Norges teknisk-naturvitenskapelige universitet
Institutt for datateknikk og informasjonsvitenskap

EKSAMENSOPPGAVE I FAG TDT4117 – INFORMASJONSGJENFINNING (BOKMÅL)

SENSURVEILEDNING

Faglig kontakt under eksamen: Heri Ramampiaro

Tlf.: 93440

Eksamensdato: ***21. desember 2012***

Eksamenstid / varighet: *09.00-13.00 / 4 timer*

Tillatte hjelpemiddel: *D: Ingen trykte eller håndskrevne hjelpemiddel tillatt. Bestemt, enkel kalkulator tillatt.*

Språkform: Bokmål

Sensurdato: 18. januar 2013

Det ønskes **korte** og **konsise** svar på hver av oppgavene.
Les oppgaveteksten meget nøye og vurder hva det spørres etter i hver enkelt oppgave.

Begrunn svar på alle oppgaver.

Dersom du mener at opplysninger mangler i oppgaveformuleringene, beskriv de antagelsene du gjør.

NorwayOil ANS er et stort oljeselskap med noen tusen ansatte. Hver ansatte produserer så mange dokumenter og data fra virksomheten at de nå har bestemt seg for å leie inn en konsulent til å hjelpe dem med organisering og søk av all materiale de har. Disse materialene består av et stort lager med tekstdokumenter, samt multimedia dokumenter som seismiske data i form av bilder som alle ligger i en server. Anta at du er den personen NorwayOil vil ha som konsulent og din oppgave vil være å tilrettelegge å sette opp et system for selskapet som skal lette tilgangen til informasjonen.

I denne oppgaven vil vi bruke *IR* for å forkorte informasjonsgjenfinning.

Oppgave I (30%)

I denne oppgaven skal vi først fokusere bare på de delene av *tekstdokumentene* som blir produsert.

a) På din første dag på jobb begynner du å undersøke hvilke generelle behov bedriften har i forhold til søk og gjenfinning av dokumenter. En del av dette er å finne ut hvilke typer søk som blir utført. Svaret du får er "vi vet ikke". Så du bestemmer deg for å forklare dem 2 alternativer "data gjenfinning" og "informasjonsgjenfinning". Ved å fokusere på karakteristikkene i disse to hvordan ville du forklare alternativene?

Svar: Her skal det forklares de viktigste karakteristikkene som kjennetegnes alternativene som full matching vs. delvis match, rangering vs. treff eller ikke treff, og feil toleranse og ikke. **(5%)**

Anta nå at ledelsen går for informasjonsgjenfinning og du får i oppgave å lage et IR-system.

b) Ledelsen vil spare penger og ønsker å velge en "open source" løsning. Du anbefaler Lucene. Forklar hvorfor Lucene ville være egnet til dette.

Svar: Her forventes at man forklarer at Lucene støtter rask indeksering og søk, standard tekstoperasjoner som leksikalanalyse via tokenizing, fjerning av stoppord og evt. stemming. **(4%)**

c) Før du i det hele tatt skal kunne indeksere data i systemet må dokumentene behandles med såkalte tekstoperasjoner. Flere av dem er viktigere enn andre. Forklar hvilken operasjon du mener bør minst være utført før du skal utføre selve indekseringsprosessen.

Svar: Det er fem operasjoner som vi har behandlet i kurset. Her kan studenten velge en av disse og begrunne hvorfor den er viktig. Jeg ville selv valgt fjerning stoppord her for å redusere indekstørrelsen og for å øke diskrimineringsgraden. **(5%)**

d) Det neste du skal velge er likhetsmodell ("Similarity model"). Forklar kort hvorfor du *ikke* vil gå, eller evt. vil gå for den bolske modellen ("Boolean similarity model").

Svar: Valgene **må** begrunnes. Her forventes at man forklarer fordeler og ulemper med den bolske modellen. **(7%)**

e) Anta at du til slutt landet på språkmodellen ("Language Model") som likhetsmodell. Forklar forskjellene på denne modellen og den sannsynlighetsbaserte modellen ("Probabilistic similarity model")?

Svar: Den største forskjellene mellom LM og PM er hvordan sannsynligheten blir beregnet. Mens PM fokuserer mye på sannsynlighet for relevans, er fokuset på LM sannsynlighet for en gitt dokumentmodell generer spørringen man har. **(9%)**

Oppgave II (25%)

I denne oppgaven skal du fortsatt fokusere på den tekstlige delen av bedriftens dokumentlager.

a) For at ledelsen skal kunne forstå hva du snakker om ønsker du å forklare dem en del typiske informasjonsgjenfinningsbegrep. Følgende skal du få dem til å forstå - dvs. du skal definere følgende begrep:

1. Indeksterm (”Index term”).
   Svar: Indeksterm er term som blir valgt til å representere et dokument. Man velger i hovedsak termer som mest mulig sier noe om innholdet i dokumentet og de som mest bidrar til høy diskirmineringsgrad (de som får dokumentet til å skille seg mest mulig fra andre dokumenter). **(2%)**

2. Mean Reciprocal Rank (MRR).
   Svar: Er et IR evaluaringsmål hvor man i hovedsak er interessert i å finne ut hvor høyt i rangeringsliste for et søkeresultat et relevant treff befinner seg. **(2%)**

3. Cosinus likhetsfunksjon (Cosine similarity function).
   Svar: Det er likhetsfunksjonen som brukes av vektor space modellen til å finne likhet mellom spørring og et dokument basert på cosinus-verdien av vinkelen mellom vektorene av vektene for disse to. **(2%)**

4. Scalar Clusters.
   Svar: Dette brukes i Local Automatic Analyse til å finne ut relasjoner mellom termer, basert på **naboskap**, for å bygge et thesaurus. Hovedideen er at jo mer 2 termer opptrer sammen i forskjellige dokumenter, desto mer er de relaterte til hverandre. **(2%)**

5. Presisjon og recall.
   Svar: Presision: Andelen av gjenfunnet dokumenter som er relevante. Recall: Andelen av relevante dokumenter totalt som er gjenfunnet. **(2%)**

b) Du angrer litt på at du ikke prøvde de forskjellige likhetsmodellene før du anbefalte språkmodellen i Oppgave I. Du bestemmer deg derfor å gjøre noe lurt og anbefaler å evaluere forskjellige IR-systemer basert på forskjellige likhetsmodeller.

1. Forklar først hvorfor er det lurt med å evaluere et IR-system.
   Svar: Hensikten med evaluering av et IR-system er for å vurdere dets evne til å oppfylle brukerens informasjonsbehov i for av hvor godt det klarer å finne relevante dokumenter. **(2%)**

2. Forklar kort minst fire evalueringmål (evaluation measures) du vil bruke.
   Svar: Her kan man forklare MAP, F-Measure, R-precision, P@n. Det er også ok å bruke precision og recall. **(3%)**

3. TREC brukes ofte i evalueringer av et IR-system eller søkealgoritmer. Hva er TREC, og hva får du ved å bruke TREC?
   Svar: TREC står for Text REtrieval Conference og har som mål å tilby test samlinger og infrastruktur for evaluering av IR-systemer og algoritmer. Ved å bruke TREC har man muligheter til å vurdere hvor bra et system er ved å bruke deres sett med spørringer og ”fasit” og standard evalueringsmetrikker. Her har

man også muligheten til sammenlikne mot andre systemer som bruker de samme datasettene. **(4%)**

4. Forklar hvorfor du vil ved å bruke thesaurus, som for eksempel synonymer, oppnå høyere recall men ikke nødvendigvis høyere presisjon.

Svar: Ved å bruke thesaurus vil man finne dokumenter som inneholder også synonymene og ikke nødvendigvis bare de ordene man har i spørringen. Med dette vi antall treff øke noe som kan bidra til økt recall. Dokumentene med synonymene er dog ikke nødvendigvis relevant og dermed vil presisjon bli lavere dersom dette er tilfelle. **(3%)**

5. Anta at i et system fikk du i gjennomsnitt en presisjon på 80 % og en recall på 60 %. Hva blir verdien av f-measure (”Harmonic mean”)?

Svar: f=2PR/(P+R) = 2*80*60 / (80 +60) = 68.57**(3%)**

Oppgave III (25%)

Anta nå at du er klar til å indekser dokumentene dine.

a) Forklar minst 3 forskjellige indekseringsmetoder som det går an å bruke til å indeksere dokumenter. Her skal du forklare prinsippet bak metodene.

Svar: Her kan studentene forklare Invertert indeks, Suffix tre, og Signaturfil. NB: Suffix Tre og Trie går ikke som 2 forskjellige indekseringsmetoder. **(8%)**

b) I et av Norway Oils dokumenter står følgende:
”New seismic tools will result in 30 million extra barrels of oil from Snorre and Grane when Norway Oil and its partners now start using permanent reservoir monitoring”.

1. Hva er suffiksstrengene for teksten over?

Svar: Her holder det at studenten genererer strengene. Det er dog viktig at de forklarer hva de gjør.
Anta at vi har fjernet alle stoppord.

5: seismic … monitoring

13: tools … monitoring

.

.

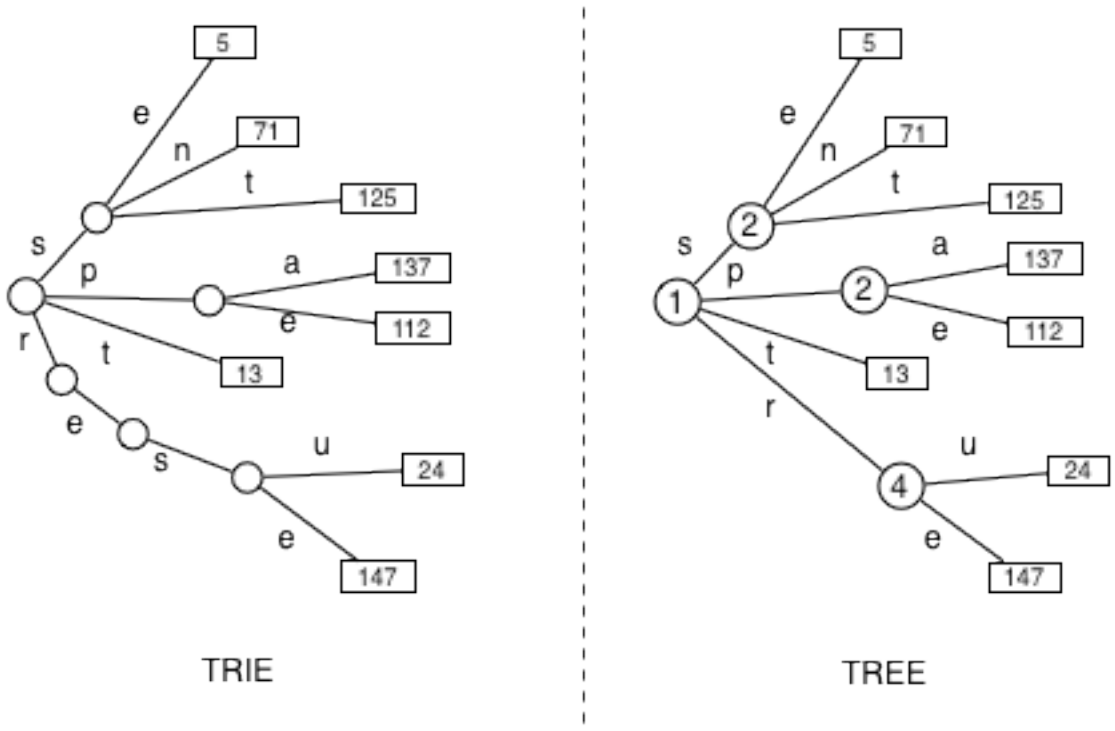
.

157: monitoring

(4%)

2. Konstruer ”suffix trie” og deretter ”suffix tree” for teksten over.

Svar: Deler av treene kan se ut som følgende **(7%)**



3. Hvar fordelene med "supra index" sammenliknet med "suffix array"?

Svar: Med supra index tar man i utgangspunktet en suffix array og deler den i blokker av et gitt antall lengder *b*. I supra index'en lager man pekere av prefixer av fast lendge *k* for hvert ord der prefisen endrer seg, og lar denne peke på starten av en blokk bestående av n ord. Fordelen med dette er at man reduserer eksterne aksesser siden denne blir lagret i hovedminnet og dermed øker ytelsen (i forhold til en vanlig suffix array). **(6%)**

Oppgave IV (20%)

a) I likhet med andre bedrifter, brukes Web-søk også til finne andre informasjon som er relevant for Norway Oil. Du mener det systemet, dvs. det sentraliserte websøkesystemet, som de fleste bruker i dag er litt oppskrytt. Du mener det søkesystemet med distribuerte arkitektur kunne være bedre. Forklar hvorfor du mener dette.

Svar: Her leges det mest vekt på fordelene med distribuerte web-søkesystemer og ikke generelle distribuerte systemer:

- Man kan fordele viktige oppgaver mellom "brokers" og dermed minker belastningene på serveren generelt.
- Man kan domenespesifikke brokers som da igjen lar oss lage indekser basert på domener (applikasjonsområder, tema, og lignende). **(8%)**

b) Du fikk en ide om at du kan forbedre søkeresultatene ved å gjøre spørringene som sendes til systemet bedre. Det du lander på er å bruke enten "Standard Rocchio" eller "Ide Regular". Hva er hovedideene bak disse? (Tips: fokuser på formål). Hva skiller disse to seg fra hverandre? (6%)

Svar: Formålene med begge er søkeforbedring gjennom utvidelse av spørringer når man bruker av vector space model (VSM) likhetsmodell. De to metodene er ganske like bortsett fra Standard Rocchio bruker antall

relevante dokumenter og antall irrelevante dokumenter fra svarsettet til å normalisere relevant- og ikke relevantleddet, mens Ide Regular ikke normaliserer disse leddene i det hele tatt. **(6%)**

c) Lag er *kanonisk huffman kode* av følgende tekst:
”for oil drilling, a tool is a tool”.

Svar: Siden dette er et IR-fag er hovedfokuset er ordbasertkomprimering og **ikke** tegnbasert.

Ordsannsynlighetene er som følgende:

for: 1/9

oil: 1/9

drilling: 1/9

,: 1/9

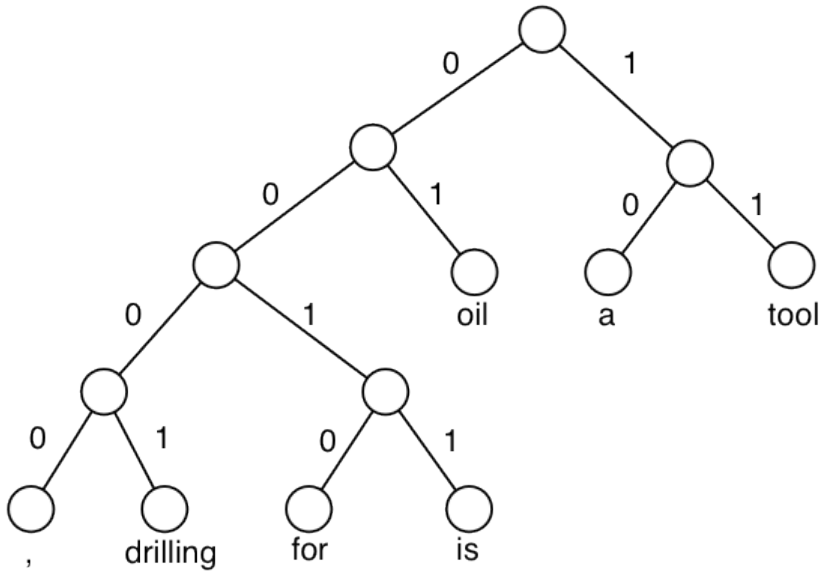
a: 2/9

is: 1/9

tool : 2/9

Kanonisk skiller seg fra det vanlige treet på den måten at ingen løver med lavere sannsynlighet kan være på høyere side av en node.

Dette gir oss treet:



Huffmankoden for teksten blir da: 0010 01 0001 0000 10 11 0011 10 11

(6%)