Forside

EKSAMENSOPPGAVE I FAG TDT4117 – INFORMASJONSGJENFINNING

Faglig kontakt under eksamen: Heri Ramampiaro

Telefon: 99027656

Eksamensdato: 12.12.2020

Eksamenstid / varighet: 09.00-13.00 / 4 timer

Løsningsskisse

Oppgave I (10%)

1. Drøft kort hvilke kriterier du ville legge til grunn for valg av indekstermer.
   Svar: kriteriene som skal forklares kort her er bl.a.
   (1) Hvor bra termen representerer innholdet.
   (2) Diskrimineringsgrad, dvs. hvordan termen gjør dokumentet unikt sammenliknet med andre dokumenter, og dermed kan øke sannsynligheten for høyre presisjon.
   (3) Hvordan termen bidrar til å dekke temaet for dominerer søkedomenet (feks. medisin, litteratur, finans, etc.)

2. Du blir bedt om å bygge et IR-system for Finn. Hovedideene er bl.a. å gjøre søkefunksjonen smartere enn det de har i dag med hensyn til rangering, spesielt. Svar på følgende spørsmål basert på dette. *Gjør de antakelsene du finner nødvendige.*
   a. Hvorfor er søk på Finn i hovedsak en informasjonsgjenfinning og ikke datagjenfinning?

   Svar: her forventes at man lister opp karakteristikkenen for IR og bruker disse som utgangspunkt for svaret.

   b. Tegn og forklar er arkitektur på ditt Finn-IR-system.

   Svar: her skal studentene forventes å tegne opp et arkitektur tilsvarende figur 1.3 i læreboka, men er tilpasset Finn.

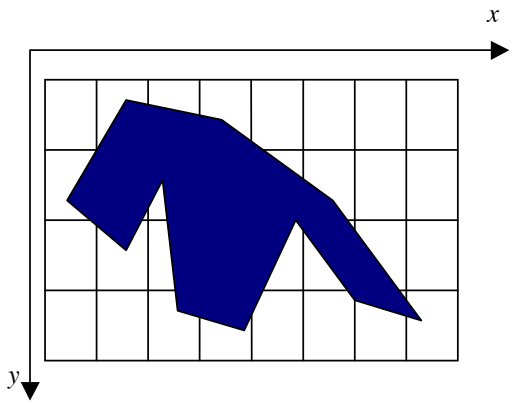
   c. Hvilke tre alternative likhetsmodeller (similarity models) ville du ha valgt for å få til rangeringen av søkeresultatene? Gjør kort rede for hvilken av disse du ville valgt selv.
   Svar: Alternative similaritetsmodeller ville være feks. VSM, Language model, og BM25. Alle disse kan argumenteres med at de tilbyr rangering og kan bruke TF og IDF i løsningen... Det viktigste er å få frem modeller som tilbyr rangering.

Oppgave II (20%)

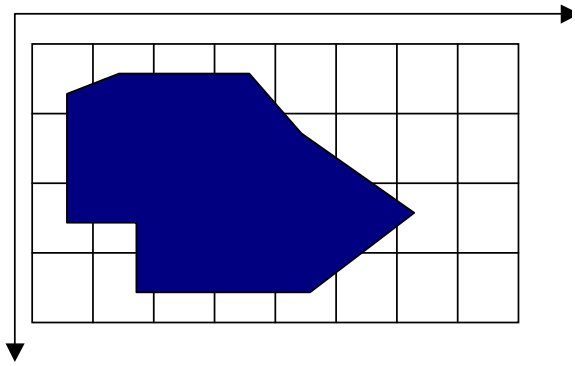
1. Drøft hvordan bildegjenfinning kan gjøres mulig ved hjelp av tekstgjenfinning.
   Svar: Bilder kan annoteres med tekst for å beskrive innholdet. Denne tekstlige beskrivelsen kan brukes i et tekstgjenfinnginssystem.

2. Innen multimedia er begrepet «features» brukt. Fargehistogram er en type *feature* som brukes til bildegjenfinning. Hvilke utfordringer eller begrensninger har histogram som feature? Begrunn svaret ditt.
   Svar: Største utfordringen med fargehistogram er at det ikke klarer å beskrive innholdet, semantisk. Videre er et standard histogram ikke bra nok til å håndtere fargenyanser, noe som igjen påvirker gjenfinningskapabilitet, og presisjon.

3. Anta at vi har følgene figurer.



Figur 1


Figur 2

   a. Hva blir er de binære sekvensene for formene (shapes) i figurene 1 og 2?

Svar: Alle ruter som er dekkes skal være 1, 0 ellers. 8 ruter pr. rad betyr 8 bits pr. gruppe, og vi har 4 grupper. Dermed får vi:
B1: 01111000 11111100 11111110 00111111
B2: 11111000 11111100 11111110 01111100

   b. Hva blir avstanden mellom figur 1 og figur 2?

Svar: Avstand d(B1, B2) er lik ant. bits som er forskjellige.
B1: 01111000 11111100 11111110 00111111
B2: 11111000 11111100 11111110 01111100

De bets-ene som er røde er forskjellige.
=> d (B1, B2) = 4

   c. Fra informasjonsgjeninningsståsted hva er begrensningene med denne metoden for å finne avstand? Forklar.
   Svar: Problemet med denne metoden er at denne tar kun hensyn til å hvilke områder som er dekket av figurene, ikke formene i seg selv. To helt forskjellige former kan fort få avstand lik 0 noe som vil føre til bl.a. dårlig presisjon. Rangering vil også være misvisende med tanke på relevans.

Oppgave II 30%

Anta at vi har følgende dokumenter:

d1 = ” India says may not need to vaccinate entire population to control COVID”

d2 = ” Fake news about a Covid vaccine has become a second pandemic”

Søkespørsmål:

q = ”covid vaccine”

Til følgende spørsmål skal du anta at du gjør leksikalanlyse, fjerner stoppordene og kjøre stemming først. Du kan i tillegg gjøre andre antakelser som du finner nødvendige.

1. Konstruer rangert liste over resultatet av spørringen q basert på ***vektormodellen*** og ved hjelp av følgende formell:

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

   Svar: Før man svarer på spørsmålet må man konstruerer K som er sett av alle termer i samlingen. Med dokumentene over, samt antakelsen over. Antar også at hjelpeverber (som may og can etc.) og adverb er en del av stoppordlista. Vi får da

   K ={become, control, covid, fake, india, pandemic, populat, new, second, vaccin}.

   Med K over, og antar at vi kun bruker rå-frekvens som vekt, får vi følgdende vektorer for dokumentene og spørringen:
   v1 = (0, 1, 1, 0, 0, 0, 1, 0, 0, 1)
   v2 = (1, 0, 1, 1, 1, 1, 0, 1, 1, 1)

   q = (0, 0, 1, 0, 0, 0, 0, 0, 0, 1)

   sim(q, d1) = [(0, 0, 1, 0, 0, 0, 0, 0, 0, 1) * (0, 1, 1, 0, 0, 0, 1, 0, 0, 1)] / [sqrt(1^2+1^2) * sqrt (1^2 + 1^2 +1^2 +1^2)] = 2/[2*sqrt(2)]= 1/sqrt(2) = 0.70

   sim(q, d1) = [(0, 0, 1, 0, 0, 0, 0, 0, 0, 1) * (1, 0, 1, 1, 1, 1, 0, 1, 1, 1)] / [sqrt(2) * sqrt(8)] = 2/4 = 0.50

   Så: svaret blir d1 rangeres høyre enn d2.

2. Konstruer rangert liste over resultatet av spørringen q basert på ***språkmodellen***. Bruk følgende formell som utgangspunkt, hvor $\lambda = ½$:

$$p(Q,d) = p(d)\prod_{t \in Q}((1-\lambda)p(t) + \lambda p(t \mid M_d))$$

Svar: Her får vi følgende.

p(d1) = p(d2) = 1/2 (2 dokumenter i samlingen og sannsynligheten for å hente ut hvert dokument er da lik 1/2).

Lamba = 0.5.

Vi er på jakt etter sannsynlighet for et dokument genererer spørringen som er «covid vaccin» (med stemming)

p(q, d1) = 0.5 * [(1-0.5)p('covid') + 0.5p('covid'| M_{d1})]* [(1-0.5)p('vaccin') + 0.5p('vaccin'| M_{d1})] = 0.5*[0.5*2/23 + 0.5*1/12]*[0.5*2/23+0.5*1/12]=0.5*0.25*0.029=0.0036.

p(q, d2) = 0.5 * [(1-0.5)p('covid') + 0.5p('covid'| M_{d2})]* [(1-0.5)p('vaccin') + 0.5p('vaccin'| M_{d2})] = 0.5*[0.5*2/23 + 0.5*1/11]*[0.5*2/23+0.5*1/11] = 0.5*0.25*0.031 = 0.0039.

Så: Her vil d2 rangeres før d1.

Oppgave III (20%)

Gitt følgende tekst:
«*CDC's team of advisers set to decide who gets coronavirus vaccine first*».

Gjør de antakelsene du finner nødvendige og svar på følgende spørsmål:

Anta at man ikke tar med stoppord som typisk vil være artikler etc.

Nødvendig første steg til spm. 1 og 2: laging av suffix-strengene med posisjon:

1: CDC's team of advisers set to decide who gets coronavirus vaccine first

7: team of advisers set to decide who gets coronavirus vaccine first

15: advisers set to decide who gets coronavirus vaccine first

24: set to decide who gets coronavirus vaccine first

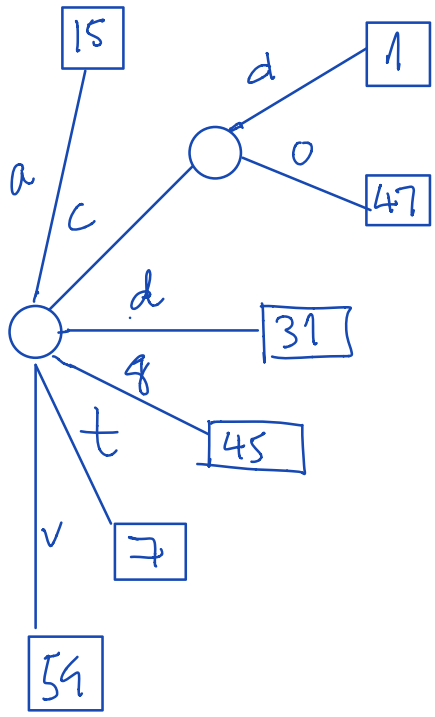
31: decide who gets coronavirus vaccine first

45: gets coronavirus vaccine first

47: coronavirus vaccine first

59: vaccine first

1. Tegn opp «**suffix trie**» basert på teksten over.
   Svar: bruker posisjonen fra suffix-strengene og plasserer de i alfabetisk rekkefølge på treet.



2. Hvordan ser «**suffix array**»-indeksen ut basert på teksten over?
   Svar: bruker posisjonen fra suffix-strengene og plasserer de i alfabetisk rekkefølge. Svar som ikke tar hensyn til alfabetisk sortering er feil.

3. Drøft kort hvor indekseringsmetoden, singaturfiler (Signature files) ikke er egnet for web-søk.
   Svar: Signatur-fil som indekseringsmetode er laget for indeksering av små og få dokumenter på grunn av hele ideen med signatur og hashkodegenerering. Hashkoder er vanskelig å lage for store dokumenter. Videre er ikke signaturfiler effektive nok for store dokumentsamlinger som web-en er. Alt dette gjør at signaturfiler ikke er egnet for web-søk.

Oppgave V (20%)

Ta utgangspunkt at en spørring q1 returnerer resultater som er vist i følgende tabell:

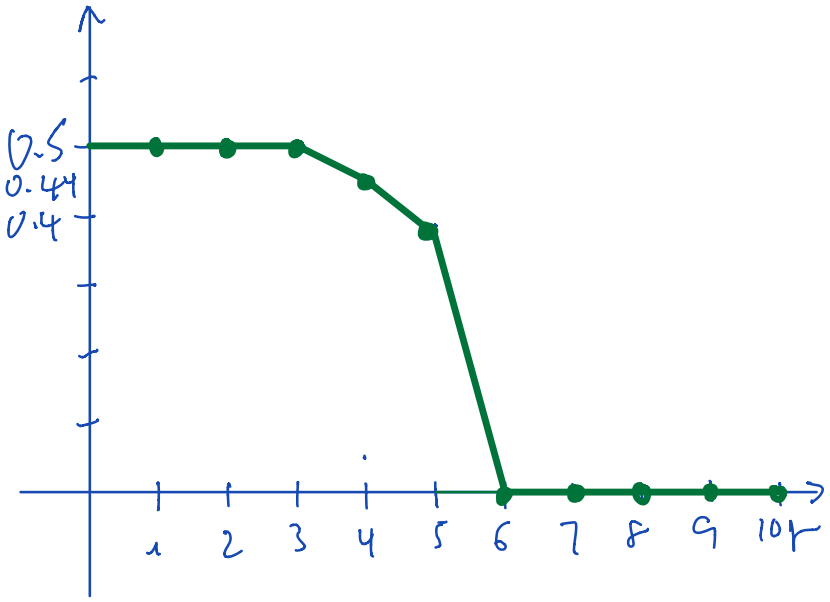
Rank	Doc ID	Relevant?
1	8	
2	9	REL
3	12	
4	5	REL
5	2	
6	17	REL
7	23	
8	10	
9	1	REL
10	4	
11	30	
12	3	
13	6	REL
14	13	

1. Vis hvordan du beregner precision- og recall-punkter for resultatet i tabellen over. Antar total ant. relevante dokumenter for spørringen er 10.
   Svar:

Rank	Doc ID	Relevant?	Precision	Recall
1	8			
2	9	REL	1/2	1/10
3	12			
4	5	REL	2/4	2/10
5	2			
6	17	REL	3/6	3/10
7	23			
8	10			
9	1	REL	4/9	4/10
10	4			
11	30			
12	3			
13	6	REL	5/13	5/10
14	13			

2. Anta at det er tre spørringer og to av disse har *average precision* (AvgP2, AvgP3) henholdsvis 0.6 og 0.5. Beregn AvgP1 for spørringen q1 og deretter regn ut **Mean Average Precision (MAP)**?
   Svar: AvgP1 = (1/2 + 2/4 + 3/6 + 4/9 + 5/13)/5 = **0.46**.
   MAP = (AvgP1 + AvgP2 + AvgP3)/3 = 0.46 + 0.6 + 0.5 = **0.52**.

3. Tegn opp grafen som viser de *interpolerte verdiene av precisions* Viktig at du forklarer fremgangsmåten du bruker.
   Svar:


0.5
0.44
0.4
1
2
3
4
5
6
7
8
9
10