NTNU
Kunnskap for en bedre verden

Institutt for datateknologi og informatikk (IDI)

Ordinæreksamen i TDT4117 Informasjonsgjenfinning

Faglig kontakt under eksamen: Heri Ramampiaro

Tlf.: 99027656

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.

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.
    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.
    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.

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

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.

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.

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”

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.
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.

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