NTNU
Kunnskap for en bedre verden

Institutt for datateknologi og informatikk (IDI)

Ordinæreksamen i TDT4117 Informasjonsgjenfinning

Faglig kontakt under eksamen: Heri Ramampiaro

Tlf.: 99027656

SENSURVEILEDNING

Eksamensdato: 18.12.2017

Eksamenstid (fra-til): 09:00-13:00

Hjelpemiddelkode/Tillatte hjelpemidler: D – Kun godkjent kalkulator tillatt

Annen informasjon:

Målform/språk: Bokmål
Antall sider (uten forside):
Antall sider vedlegg:

Informasjon om trykking av eksamensoppgave
Originalen er:
1-sidig ☐ **2-sidig** ☐
sort/hvit ☐ **farger** ☐
skal ha flervalgskjema ☐

Kontrollert av:

_____

Dato Sign

Oppgave I – 40%

John Smith er ekspert i informasjonssøk og gjenfinning. Han er nytilsatt i et firma som selger varer og tjenester på nettet. Hans oppgave er å bistå ledelsen i å lage en god løsning for lagring og gjenfinning av informasjon om varene de skal selge. Han skal også bygge en kunnskapsbase som inneholder all informasjon (i form av tekstdokumenter) om tidligere varesalg og erfaringene med dette. Anta at du nå skal fungere som rådgiver for John.

1. Forklar hvorfor søk i informasjonen om varene både kan være informasjonsgjenfinning og datagjenfinning.

   Svar: For å få full pott må svaret inneholde en kort beskrivelse på karakteristikkene til data- og informasjonsgjenfinning:

	Data retrieval	Information retrieval
Content	Data	Information
Data object	Table	Document
Matching	Exact match	**Partial match, best match**
Items wanted	Matching	**Relevant**
Query language	SQL(artificial)	Natural
Query specification	Complete	**Incomplete**
Model	Deterministic	Probabilistic
	Highly structured	**Less structured**

   Table by Xin Xao, Drexel University

   Vareinformasjonen kan lagres som en strukturert informasjon i en relasjonsdatabase og kan oppfylle alle karakteristikkene til datagjenfinning. Beskrivelsene av varene kan også lagres som ustrukturert (og/eller semi-strukturert) tekst og kan derfor være indeksert i et IR-system som kan oppfylle de viktigste karakteristikkene for informasjonsgjenfinning.

2. I følgende deloppgaver skal du forklare hvordan du ville gå frem for å hjelpe John til å få indeksert dokumentene som skal inn i kunnskapsbasen. Bruk de antakelsene du finner nødvendige.
   a. Hvilke ***fem*** tekstoperasjoner trenger du å bruke for å forberede dokumentene til indekseringen. Forklar kort hver av disse.

   Svar:
   (1) Leksikalanalyse – Gjøre teksten om til tokens eller sett av termer, fjerner evt. unødvendige tall, bindestrek, etc.
   (2) Stopporfjerning – Fjerne alle vanlige termer som feks. artikler (”the”, ”a”, etc.), samt de som nevnes i de fleste dokumentene i dokumentsamlingen.
   (3) Stemming – Gjøre om alle termer til ordstamme (feks. conntected, connecting, connector til connect).
   (4) Bygge thethaurus – Lage thesaurus av termer slik at dokumenter kan indekseres med termer med samme betydning også, i tillegg til de fra dokumentene.
   (5) Valg indekstermer – Velge termer som skal indekseres manuelt av en domeneekspert og/eller automatisk som navn, spesielle termer, etc.

   NB: *Det er viktig at man viser forståelse på disse operasjonene og ikke bare lister de opp uten å ta hensyn til feks. rekkefølgen. For eksempel gir det ikke mening å kjøre stoppordfjerning før man har gjort leksikalanalyse, eller kjøre stemming før man har fjernet stoppordene først, osv.*

   b. Anta at du har valget mellom sannsynlighetsmodellen (probabilistic similarity model) og vektormodellen (vector space model) som skal implementeres som likhetsmodell. Drøft hvilken av disse modellene du ville anbefalt John. For å overbevise ham er det viktig at du gir ham informasjon om både fordeler og ulemper med hver av modellene, samt en kort beskrivelse av prinsippene bak dem.

Svar:

For å få fullpott forventes at svaret viser at man har forstått prinsippene med begge modellene, dvs. at svaret må inneholde en kort forklaring av hva disse to modellene går ut på.

Vektormodellen	**Sannsynlighetsmodellen**
Fordeler: - Delvis søk tillatt (bøhever ikke å finne eksakt match) - Rangering av resultater - Veldig enkel	- Delvis match tillatt - Rangering av søkeresultater basert på estimering av relevanssannsynlighet som basis
Ulemper: - Antar at alle termer er uavhengig - Kan ta med mange dokumenter som brukeren ikke mener er relevante	- Må estimere den initielle sannsynligheten noe som ikke alltid er rettfrem - Tar ikke hensyn til TF og IDF

c. Anta at John ikke er lett å overbevise siden han mener begge er jevngode, så du må ty til praktisk evaluering for vise ham hvilken modell som dere bør velge. Drøft først ***tre*** andre forskjellige evalueringsmål (evaluation measures), *i tillegg* til precision og recall som du kan bruke. Forklar deretter hvordan du ville gå fram med evalueringen.

Svar:
(1) MAP – Mean Average Precision som går ut på å finne snitt av alle precisionverdiene for hver spørring og deretter ta snittet av disse verdiene igjen. Precision verdier her kan regnes ut ved å bruke precision-recall-punkttabell eller interpolert precisionverdier fra hvert recall-punkt.
(2) F-measure – Verdien av harmonic means mellom precision (P) og recall (R): (2PR/(P+R).
(3) R-precision – P@R hvor R er verdien av totale relevante dokumenter

Det som spørres her er egentlig hvordan man evaluere et IR-system: En måte å evaluere et IR-systemet er: gitt en samling av dokumenter, lag et sett med spørringer og fasit. Deretter prøver man å finne ut ved å bruke modellene og spørringene hvor mange av returnerte dokumenter er relevante. Bruker deretter dette til å regne ut precision og recall, og/eller de andre målene over. Den modellen som har beste MAP-verdi er den beste modellen. (Bonus gis til besvarelse som nevner bruken av testsamlinger som TREC og/eller GOV2).

NB: Det er viktig at studentene fokuserer på at evalueringen skal finne ut i hvilken grad et IR-system imøtekommer brukerens informasjonsbehov og evnen til å gjenfinne relevante dokumenter.

3. For å effektivisere indekseringen trenger John råd om hvilken indekseringsmetode dere skal benytte. John mener at et alternativ for dere er å benytte *signaturfil*. Du er derimot uenig med ham og foreslår en annen metode.
   a. Forklar hvorfor du mener indekseringsmetode med *signaturfil* ikke egner seg så godt for oppgaven.

Svar: Signaturfil egner seg ikke godt her da vi kan anta at datamengden er stor og signaturfil passer best når datamengden er liten. Dette fordi signaturfil baserer seg på hash-kode som blir generert for

hvert ord i vokabularet som skal indekseres, som igjen brukes til å lage signatur for hver blokk i en setning og dokument.

b. Hvilken annen metode ville du heller ha valgt? Begrunn svaret ditt.

Svar: Den vanligste og den mest effektive indekseringsmetoden er invertertindeks (inverted index).

Full pott gis til svar som inneholder en kort forklaring på prinsippet bak inverted index.

4. Vareinformasjonen inneholder både bilder og video, i tillegg til tekst. Anta at systemet deres skal tillate søk på bilder av varene, og du velger å bruke bildehistogram som hoved-feature for dette. Hva menes med "feature" i denne sammenheng? Forklar med eksempel hvordan du kan bruke histogrammet til å sammenlikne to bilder. Gjør de antakelsene du finner nødvendige.

Svar: Feature er noe som best representer og/eller karakteriserer et media objekt ifm innholdsbasert gjenfinning av multimedia informasjon. Dette kan være farger, fargehistogram, form (shape), og texture for bilder. For video kan dette være R-frame, objekter, bevegelsesinformasjon, tekstannotering, osv.

Eks. på hvordan histogram kan brukes:
Anta at de 2 bildene vi har er med størrelse på 9x9 piksler. Anta videre at pikslene kan ha en av disse 9 fargene C1 til C9 og er fordelt på følgende måte: Bilde 1: 7 piksler i hver av fargene C1, C2, C6, C7, 11 piksler av fargene C3 til C5 og C8, 9 piksler av C9.
Bilde 2: 3 piksler i hver av fargene C1 til C3 og 12 piksler i hver av fargen C4 til C9

Histogrammene til bildene blir da:
H1= {7, 7, 11, 11, 11, 7, 7, 11, 9}, H2= {3, 3, 3, 12, 12, 12, 12, 12, 12}.

Oppgave II – 30%

I hver av følgende deloppgaver er det gitt flere alternative påstander. Du skal velge kun ***en*** riktig påstand. Dersom du synes flere enn en påstand er riktig, velger du den som du mener er mest riktig. Svar kun med spørsmålnummer og nummer på riktig svaralternativ (f. eks. 11.a, etc.). Du skal ikke begrunne svaret ditt. Hvert riktig svar gir **tre** poeng, mens feilsvar gir **ingen** poeng.

1.
   a. Fargepiksler egner seg ikke til å sammenlikne to bilder siden de ikke tar hensyn til nyansene i fargene.
   b. **Fargepiksler kan fint brukes til å sammenlikne to bilder siden det er de som danner grunnlaget for å lage fargehistogram.**
   c. Fargepiksler kan ikke brukes til å sammenlikne to bilder siden piksler bare inneholder informasjon om fargepunkter og ikke noe annet.
   d. Fargepiksler kan fint brukes til å sammenlikne to bilder siden hvert pikselpunkt ikke kan påvirkes av bildestøy.
2.
   a. R-frame er en betegnelse for gjennomsnittsbildet i en bildesamling.
   b. **R-frame kan være en betegnelse for gjennomsnittsbildet i en videosekvens.**
   c. R-frame har ikke bare med videogjenfinning å gjøre, men er også et evalueringsmål i IR.
   d. R-frame er en betegnelse for gjennomsnittsbildet i en videosamling.
3.

a. Thesaurus kan ikke brukes til utvidelse av spørringer da det kun er ”global automatic analysis” som kan bruke det.
b. ”Global automatic analysis” er en metode som ikke har noe med informasjonsgjenfinning å gjøre, men med dataanalyse generelt.
c. ”Global automatic analysis” er en metode for å utvide spørringer der man bruker det returnerte resultatet fra et søk som grunnlag.
d. **”Global automatic analysis” er en metode for å utvide spørringer der man bruker hele samlingen for å lage thesaurus.**

4.

a. Crawlers brukes i distribuerte web-søkemotorer mens gatherers brukes i sentraliserte web-søkemotorer.
b. **Crawlers brukes i sentraliserte web-søkemotorer mens gatherers brukes i distribuerte web-søkemotorer.**
c. Crawlers brukes i både sentraliserte og distribuerte web-søkemotorer.
d. Gatherers har ingen ting med web-søk å gjøre, bare crawlers.

5.

a. ”Micon” er en viktig feature for bilder og brukes i bildegjenfinning. Det kan sidestilles med ”index terms” for tekstgjenfinning.
b. ”Micon” er en viktig feature for video men brukes også i bildegjenfinning. Det kan sidestilles med ”index terms” for tekstgjenfinning.
c. **”Micon” er en viktig feature for video og brukes i videogjenfinning. Det kan sidestilles med ”index terms” for tekstgjenfinning.**
d. ”Micon” har ingenting med hverken bildegjenfinning eller videogjenfinning å gjøre.

6.

a. R-Precision er en forkortelse for Recall-Precision.
b. R-Precision er en forkortelse for Rounded-Precision.
c. **R-Precision er definert som andelen av topp-R dokumenter som er gjenfunnet som er relevante, hvor ”R” er det total antall relevante dokumenter i samlingen.**
d. R-Precision er definert som andelen av topp-R dokumenter som er gjenfunnet som er relevante, hvor ”R” er recall-verdien for spørringen.

7.

a. “Vocabulary Trie” og “Suffix Trie” er to begrep som beskriver samme indekseringsmetode.
b. “Vocabulary Trie” og “Suffix Trie” er to helt forskjellig type trær som ikke har noe med indeksering å gjøre.
c. **“Vocabulary Trie” og “Suffix Trie” er to konsept som brukes i to forskjellige indekseringsmetoder, hvor “Suffix Trie” i seg selv er en indekseringsmetode.**
d. “Vocabulary Trie” og “Suffix Trie” er to konsept som brukes i to forskjellige indekseringsmetoder, hvor “Vocabulary Trie” er i seg selv er en indekseringsmetode.

8.

a. Web-søkemotorer bruker ikke ”stemming” fordi stemming gjør at man bruker for mye ressurser uten å få noen høyere recall.
b. Web-søkemotorer bruker ikke ”stemming” fordi stemming gjør at man bruker for mye ressurser uten å få hverken høyere recall eller høyere precision.
c. **Web-søkemotorer bruker ikke ”stemming” fordi at selv om stemming kan gi høyere recall gir dette ikke nødvendigvis noen gevinst i form av høyere precision.**
d. Alle Web-søkemotorer må bruke ”stemming” fordi stemming kan gi høyere recall, og dette er veldig viktig i web-søk generelt.

9.

a. Ingen komprimeringsmetode kan brukes i informasjonsgjenfinning siden den komprimerte teksten da alltid må dekomprimeres først.
b. Det er flere komprimeringsmetoder som kan brukes i informasjonsgjenfinning. Disse kjennetegnes ved at de alltid kan bygges som et tre, som f. eks. et Huffman-tre, etc.
c. Det er flere komprimeringsmetoder som kan brukes i informasjonsgjenfinning. Disse kjennetegnes ved at de er statiske komprimeringsmetoder.
d. **Det er flere komprimeringsmetoder som kan brukes i informasjonsgjenfinning. Disse kjennetegnes ved at de er semi-statiske og tillater direkte aksess.**

10.

a. Den største likheten mellom "Probabilistic Similarity Model" og "Language Model" er måten sannsynligheten blir beregnet.
b. Den største likheten mellom "Probabilistic Similarity Model" og "Language Model" er at begge bruker TF-IDF i estimeringen.
c. **Den største likheten mellom "Probabilistic Similarity Model" og "Language Model" er at begge bruker sannsynlighet som basis. Måten denne blir beregnet på er imidlertid forskjellige.**
d. Den største likheten mellom "Probabilistic Similarity Model" og "Language Model" er at begge har fokus på beregning av sannsynlighet for relevans, gitt en spørring.

Oppsummert svar:
1. b,
2. b,
3. d,
4. b,
5. c,
6. c,
7. c,
8. c,
9. d,
10. c.

Oppgave III – 30%

1. Anta at vi har følgende par av tekststrenger. Finn **edit distance**-verdiene mellom parene.
   a. "levenshtein" vs. " lichtenstein"
   b. "stonebrook" vs. "steinberg"

   Svar: Edit distance defineres som minimum antall operasjoner (sletting av tegn, bytting av tegn, og legg til tegn) som trengs for å få to strenger til å bli lik hverandre (se også s. 222 i læreboka).
   a. edit distance("levenshtein" vs. " lichtenstein") = **5**
   b. edit distance("stonebrook" vs. "steinberg") = **7**

   NB: For å få poeng i det hele tatt må man vise at man vet hva en edit distance er og hvordan den blir beregnet. Det holder ikke å bare skrive et tall.

2. Anta følgende tabell som viser rangering av et søkeresultat:

Rank	Doc ID	Relevant?
1	8	REL

2	9	REL
3	12	
4	5	REL
5	2	
6	17	
7	23	
8	10	REL
9	1	
10	4	
11	30	
12	3	
13	6	
14	13	REL

Anta videre at totalt antall relevante dokumenter for denne spørringen er 8.

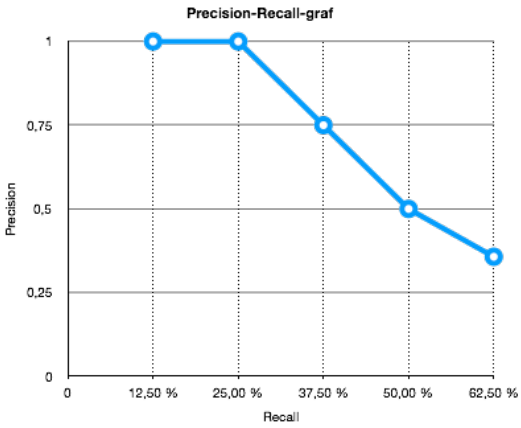
a. Bruk tabellen over til å regne ut precision- og recall-punktene.

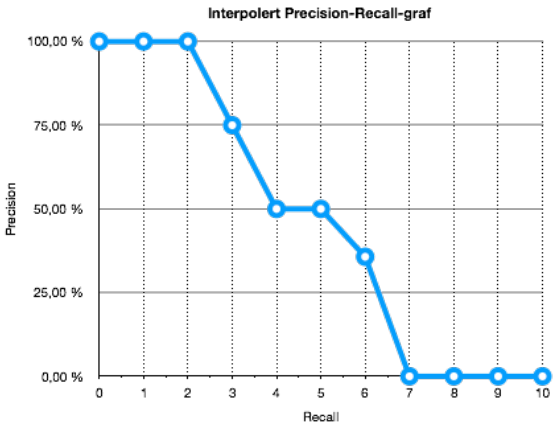
Svar:

NB: Det gir ikke mening å beregne Precision- og recall-punktene der det ikke relevante dokumenter. Full pott gis derfor kun til de som har forstått dette.

Rank	Doc ID	Relevant	Precision	Recall
1	8	REL	100,00 %	12,50 %
2	9	REL	100,00 %	25,00 %
3	12			
4	5	REL	75,00 %	37,50 %
5	2			
6	17			
7	23			
8	10	REL	50,00 %	50,00 %
9	1			
10	4			
11	30			
12	3			
13	6			
14	13	REL	35,71 %	62,50 %

b. Tegn opp en graf som viser verdiene, dvs. en precision-recall-graf. Vis deretter hvordan den interpolerte (interpolated) versjonen av denne grafen ser ut.





I den interpolerte grafen tilsvarer tallene 0, 1, 2, 3, ... hhv. 0, 10%, 20%, 30% recall-verdier. (Figur presentert som "trapp" er også ok).

3. Følgende formler brukes til å forbedre spørringer i forbindelse med brukerrelevans-feedback (User relevance feedback):

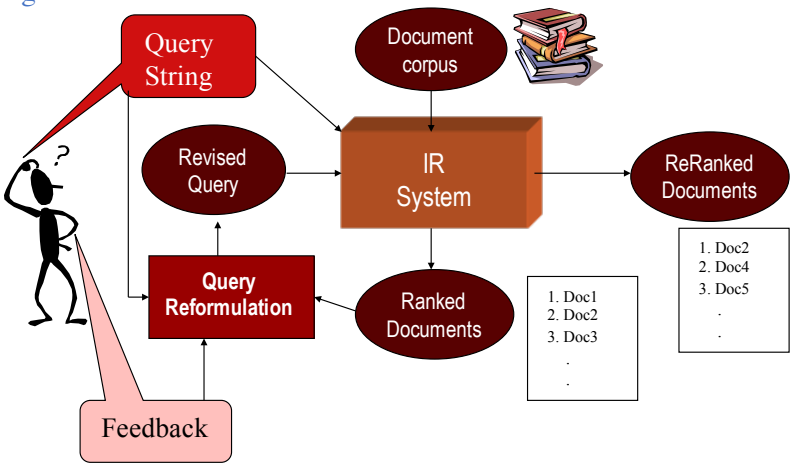
$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \frac{\gamma}{|D_n|}\sum_{\vec{d}_j \in D_n}\vec{d}_j \quad (1)$$

$$\vec{q}_m = \alpha\vec{q} + \beta\sum_{\forall\vec{d}_j \in D_r}\vec{d}_j - \gamma\sum_{\forall\vec{d}_j \in D_n}\vec{d}_j \quad (2)$$

$$\vec{q}_m = \alpha\vec{q} + \beta\sum_{\forall\vec{d}_j \in D_r}\vec{d}_j - \gamma\max_{non-relevant}(\vec{d}_j) \quad (3)$$

a. Hva heter disse formlene (1), (2), og (3), og forklar **kort** hvordan de brukes.

Svar: Alle tre formlene brukes til URF for vektormodellen som kan forklares ved hjelp av følgende figur:



Formelen brukes da til "query refomulation"-delen for å produsere nye rangeringer.

(1) Rocchios update method
(2) Ide Regular
(3) Ide "Dec hi"

Dr her er sett av alle relevante dokumenter, Dn er sett av alle irrelevante dokumenter. Alfa, gamma, og beta er alle konstanter som brukes til å bestemme leddenes viktighet.

b. Forklar **kort** hva som menes med begrepet ”implicit relevance feedback”.

Svar: Implisit relevance feedback er feeback fra brukeren som ikke er direkte valg av relevans. Dette kan feks. være at man bruker klikk-informasjon som indirekte angivelse av relevans (dvs. hvis en bruker åpner og leser på et dokument som er returnert, kan dette ses på som relevant etc.).