NTNU
Kunnskap for en bedre verden

Department of Computer and Information Science

TDT4117 Information Retrieval Exam

Contact during the exam: Heri Ramampiaro

Phone: 990 27 656

Exam date: 08.12.2015

Exam time (from-to): 09 - 13

Allowed aid: D – No printed or written materials allowed. Only approved calculator is allowed.

Other Information:

Language: English

Number of pages: 8

Number of appendices: 0

Controlled by:

_____

Date Sign

Please give short and concise answers *to all questions.* ***Short sentences*** *are preferred rather than long explanation. It is highly recommended that you read the whole exam set before starting answering the questions. Gjør rimelige antagelser der du mener oppgaveteksten er ufullstendig og skriv kort hva du antar. Lykke til!*

Oppgave I - Lett blanding (20%)

Ola Nordman er nettopp ferdig med masterutdannelsen sin. Han fikk en ide om å starte en egen bedrift som et enmannsforetak. Før han kunne starte opp må han sette seg inn i en del lover og regler i forhold til skatt, plikter og ansvarsforhold. Ola vet at Brønnøysundregistrene var et perfekt sted å starte, i stedet for å lese gjennom boka ”Norges Lover".

Basert på denne lille historien skal du svare på følgende delspørsmål:

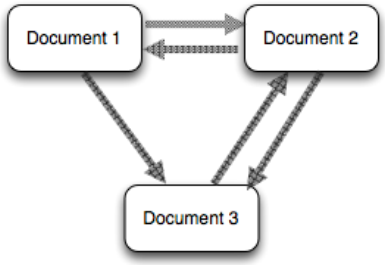
1. Ola ønsker å finne all relevant informasjon som omhandler skatteregler for enmannsforetak. Drøft kort hvorfor det er viktig for Ola at de som har utviklet søkesystemet for Brønnøysundregistrene har fokusert på høyst mulig "recall" fremfor “precision".
2. Ola vil etterhvert trenge å registrere bedriften sin i Enhetsregisteret, men før han gjør dette bestemte han seg for å finne ut hvordan andre har gjort dette før. Er problemet til Ola en "*datagjenfinning*" eller "*informasjonsgjenfinning*"? Begrunn svaret ditt.
3. Ola finner ut at Brønnøysundregistrene lagrer informasjonen både veldig strukturert og ustrukturert. Han finner også ut at de har gjort det veldig enkelt å bla gjennom dokumenter via linker. Han mistenker at når han søker generelle informasjon om bedrifter så bruker de nettopp denne *linkinformasjonen* til å rangere søkeresultatene. Anta at den som er blitt brukt er "*page rank*”, som er angitt i følgende formell:

$$PR\left(p_i\right) = \frac{1-d}{N} + d \cdot \sum_{p \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

Anta at d = 0.15.

a) Forklar hovedideene bak ”page rank” ved å tolke formelen ovenfor.

b) Se for deg at du har en linkstruktur som vist på Figur 1 på de dokumentene som er returnert fra søket. Vis hvordan du regner ut og rangerer dokumentene ved hjelp av page rank. Bruk formelen ovenfor. Gjør også de antakelsene du finner nødvendig.



Figur 1 Linker mellom dokumentene

c) Nevn og forklar kort to andre metoder som utnytter linkstruktur til å rangere søkeresultater.

4. Ola stusser litt på at man bruker typiske websøkemetoder til å håndtere søk i registrene. Han tenker at det ville være best med tradisjonelle informasjonsgjenningsbaserte metoder. Drøft kort hovedforskjellene mellom et typisk websøkesystem og et tradisjonell informasjonsgjenningssystem.

Oppgave II Tekstoperasjoner og Indekseringsteknikker (30%)

1. Forklar hvorfor man mener "stemming" er noe kontroversiell spesielt i forbindelse med websøk. Forklar tre andre tekstoperasjonsmetoder som man kan bruke.
2. Anta at vi allerede har fjernet alle stoppord fra følgende tekst. Konstruer invertert file/liste for følgende tekst:

''President Barack Obama warned his Russian counterpart Tuesday against intervening in Syria's civil war, suggesting that Vladimir Putin is aware of the dangers his country faces by entering the conflict''

a) Bruk teksten over til å konstruere invertert liste med blokkadressering.

b) Konstruer et partielt vokabular "trie" av teksten over.

3. Forklar hvordan signaturfilindekseringsteknikken fungerer. Bruk gjerne eksempel til å støtte forklaringen din.

Oppgave III Similaritetsmodeller og Evaluering (30%)

1. Anta at vi har følgende dokumenter:

D1 =''George Bush is former American President, but he is still called president''

D2 =''President Barak Obama's presidential period will soon be over, and yet another Bush may well become a new president''

Anta at spørresetningen q= "american president bush" er brukt til å søke på dokumentene.

a) Finn alle nøkkelordene (som kan brukes som index terms) i de to dokumentene og sett opp vokabularsettet K. Gjør de antakelsene du finner nødvendige.

b) Bruk formelen nedenfor og vis hvordan du regner ut likhetsfaktoren (similarity), *sim(q, D),* mellom dokumentene og spørresetningen ved hjelp av "vector space" modellen. Hvilket av de to dokumentene blir rangert først?

$$Sim(q,d_j) = \cos(\theta) = \frac{\vec{d_j} \bullet \vec{q}}{\left|\vec{d_j}\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{ij} w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

c) Hva tror du er grunnene til a denne modellen er mer populær enn sannsynlighetsmodellen?

2. Forklar hovedforskjellene mellom språkmodellen (Language model for information retrieval) og sannsynlighetsmodellen.
3. Du skal gjennomføre en evaluering av et gjenfinningssystem, og etablerer en eksperimentsamling på 1000 dokumenter. Evalueringen skjer ved at det utarbeides et sett med spørringer som kjøres mot eksperimentsamlingen, og at det utarbeides en såkalt "ground truth" eller "fasit" med dokumenter som anses for relevante for hver spørring. I forhold til et gitt spørsmål finner du at 15 av dokumentene i eksperimentsamlingen er relevante. Vi antar at spørringen har i alt 20 relevante dokumenter. Du finner disse dokumentene på rangnummer 1, 3, 6, 8, 9, 12, 15, 24, 36, 42, 43, 45, 50 , 54, og 60 i en rangert treffliste. Ta utgangspunkt i første 15 returnerte dokumenter i resultatlista og regn ut precision- og recall-punktene. Hva blir *f-measure* (hamonic means) verdien?
4. Forklar kort hvorfor "recall" er viktigere enn "precision" for Lovdatasøk, mens “precision” er viktigere enn “recall” for søk i Gulesider o.l .

Kandidatnr.:_____

Dette arket skal leveres sammen med besvarelsen. Husk derfor å føre på ditt kandidatnummer.

Oppgave IV (20%)

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. Riktig svar gir 2 poeng**)

1.

- ☐ Micon og videogjenfinning hører sammen
- ☐ Micon og lydgjenfinning fra video shots hører sammen
- ☐ Micon kan brukes av PST til å lett gjenfinne bilder med terrorister
- ☐ Micon har ingen ting med informasjonsgjenfinning å gjøre

2.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

- ☐ Denne formelen er et likhetsmål som egner seg veldig godt til bruk i rangering av resultater med søk av store strukturerte dokumenter
- ☐ Denne formelen forteller noe om mengden av informasjon som er i dokumentene
- ☐ Denne formelen er en Jaccard likhetsmål som ikke kan brukes til å vurdere om to ord er like eller ikke
- ☐ Denne formelen er en Jaccard likhetsmål som kan brukes til å vurdere om hvor stor grad to dokumenter er like

3.

- ☐ Et Huffman tre er en variant av et suffix tre

☐ Et Huffman tre er et komprimeringstre som kan være både ”bit” basert (binært) og ”byte” basert samtidig.

☐ Et Huffman tre en variant av et suffix trie

☐ Et Huffman tre konstruksjon bruker informasjon om termdistribusjon i et dokument for å kunne lage en så effektiv komprimering som mulig

4.

☐ Automatic Global Analysis og Automatic Local Analysis bruker to motsatte metoder få utvidede spørringer

☐ Pseudo relevance feedback er en metode for å utvide spørringer

☐ I Pseudo relevance feedback trenges tilbakemelding fra brukeren for at det skal fungere optimalt

☐ Med Pseudo relevance feedback trenger man alltid å bruke Rocchios standard metode for få det til å fungere

5.

☐ IDF står for “Information of Document Frequency”

☐ IDF står for “Invariant Document Frequency” og brukes til å måle hvor mye svingninger er det i antall termer per dokument

☐ IDF står for Inverse Document Frequency og kan brukes til å straffe termer som nevnes ofte i et dokument

☐ IDF står for Inverse Document Frequency og kan brukes til å straffe termer som nevnes ofte i en samling av dokumenter

6.

☐ Zipf's law brukes til å vurdere hvilke ord som kan være gode kandidater til stoppordlista

☐ Zipf's og Heap's law er viktig for genfinningsregler i websøkesystemer

☐ En crawler bruker Zipf's law til å avgjøre om den skal følge en link videre til neste dokument, for indeksering

☐ Zipf's law sier ingenting om termdistribusjon i en samling

7.

☐ Harverst websøkesystem er basert på et sentralisertsystemarkitektur med koordinerte crawlers

☐ Harverst websøkesystem er basert på et distribuertsystemarkitektur med koordinerte gatherers

☐ Harverst er ingen websøkesystem men en proprietær systemarkitektur for datahøsting

☐ Harverst-arkitekturen er helt lik den arkitekturen som Google har brukt som basis for deres websøkearkitektur

8.

☐ Et bildegjenfinningssystem består av en "feature extractor"-del som gjør det enkelt å automatisere lagring av informasjon om innholdet i et bilde

☐ Et bildegjenfinningssystem må bruke alle teknikker fra databaseteknikk for å kunne enkelt utføre spørringer

☐ Det er et krav at et bildegjenfinningssystem tilbyr støtte til nøkkelordsøk

☐ De fleste bildegjenfinningssystem bruker piksel-til-piskelsammenlikninger til å finne likhet mellom bilder og til rangering

9.

☐ Okapi BM25 er en variant av sannsynlighetsmodellen men bruker både TF og IDF til å estimere sannsynlighetene

☐ Okapi BM25 er en variant av språkmodellen (Language model for information retrieval)

☐ Ifølge forskningen fungere Okapi BM25 mye dårligere enn boolskmodellen

☐ Okapi BM25 har ingenting med søkeresultatrangeringer å gjøre

10.

- ☐ Precision og recall er like viktige uavhengig av søkeapplikasjoner
- ☐ Recall er typisk viktigere enn precision for søk i Gulesider
- ☐ MAP bruker ikke recall i det hele tatt
- ☐ Interpolering er nyttig dersom man har for få recall-punkter