Framside

Department of Computer Science

Examination paper for TDT4305 Big Data Architecture

Academic contact during examination: Kjetil Nørvåg and Heri Ramampiaro
Phone: 41440433 and 99027656

Examination date: May 28th

Examination time (from-to): 1500-1900

Permitted examination support material: D: No tools allowed except approved simple calculator.

Other information:

Students will find the examination results in Studentweb. Please contact the department if you have questions about your results. The Examinations Office will not be able to answer this.

1 Problem 1 – Hadoop – 10 % (all having same weight)

1. Explain two techniques that together make fault tolerance on data nodes (DataNodes) unnecessary.
2. Explain *combiner* in MapReduce, and what is the purpose of this.
3. What is the purpose of using data pipelines when writing blocks in HDFS?

Fill in your answer here

Maximum marks: 10

2 Problem 2 – Spark – 10 % (all having same weight)

Tuddel is a subscription based streaming service for music. Information about all songs that are streamed is stored in a log, in order to be able to analyze, perform recommendations, and calculate royalties to artists. For each song streamed, a line will be generated in the log, in the following format (comma separated):
TimeStamp,UserName,Artist,SongName

Assume the following example dataset stored in the file streamed.csv:

1,u1,a1,s1
2,u2,a1,s2
3,u1,a1,s2
4,u3,a1,s1
5,u1,a2,s3
6,u2,a1,s2
7,u2,a1,s2
8,u2,a1,s2

This dataset has already been loaded into an RDD named s:
val s = sc.textFile("streamed.csv").map(_.split(","))

For each of the subproblems below, you should show how they can be solved using Spark transformations/actions (Scala, Python or Java).

1. Create an RDD containing number of songs streamed for each artist. Example results:
   (a2,1)
   (a1,7)

2. Create an RDD that for each line has name of user and number of song he/she has streamed. Example results:
   u3 1
   u2 4
   u1 3
3. Find number of *distinct* songs that have been played. Example results:
   3

Fill in your answer here

Maximum marks: 10

3 Problem 3 – NoSQL – 15 %

1. Explain briefly 4 categories of NoSQL systems.
2. Explain how scalability is achieved in Voldemort.

Fill in your answer here

Maximum marks: 15

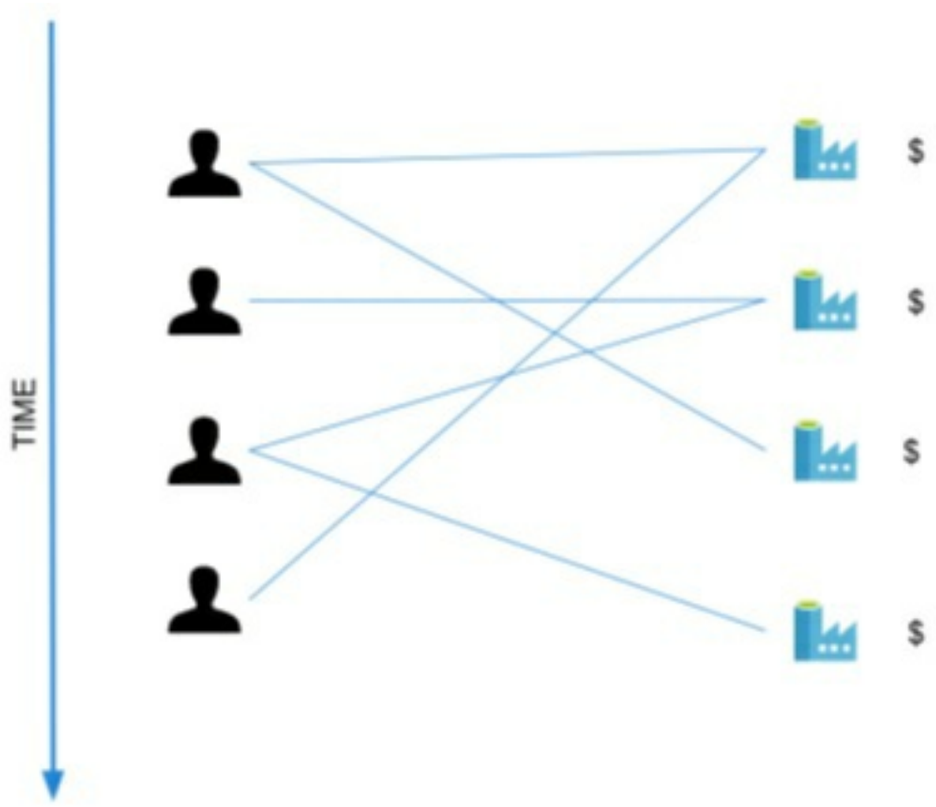
4 Problem 4 – MinHashing – 10 %

Explain the main features of LSH (locality sensitive hashing) for documents, incl.:
1. The purpose of LSH.
2. Input and output.
3. Contents of important data structure(s), and algorithm.

Fill in your answer here

Maximum marks: 10

5 Oppgave 5 – Adwords – 5 %



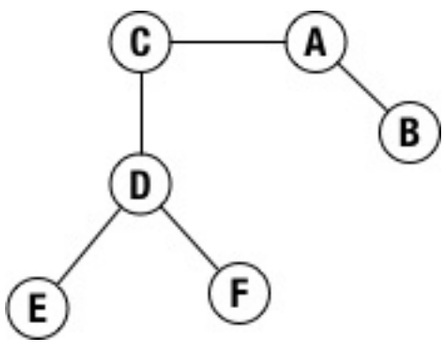
Given the figure above, explain the Adwords problem and a Greedy algorithm for solving this problem. What is the disadvantage of using this algorithm versus an optimal solution? A minor change to the greedy algorithm can improve results. Explain this change.

Fill in your answer here

Maximum marks: 5

6 Oppgave 6 – 15%

1. Explain briefly what steps you need to execute to find "communities" in a graph. Use the following figure to support your explanation. (6%)



2. Explain the main purpose of analyzing social graphs. (4%)
3. Explain the main differences between **Storm** and **Spark**. Explain the advantages of the **AsterixDB**'s Feed System has compared to both **Storm** and **Spark**. (5%)

Fill in your answer here

Maximum marks: 15

7 Oppgave 7 – 20%

1. a. What is "Bloom Filter" and what is it used for? Use an example to support your explanation. (4%)

   b. Use Bloom filter to complete the following table. Assume that we use h as a hash-function, and that it is defined as h(x)= y mod 11, where y is extracted from the odd number bits from xor even number bits from x.
   E.g., h1(39) = 011 = 3,
   h2(39) = 101 = 5, osv. (6%)

Strømelement	Hash-funksjon - h_1	Hash-funksjon - h_2	Filtrere Innhold
			000 0000 0000
85 = **101 0101**			
214 = **1111 1010**			
353 = **01 0110 0011**			

2. Assume that you will find fractions of unique queries from the last month in a stream of search queries. Due to space and time limitation, you cannot check all incoming queries and must use sampling and decide to read only every 10th of the questions (i.e. 10% sampling). Explain why this will not give the correct answer? What would be a more correct approach to sample the stream of search queries? (5%)

3. Imagine you are using social media like Twitter to analyze, e.g., trends. In particular, you are interested in knowing when a product is mentioned by people and how often they are mentioned. You propose using the bucket principle to count how many times this product is mentioned. Explain how the bucket principle on data stream works in this case. (5%)

Fill in your answer here

Maximum marks: 20

8 Oppgave 8 – 15%

1. So-called cold start is a challenge when using collaborative filtering in recommended systems? Explain what is meant by cold start and when this can occur. Further, discuss possible solutions for the cold start problem. (6%)

2. Assume the following user rating table:

USERS

PRODUCTS	1	2	3	4	5	6	7	8
1	1		3			5		
2			5	4			4	
3	2	4		1	2		3	
4		2	4		5			4
5			4	3	4	2		
6	1	P	3		3			2

Assume further that you will use "item-item collaborative filtering".

a. Explain what is the difference between item-item collaborative filtering and user-item collaborative filtering. (4%)

b. Explain how you will proceed to compute the predicted / estimated value of P. Make the assumptions you find necessary. (5%)

Fill in your answer here

Maximum marks: 15

Framside

Institutt for datateknologi og informatikk

Eksamensoppgåve i TDT4305 Big Data-arkitektur

Fagleg kontakt under eksamen: Kjetil Nørvåg og Heri Ramampiaro
Tlf.: 41440433 og 99027656
Eksamensdato: 28. mai 2018
Eksamenstid (fra-til): 1500-1900
Hjelpemiddelkode/Tillatne hjelpemiddel: D: Ingen trykte eller handskrivne hjelpemiddel tilletne. Bestemt, enkel kalkulator tillate.
Annan informasjon:

Merk! Studentane finn sensur i Studentweb. Har du spørsmål om sensuren må du kontakte instituttet ditt. Eksamenskontoret vil ikkje kunne svare på slike spørsmål.

1 Problem 1 – Hadoop – 10 % (all having same weight)

1. Forklar to teknikkar som tilsaman gjer at det ikkje er naudsynt med feiltoleranse på datanodar (DataNodes).
2. Forklar *combiner* i MapReduce, og kva som er hensikta med denne.
3. Kva er hensikta med bruk av data-samleband (pipeline) ved skriving av blokker i HDFS?

Skriv svaret ditt her...

Maks poeng: 10

2 Problem 2 – Spark – 10 % (all having same weight)

Tuddel er ein abonnementsbasert straumeteneste for musikk. Informasjon om alle songar som er strauma vert lagra i ein logg for å kunne utføre analyse av avspelingar, gjere anbefalingar, og rekne ut royalties til artistar. For kvar avspeling vert det generert ei linje i denne loggen, på følgjande format (komma-separert):
TimeStamp,UserName,Artist,SongName

Anta følgjande eksempel-datasett lagra i fila streamed.csv:

1,u1,a1,s1
2,u2,a1,s2
3,u1,a1,s2
4,u3,a1,s1
5,u1,a2,s3
6,u2,a1,s2
7,u2,a1,s2
8,u2,a1,s2

Dette datasettet er allereie lasta inn i ein RDD med navn s:
val s = sc.textFile("streamed.csv").map(_.split(","))

Dykk skal for kvar av deloppgåvene under vise korleis dei kan løysast vha. Spark- transformasjonar/aksjonar (Scala, Python eller Java).

1. Lag ein RDD som inneheld tal på songar strauma for kvar artist. Eksempelresultat:
   (a2,1)

(a1,7)

2. Lag ein RDD som for kvar linje har namn på brukar og tal på songar han/ho har strauma. Eksempelresultat:
   u3 1
   u2 4
   u1 3
3. Finn tal på *distinkte* songar som er spela. Eksempelresultat:
   3

Skriv svaret ditt her...

Maks poeng: 10

3 Problem 3 – NoSQL – 15 %

1. Forklar kort 4 kategoriar av NoSQL-system.
2. Forklar korleis skalering vert oppnådd i Voldemort.

Skriv svaret ditt her...

Maks poeng: 15

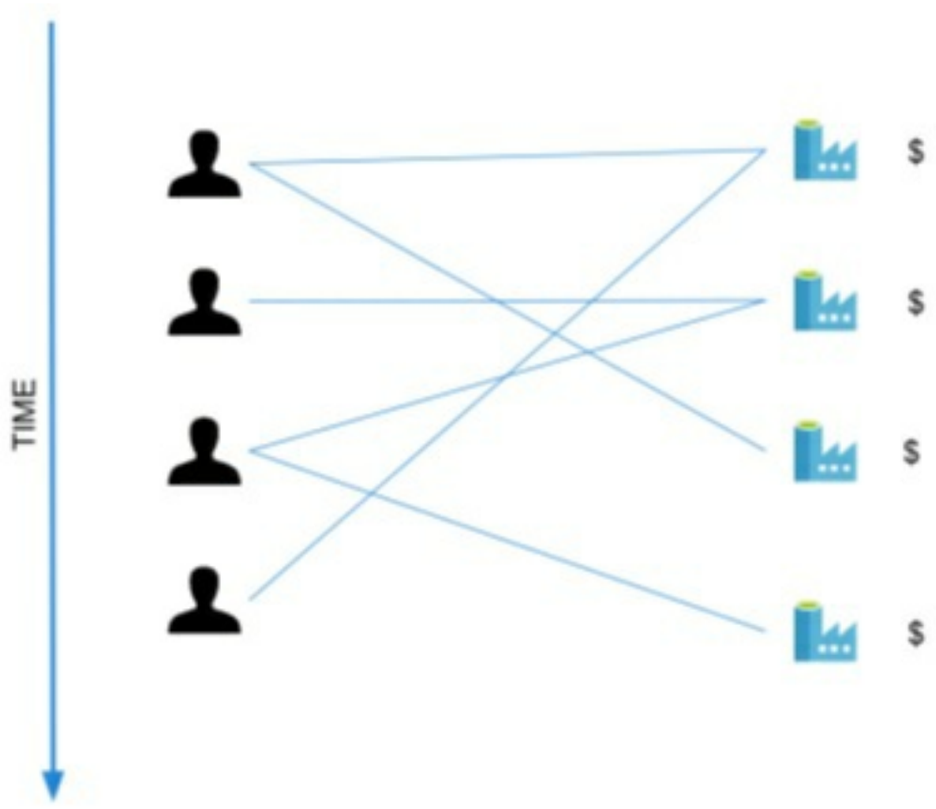
4 Problem 4 – MinHashing – 10 %

Forklar hovedtrekka bak LSH (locality sensitive hashing) for dokument, inkl.:
1. Hensikta med LSH.
2. Input og output.
3. Innhald i viktige datastruktur(ar), og algoritme.

Skriv svaret ditt her...

Maks poeng: 10

5 Oppgave 5 – Adwords – 5 %



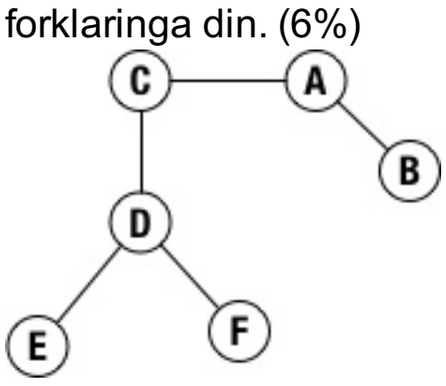
Med utgangspunkt i figuren ovenfor, forklar Adwords-problemet og ein grådig (Greedy) algoritme for å løyse dette problemet. Kva er ulempen med å bruke denne algoritmen versus ei optimal løysing? Ei lita endring i den grådige algoritmen kan gjere at ein oppnår betre resultat. Forklar denne endringa.

Skriv svaret ditt her...

Maks poeng: 5

6 Oppgave 6 – 15%

1. Forklar kort kva steg du treng å utføre for å finna «community» i ein graf. Bruk følgjande figur til å støtta forklaringa din. (6%)



2. Forklar hovudføremåla med å analysar sosiale grafar. (4%)
3. Forklar hovudskilnadene mellom **Storm** og **Spark**. Forklar kva for fordeler **AsterixDB** sitt Feed-system har i tilhøve til både **Storm** og **Spark**. (5%)

Skriv svaret ditt her...

Maks poeng: 15

7 Oppgave 7 – 20%

1. a. Kva er «Bloom Filter» og kva brukast det til? Bruk eksempel til å støtta forklaringa di. (4%)

   b. Bruke bloom filter til å fylle ut følgjande tabell. Gå ut frå at vi skal bruke h som hash-funksjon, og at den er definert som h(x)= y mod 11, der y er henta høvesvis frå oddetalls-bitar frå x eller partalls-bitar frå x.
   Eksempel: h1(39) = 011 = 3,
   h2(39) = 101 = 5, osb. (6%)

Strømelement	Hash-funksjon - h_1	Hash-funksjon - h_2	Filtrere Innhold
			000 0000 0000
85 = **101 0101**			
214 = **1111 1010**			
353 = **01 0110 0011**			

2. Anta at du skal finne andelen av spørjingar (search queries) den siste månaden frå ein straum av spørjingar som er unike. På grunn avgrensing på plass og tid kan du ikkje sjekka alle innkomne spørjingar og må difor bruke sampling. Du bestemmer deg for å lese kvar 10. spørjing (dvs. 10% sampling). Forklar kvifor dette ikkje vil gje korrekt svar? Kva ville vera ein meir korrekt måte å sampla straumen av spørjingane på? (5%)
3. Tenk at du skal bruke sosiale media som Twitter for å analysere trendar o.l. Du er spesielt interessert i å vite når eit produkt vert nemnt av folk og kor ofte dei vert nemnt. Du foreslår å bruke «bucket»-prinsippet til å telja kor mange gonger dette produktet vert nemnd. Forklar korleis «bucket»-prinsippet på datastraum fungerer i dette tilfellet. (5%)

Skriv svaret ditt her...

Maks poeng: 20

8 Oppgave 8 – 15%

1. Såkalt «*cold start*» er ei utfordring når ein brukar «collaborative filtering» i eit anbefalingssystem. Forklar kva som vert med meint med «cold start» og når dette kan oppstå. Drøft vidare kva for moglege løysingar som finst for cold start-problemet.

2. Gå ut frå følgjande brukar-rating tabell.

USERS

PRODUCTS	1	2	3	4	5	6	7	8
1	1		3			5		
2			5	4			4	
3	2	4		1	2		3	
4		2	4		5			4
5			4	3	4	2		
6	1	P	3		3			2

Anta vidare at du skal bruke «item-item collaborative filtering».

a. Forklar kva er skilnaden(e) mellom «item-item collaborative filtering» og ««user-item collaborative filtering». (4%)

b. Forklar korleis du vil gå fram for å rekne predikerte/estimerte verdien av **P**. Gjer dei føresetnadane du finn naudsynte. (5%)

Skriv svaret ditt her...

Maks poeng: 15

Framside

Institutt for datateknologi og informatikk

Eksamensoppgave i TDT4305 Big Data-arkitektur

Faglig kontakt under eksamen: Kjetil Nørvåg og Heri Ramampiaro
Tlf.: 41440433 og 99027656
Eksamensdato: 28. mai 2018
Eksamenstid (fra-til): 1500-1900
Hjelpemiddelkode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrevne hjelpemiddel tillatt. Bestemt, enkel kalkulator tillatt.
Annen informasjon:

Merk! Studenter finner sensur i Studentweb. Har du spørsmål om din sensur må du kontakte instituttet ditt. Eksamenskontoret vil ikke kunne svare på slike spørsmål.

1 Problem 1 – Hadoop – 10 % (all having same weight)

1. Forklar to teknikker som tilsammen gjør at det ikke er nødvendig med feiltoleranse på datanoder (DataNodes).
2. Forklar *combiner* i MapReduce, og hva som er hensikten med denne.
3. Hva er hensikten med bruk av data-samlebånd (pipeline) ved skriving av blokker i HDFS?

Skriv ditt svar her...

Maks poeng: 10

2 Problem 2 – Spark – 10 % (all having same weight)

Tuddel er en abonnementsbasert strømmetjeneste for musikk. Informasjon om alle sangene som er strømmet blir lagret i en logg for å kunne utføre analyse av avspillinger, gjøre anbefalinger, og regne ut royalties til artister. For hver avspilling vil det genereres en linje i denne loggen, på følgende format (komma-separert):
TimeStamp,UserName,Artist,SongName

Anta følgende eksempel-datasett lagret i filen streamed.csv:

1,u1,a1,s1
2,u2,a1,s2
3,u1,a1,s2
4,u3,a1,s1
5,u1,a2,s3
6,u2,a1,s2
7,u2,a1,s2
8,u2,a1,s2

Dette datasettet er allerede lastet inn i en RDD med navn s:
val s = sc.textFile("streamed.csv").map(_.split(","))

Dere skal for hver av deloppgavene under vise hvordan de kan løses vha. Spark- transformasjoner/aksjoner (Scala, Python eller Java).

1. Lag en RDD som inneholder antall sanger strømmet for hver artist. Eksempelresultat:
   (a2,1)
   (a1,7)
2. Lag en RDD som for hver linje har navn på bruker og tall på sanger han/hun har strømmet. Eksempelresultat:
   u3 1
   u2 4

u1 3

3. Finn antall *distinkte* sanger som er spilt. Eksempelresultat:
   3

Skriv ditt svar her...

Maks poeng: 10

3 Problem 3 – NoSQL – 15 %

1. Forklar kort 4 kategorier av NoSQL-system.
2. Forklar hvordan skalering oppnås i Voldemort.

Skriv ditt svar her...

Maks poeng: 15

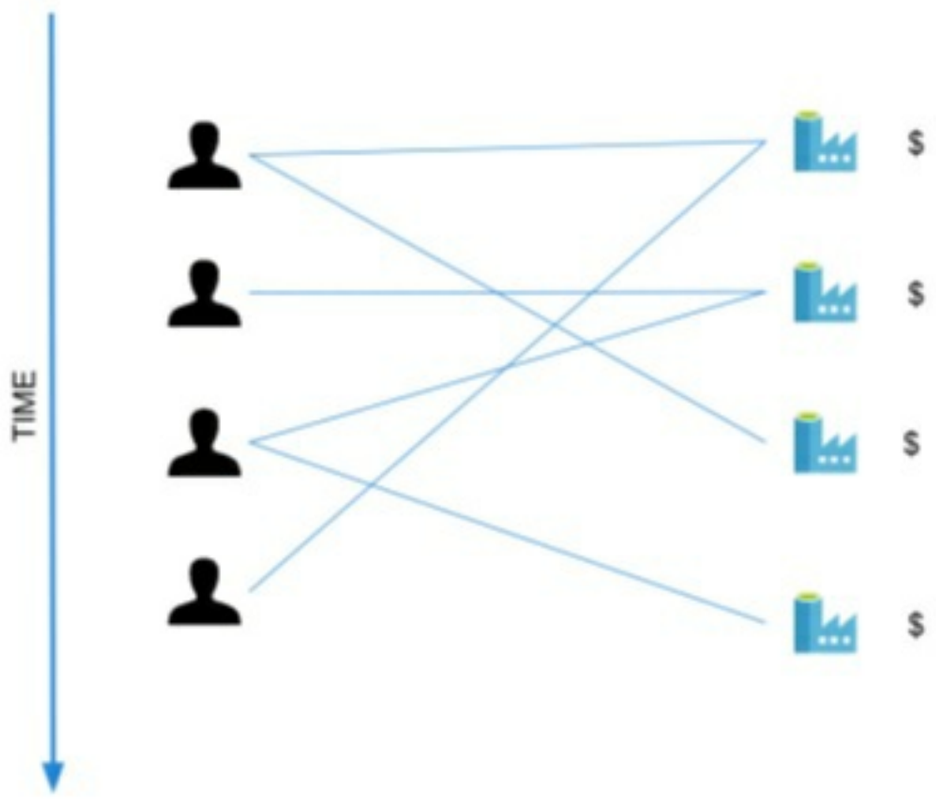
4 Problem 4 – MinHashing – 10 %

Forklar hovedtrekkene bak LSH (locality sensitive hashing) for dokumenter, inkl.:
1. Hensikten med LSH.
2. Input og output.
3. Innhold i viktige datastruktur(er), og algoritme.

Skriv ditt svar her...

Maks poeng: 10

5 Oppgave 5 – Adwords – 5 %



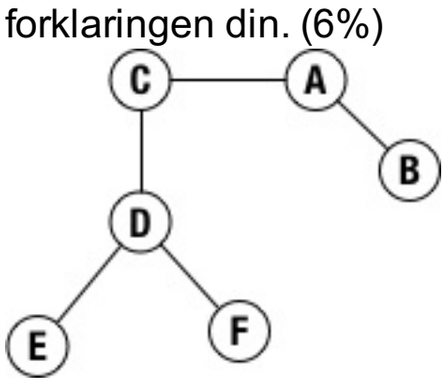
Med utgangspunkt i figuren ovenfor, forklar Adwords-problemet og en grådig (Greedy) algoritme for å løse dette problemet. Hva er ulempen med å bruke denne algoritmen versus en optimal løsning? En liten endring i den grådige algoritmen kan gjøre at man oppnår bedre resultat. Forklar denne endringen.

Skriv ditt svar her...

Maks poeng: 5

6 Oppgave 6 – 15%

1. Forklar kort hvilke steg du trenger å utføre for å finne «*community*» i en graf. Bruk følgende figur til å støtte forklaringen din. (6%)



2. Forklar hovedhensiktene med å analyser sosiale grafer. (4%)

3. Forklar hovedforskjellene mellom **Storm** og **Spark**. Forklar hvilke fordeler **AsterixDB** sitt Feed-system har i forhold til både **Storm** og **Spark**. (5%)

Skriv ditt svar her...

Maks poeng: 15

7 Oppgave 7 – 20%

1. a. Hva er «Bloom Filter» og hva brukes det til? Bruk eksempel til å støtte forklaringen din. (4%)

   b. Bruke bloom filter til å fylle ut følgende tabell. Anta at vi skal bruke h som hash-funksjon, og at den er definert som h(x)= y mod 11, der y er hentet henholdsvis fra oddetalls-bits fra xeller partalls-bits fra x. Eksempel: h1(39) = 011 = 3, h2(39) = 101 = 5, osv. (6%)

Strømelement	Hash-funksjon - h_1	Hash-funksjon - h_2	Filtrere Innhold
			000 0000 0000
85 = **101 0101**			
214 = **1111 1010**			
353 = **01 0110 0011**			

2. Anta at du skal finne andelen av spørringer (search queries) den siste måneden fra en strøm av spørringer som er unike. På grunn av plass- og tidsbegrensninger kan du ikke sjekke alle innkomne spørringer og må derfor bruke sampling. Du bestemmer deg for å lese hver 10. spørring (dvs. 10% sampling). Forklar hvorfor dette ikke vil gi korrekt svar? Hva ville være en mer korrekt måte å sample strømmen av spørringene på? (5%)
3. Tenk at du skal bruke sosiale media som Twitter for å analysere trender o.l. Du er spesielt interessert i å vite når et produkt nevnes av folk og hvor ofte de nevnes. Du foreslår å bruke «bucket»-prinsippet til å telle hvor mange ganger dette produktet blir nevnt. Forklar hvordan «bucket»-prinsippet på datastrøm fungerer i dette tilfelle. (5%)

Skriv ditt svar her...

Maks poeng: 20

8 Oppgave 8 – 15%

1. Såkalt «*cold start*» er en utfordring når man bruker «collaborative filtering» i et anbefalingssystem. Forklar hva som menes med «cold start» og når dette kan oppstå. Drøft videre hvilke mulige løsninger som finnes for cold start-problemet. (6%)
2. Anta følgende bruker-rating tabell.

PRODUCTS \ USERS	1	2	3	4	5	6	7	8
1	1		3			5		
2			5	4			4	
3	2	4		1	2		3	
4		2	4		5			4
5			4	3	4	2		
6	1	P	3		3			2

Anta videre at du skal bruke «item-item collaborative filtering».

a. Forklar hva er forskjellen(e) mellom «item-item collaborative filtering» og «user-item collaborative filtering». (4%)

b. Forklar hvordan du vil gå frem for å beregne predikert/estimert verdien av **P**. Gjør de antakelsene du finner nødvendige. (5%)

Skriv ditt svar her...

Maks poeng: 15