen** | | **BM25** | |
|---|---|---|---|
| **Fordeler** | **Ulemper** | **Fordeler** | **Ulemper** |
| Enkel å bergne | Alle termen antas å vær uavhengig av hverandre. Tar ikke hensyn til semantikk | Bygger på sannsynlighetsteori om relevans | Antar termene er uavhengige |
| Bruker ikke binære relevans | | | Estimering ikke presis nok |
| Kan regnes ut veldig effektivt | | | |

**Oppgave 5 (15%)**

1. Forklar ulempene med det distribuerte websøkesystemet sammenliknet med det sentraliserte crawler-baserte systemet. (5%)
   Svar: Ulempene er at det distribuerte krever koordinering av gatherers via brokers, noe som kan føre til overhead. I tillegg er ikke distribuerte websøkesystem utbredt nok, noe som gjøre de litt vanskelig å videreutvikle systemet effektivt nok.

2. Det finnes flere forskjellige rangeringsmetoder for websøkesystem. Forklar kort to av disse. (5%)
   Svar: Her skal man forklare kort feks. HITS og Page Rank.

3. Forklar hvorfor precision-recall-grafen ikke mulig bruke i evalueringer innen websøk. (5%)
   Svar: Precision-recall-graf kan ikke brukes innen websøk da det ikke er mulig å regne ut recall.

**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.