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

Tillatte hjelpemiddel: A: Alle trykte og håndskrevne hjelpemidler tillatt. Alle kalkulatorer tillatt. Eksamen er å anse som *individuell oppgave*. Samarbeid er derfor *ikke* tillatt.

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

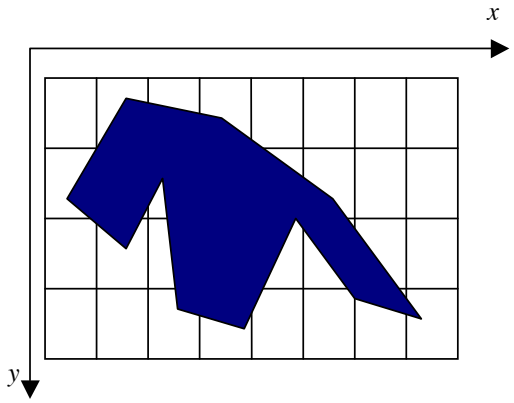
Dersom du mener at opplysninger mangler i oppgaveformuleringene, beskriv de antagelsene du gjør.

Oppgave I (10%)

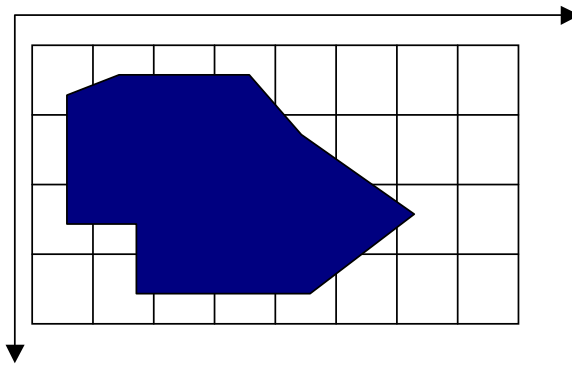
1. Drøft kort hvilke kriterier du ville legge til grunn for valg av indekstermer.
2. Du blir bedt om å bygge et IR-system for Finn. Hovedideene er bl.a. å gjøre søkefunksjonen smartere enn det de har i dag med hensyn til rangering, spesielt. Svar på følgende spørsmål basert på dette. *Gjør de antakelsene du finner nødvendige.*
    a. Hvorfor er søk på Finn i hovedsak en informasjonsgjenfinning og ikke datagjenfinning?
    b. Tegn og forklar er arkitektur på ditt Finn-IR-system.
    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.

Oppgave II (20%)

1. Drøft hvordan bildegjenfinning kan gjøres mulig ved hjelp av tekstgjenfinning.
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.
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?
b. Hva blir avstanden mellom figur 1 og figur 2?
c. Fra informasjonsgjeninningsståsted hva er begrensningene med denne metoden for å finne avstand? Forklar.

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:

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

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

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

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:

1. Tegn opp «**suffix trie**» basert på teksten over.
2. Hvordan ser «**suffix array**»-indeksen ut basert på teksten over?
3. Drøft kort hvor indekseringsmetoden, singaturfiler (Signature files) 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.
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)**?
3. Tegn opp grafen som viser de *interpolerte verdiene av precisions* Viktig at du forklarer fremgangsmåten du bruker.