NTNU
Kunnskap for en bedre verden

Institutt for datateknikk og informasjonsvitenskap (IDI)

# Eksamensoppgave i TDT4117 Informasjonsgjenfinning

**Faglig kontakt under eksamen: Heri Ramampiaro**
**Tlf.: 73591459**

**Eksamensdato: 07.12.2016**
**Eksamenstid (fra-til): 09:00-14:00 (4 timer)**
**Hjelpemiddelkode/Tillatte hjelpemidler: D – Kun godkjent kalkulator tillatt**

**Annen informasjon:**

**Målform/språk: Bokmål**
**Antall sider (uten forside): 2**
**Antall sider vedlegg: 0**

**Informasjon om trykking av eksamensoppgave**
**Originalen er:**

**1-sidig** ☐ **2-sidig** ☐

**sort/hvit** ☐ **farger** ☐

**skal ha flervalgskjema** ☐

**Kontrollert av:**

Dato Sign

*Svar **kort og konsist** på alle spørsmålene.*
*Les igjennom alle oppgavesett før du begynner å løse oppgavene. Forklar kort eventuelle antakelser der du mener oppgaveteksten ikke er fullstendig. Lykke til!*

Oppgave I (25%)

1. Her er en påstand: ”Den boolske likhetsmodellen (boolean similarity model) er egentlig ikke en informasjonsgjenfinningsmodell men en datagjenfinningsmodell”. Forklar hvorfor dette kan være sant.

2. ”Page Rank” er en rangeringsmetode som kan brukes til å rangere søkeresultat. Forklar når ”Page Rank” ikke kan brukes.

3. Forklar med eksempel og figur forskjellene mellom ”suffix trie” og ”vocabulary trie”.

4. Tegn et blokkdiagram (med firkanter og piler) som forklarer hvordan informasjonsgjenfinningsprosessen er bygd opp. Tips: Dette er ikke tekstoperasjoner.

5. To bilder med størrelse 4x4 med inneholder 4 forskjellige pikselfarger (C1, C2, C3, C4) fordelt på følgende måte:
   Bilde 1: 2 av C1, 3 av C2, 6 av C3 og 5 av C4.
   Bilde 2: 4 av C1, 1 av C2, 8 av C3 og 3 av C4.
   Finn forskjellen (eller avstanden) mellom de to bildene ved hjelp av bildenes *histogram*.

## Oppgave II (30%)

Anta vi har følgende tekst:

”Political differences can break family ties, but family heals”.

Gjør (og forklar) ellers de antakelsene du finner nødvendig når du svarer på følgende spørsmål.

1. Utfør tekstoperasjonene: leksikalanslyse (lexical analysis), stoppordfjerning, stemming. Hvilke liste av termer sitter man igjen med?

2. Konstruer en signaturfil av teksten over. Anta at du kan dele teksten i 3 blokker. Du må ellers gjøre dine egne antakelser angående **hash-funksjon**, og liknende.

3. Konstruer invertert-indeks (inverted index) av teksten over.

4. Lag et ”suffix array” av teksten over. Hvorfor er denne indekseringsmetoden mindre egnet enn ”supra index”?

## Oppgave III (25%)

Ola skal evaluere sitt informasjonsgjenfinningssystem (IR-system). Han tester forskjellige evalueringsmetrikker. Han setter en grense på 20 (rangerte) resultatdokumenter for hver spørring. Han har i alt 4 spørringer (q1 – q4) med fasit han skal bruke til evalueringen. Når han sjekker fasiten finner han relevante dokumenter på følgende plasser i resultatlista for hver spørring:

q1: 1, 3, 4, 5, 6, 10, 12, 15.
q2: 2, 4, 5, 12, 13.
q3: 1, 2, 4, 6, 13, 18.
q4: 1, 4, 7, 8, 9, 10, 14, 20.

For q1 er totalt antall relevante dokumenter 10. For q2 er dette 15. For q3 er det totalt 16 relevante dokumenter og for q4 er det 13.

1. Beregn presisjon (precision) og recall for hver spørring.
2. Hva blir R-precision for hver av spørringene.
3. Bruk resultatene til q1 til å illustrere hvordan du kan beregne precision- og recall-punktene.
4. Hva blir Mean Average Precision (MAP)-verdien for Ola sitt IR-system? NB: For å få poeng må du vise hvordan du kom fram til svaret ditt.

## Oppgave IV (20%)

Svar rett/galt med begrunnelse på følgende utsagn. Hvert **riktig** og **begrunnet** svar belønnes med **2** poeng. **Feilsvar**, **ubegrunnet** eller **ingen svar** gir ingen poeng.

1. Fargehistogrammer kan brukes i forbindelse med gjenfinning av både bilder og videosnutter.
   (Rett/Galt)
2. Google bruker ikke ”stemming” fordi stemming ikke passer til web-søk generelt.
   (Rett/Galt)
3. Søkemotorer med ”Harvest”-arkitektur er en variant av distribuert web-søkemotor arkitektur.
   (Rett/Galt)
4. Thesaurus-bygging er naturlig del i automatisk lokal analyse (automatic local analysis), og bruker hele dokumentsamlingen til å gjøre dette.
   (Rett/Galt)
5. Den største forskjellene mellom ”Probabilistic Similarity Model” og ”Language Model” er måten sannsynligheten blir beregnet.
   (Rett/Galt)
6. Hvis man ser på tekst som et multimediaobjekt ville indekstermene (index terms) være ”features”.
   (Rett/Galt)
7. En SQL-database er for datagjenfinning og derfor kan den ikke brukes til bildegjenfinning.
   (Rett/Galt)
8. Komprimering kan ikke alltid brukes i informasjonsgjenfinning siden dataene da alltid må dekorprimeres først.
   (Rett/Galt)
9. Treet for Huffaman-kode er en spesial-versjon av et indekskomprimeringstre.
   (Rett/Galt)
10. Videogjenfinning kan bruke ”features ” fra bildegjenfinning.
    (Rett/Galt)