

# □ TDT4305 vår 2021

Institutt for datateknologi og informatikk

Eksamensoppgave i TDT4305 Big Data Architecture

Eksamensdato: 26. mai 2021

Eksamenstid (fra-til): 09:00 – 13:00

Hjelpemiddelkode/Tillatte hjelpemidler: A / Alle hjelpemidler tillatt

Faglig kontakt under eksamen: Heri Ramampiaro

Tlf.: 99 02 76 56

Teknisk hjelp under eksamen: [NTNU Orakel](#)

Tlf: 73 59 16 00

Får du tekniske problemer underveis i eksamen, må du ta kontakt for teknisk hjelp snarest mulig, og senest innen eksamenstida løper ut. Kommer du ikke gjennom umiddelbart, hold linja til du får svar.

## ANNEN INFORMASJON

**Gjør dine egne antagelser** og presiser i besvarelsen hvilke forutsetninger du har lagt til grunn i tolkning/avgrensing av oppgaven. Faglig kontaktperson skal kun kontaktes dersom det er direkte feil eller mangler i oppgavesettet.

**Juks/plagiat:** Eksamen skal være et individuelt, selvstendig arbeid. Det er tillatt å bruke hjelpemidler, men vær obs på at du må følge eventuelle anvisningen om kildehenvisninger under. Under eksamen er det ikke tillatt å kommunisere med andre personer om oppgaven eller å distribuere utkast til svar. Slik kommunikasjon er å anse som juks.

Alle besvarelser blir kontrollert for plagiat. [Du kan lese mer om juks og plagiering på eksamen her.](#)

**Kildehenvisninger:** Alle oppgaver skal besvares "med egne ord" for å vise forståelse.

**Varslinger:** Hvis det oppstår behov for å gi beskjeder til kandidatene underveis i eksamen (f.eks. ved feil i oppgavesettet), vil dette bli gjort via varslinger i Inspira. Et varsel vil dukke opp som en dialogboks på skjermen i Inspira. Du kan finne igjen varselet ved å klikke på bjella øverst i høyre hjørne på skjermen. Det vil i tillegg bli sendt SMS til alle kandidater for å sikre at ingen går glipp av viktig informasjon. Ha mobiltelefonen din tilgjengelig.

**Vekting av oppgavene:** Som vist i oppgavesettet.

## OM LEVERING

**Slik svarer du på oppgavene:** Alle oppgaver som *ikke* er av typen filopplasting, skal besvares direkte i Inspira. I Inspira lagres svarene dine automatisk hvert 15. sekund.

NB! Klipp og lim fra andre programmer frarådes, da dette kan medføre at formatering og elementer (bilder, tabeller etc.) vil kunne gå tapt.

**Automatisk innlevering:** Besvarelsen din leveres automatisk når eksamenstida er ute og prøven stenger, forutsatt at minst én oppgave er besvart. Dette skjer selv om du ikke har klikket «Lever og gå tilbake til Dashboard» på siste side i oppgavesettet. Du kan gjenåpne og redigere besvarelsen din så lenge prøven er åpen. Dersom ingen oppgaver er besvart ved prøveslutt, blir ikke besvarelsen din levert. Dette vil anses som “ikke møtt” til eksamen.

**Trekk/avbrutt eksamen:** Blir du syk under eksamen, eller av andre grunner ønsker å levere blankt/avbryte eksamen, gå til “hamburgermenyen” i øvre høyre hjørne og velg «Lever blankt». Dette kan ikke angres selv om prøven fremdeles er åpen.













**Tilgang til besvarelse:** Du finner besvarelsen din i Arkiv etter at sluttida for eksamen er passert.

# **1 Opppg. 1 – Big Data-rammeverk – 15 % (lik vekting for a, b og c)**

- a) Diskuter i hvilken grad rammeverkene MapReduce, Spark, og Storm er egnet til sanntids-prosessering av innkommende data.
- b) Hva kjennetegner de av transformasjonene («transformations») i Spark som har høy kostnad, og hva er årsaken/årsakene til høy kostnad?
- c) Anta at du er ansvarlig for en Hadoop-klynge, bestående av 30 maskiner. Brukerne klager over at applikasjonene deres har lengre responstid enn før. Diskuter hva som kan være potensielle årsaker til dette, og hvor høy grad av sannsynlighet hver årsak har (høg/lav).

**Skriv ditt svar her**

Format ▾

**B** *I* U  $\times_e$   $\times^e$   $\mathcal{I}_x$             

Words: 0

Maks poeng: 15

## 2 Oppg. 2 – Shingles, minhashing, og LSH – 20 % (lik vekting på a og b)

Vi har i denne oppgaven fem sett ( $S_0, S_1, S_2, S_3, S_4$ ), og ønsker å bruke LSH for å finne like sett. Anta følgende tabell, som viser de fem settene og 2-shingler som er inneholdt i disse:








Row/x	k-shingle	$S_0$	$S_1$	$S_2$	$S_3$	$S_4$
0	aa	1	0	1	1	1
1	ab	1	1	0	0	0
2	ac	0	0	0	1	1
3	ba	1	1	0	0	0


Anta at vi ønsker å generere MinHash-signaturer med lengde 4, med permutasjoner gitt av fire hash-funksjoner  $h_1(x)$ ,  $h_2(x)$ ,  $h_3(x)$  og  $h_4(x)$ , som har følgende verdier for  $x$ :

x	$h_1(x)$	$h_2(x)$	$h_3(x)$	$h_4(x)$
0	1	3	0	2
1	3	1	2	0
2	2	0	3	1
3	0	2	1	3

- a) Hva blir signaturene for de 5 settene? Forklar hvordan du kommer frem til disse.
- b) Bruk LSH med to bånd og signaturene du nettopp har kommet frem til, for å finne sett med Jaccard-likhet  $\text{SIM}(S_i, S_j) = 1$ . Forklar hvordan du kommer frem til disse.

**Skriv ditt svar her**

Format ▾ | **B** *I* U  $\times_2$   $\times^2$  |  $\int_x$  |   |    |  $\frac{1}{2} =$   $\frac{3}{2} =$  |  $\Omega$   |  |  $\Sigma$  |



Words: 0

---

Maks poeng: 20

### 3 Oppg. 3 – Adwords – 15 % (5% på a, 10 % på b)

- a) Hvorfor kan «competitive ratio» ofte forventes å være mindre enn 1 for on-line-algoritmer?  
 b) Gitt følgende tabeller med 1) annonsører og deres bud på spørringer, og 2) annonsører og deres budsjett:








Annonsør	Spørring	Bud
a1	q1	1
a2	q2	0.5
a2	q3	0.5
a3	q2	1
a4	q1	0.75
a4	q2	0.5
a4	q4	0.5


Annonsør	Budsjett
a1	1
a2	3
a3	1
a4	2

Anta følgende spørringer, i gitt rekkefølge: q1, q2, q4, q3, q2, q2, q3, q2, q2

Finn annonsør-spørring-par ved å bruke «Balance»-algoritmen. Forklar hvordan du bruker algoritmen for å komme frem til resultatene, og hva som blir akkumulert inntekt.

**Skriv ditt svar her**

Format ▾ | **B** *I* U  $\times_2$   $\times^2$  |  $\int_x$  |   |    |  $\frac{1}{2} =$   $\frac{3}{2} =$  |  $\Omega$   |  |  $\Sigma$  |



Words: 0

---

Maks poeng: 15

#### 4 Opppg. 4 Systemer for datastrømmer/datastraumar (lik vekting for a, b, c og d)

a) Drøft hvorfor system som f.eks. AsterixDB Feeds, Spark eller Storm er nødvendige for at håndtering av datastrøm skal være mulig. Forklar deretter hvilke fordeler og ulemper hver av disse nevnte systemene har.

b) Tegn og forklar *systemarkitekturen* til Storm (Tips: ikke topologi).

c) Forklar hva som menes med «Delivery Semantics (Message Guarantees)» for datastrøm. Bruk eksempler til å støtte forklaringen din.

d) Vi skiller mellom «soft failure» og «hard failure» når vi snakker om feiltoleranse for et system for datastrøm. Forklar forskjellene mellom disse.

**Skriv ditt svar her**

Format

**B**

*I*

U

$x_2$

$x^2$

$I_x$

Words: 0

Maks poeng: 15



## 5 Oppgave 5 Håndtering/handsaming av datastrøm/datastraum (lik vekting for a, b og c)

For å overvåke uønskede aktiviteter bruker sosialemedia-plattformer som Instagram og Twitter algoritmer som analyserer bruksmønstre, feks. teller hvor ofte brukerne trykker på likes eller kommenterer på en post/melding/bilde i løpet av en gitt periode. Dette gjør de for å avgjøre om disse brukerne er reelle brukere eller «bots», og deretter evt. straffe dem med utestengelse. For eksempel utestenger Instagram ofte brukerne sine midlertidig hvis de kommenterer på for mange poster i løpet av en time eller i løpet av et døgn, der målet er å rense plattformen for spammere.

Anta at du skal hjelpe Instagram med oppgaven over, dvs. å finne ut hvilke brukere som kan være bots. *Gjør ellers de antakelsene som du mener er nødvendige* og svar på følgende spørsmål:

- a) Instagram-oppgaven kan løses ved hjelp av både stående spørring (standing queries) og ad-hoc spørring. Forklar hvordan.
- b) Skisser en løsning for Instagram basert på glidende vinduer (sliding windows).
- c) Kan Flajolet-Martin-algoritmen brukes her til å estimere antall «likes» til en bruker? Forklar.

**Skriv ditt svar her**

Format
|
B
I
U
x<sub>2</sub>
x<sup>2</sup>
I<sub>x</sub>
|
📄
📋
|
↶
↷
□
|
☰
☷
|
Ω
🔢
|
□
|
Σ
|

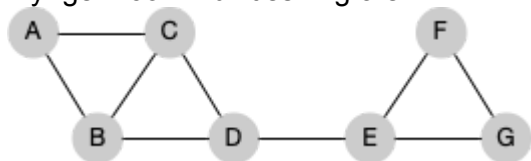
□

Words: 0

Maks poeng: 15

## 6 Oppgave 6 Sosiale grafer og anbefalingssystemer (lik vekting for a, b og c)

a) Gitt følgende forenklet sosiale graf. Bruk Girvan-Newman-metoden til å finne klynger/«communities» i grafen.



b) Vi velger ofte item-item collaborative filtering fremfor user-user collaborative filtering fordi den er mer effektiv. Er du enig? Hvorfor/hvorfor ikke?

c) Det finnes tre likhetsmål (similarity measures) som vi kan bruke for å sammenlikne brukere eller produkt i forbindelse med anbefaling av et produkt til en bruker. Drøft hvilke likhetsmål disse er. Forklar kort hvilket av disse likhetsmålene er minst egnet til bruk til anbefaling. Bruk et konkret eksempel til å støtte forklaringen din.

**Skriv ditt svar her**

Format
|
**B**
*I*
U
 $x_2$ 
 $x^2$ 
 $I_x$ 
|


|


|


|


|

|
Σ
|

□

Words: 0

Maks poeng: 20