NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer and Information Science

Examination paper for TDT4305 Big Data Architecture

Academic contact during examination: Kjetil Nørvåg/Heri Ramampiaro

Phone: 73596755/73591459

Examination date: May 16th 2017

Examination time (from-to): 09.00-13.00

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

Other information:

Language: English

Number of pages (front page excluded): 4

Number of pages enclosed: 0

Checked by:

Date Signature

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig X 2-sidig □

sort/hvit X farger □

Problem 1 – Hadoop – 15 % (all having same weight)

a) Explain replication in HDFS.

b) What is the purpose of the NameNode, and what information does it store?

c) What is input and output of, respectively, the Map and Reduce functions?

Problem 2 – Spark – 20 % (all having same weight)

Assume a file *countries.tsv* (tsv = tab separated) which contains information about countries and characteristic terms for these countries (in this case, you can assume one tab "\t" per line, between country and the term list):

```
France     wine,art,trains,wine,trains
Canada     glaciers,lakes,hockey,maple,grizzly
Norway     fjords,fjords,glaciers,trolls
Japan      trains,sushi,origami,sushi,fuji
Argentina  wine,glaciers,football,glaciers
```

You are now, for each of the problems below, to show how they can be solved using Spark transformations /actions (Scala, Python, or Java).
Hint: Split a text string x based on "\t" (tab): val y = x.split("\t")

a) Create an RDD with name ”data” based on the file, where each record/object is a text string (String) containing one line from the file.

b) Create a new RDD ”data2” where each object in ”data” is an ”array/list of strings”, where the first text string is country and text string number two contains the terms. Example result (Python):

```
[['France', 'wine,art,trains,wine,trains'],
 ['Canada', 'glaciers,lakes,hockey,maple,grizzly'],
 ['Norway', 'fjords,fjords,glaciers,trolls'],
 ['Japan', 'trains,sushi,origami,sushi,fuji'],
 ['Argentina', 'wine,glaciers,football,glaciers']]
```

c) Create a new pair RDD ”data3” where (key is country, and value is an “array/list of strings” of characteristic terms, example result:

```
[('France', ['wine', 'art', 'trains', 'wine', 'trains']),
 ('Canada', ['glaciers', 'lakes', 'hockey', 'maple', 'grizzly']),
 ('Norway', ['fjords', 'fjords', 'glaciers', 'trolls']),
 ('Japan', ['trains', 'sushi', 'origami', 'sushi', 'fuji']),
 ('Argentina', ['wine', 'glaciers', 'football', 'glaciers'])]
```

d) Find number of characteristic terms in the dataset (not including name of the countries).

e) Find the number of *distinct* characteristic terms in the dataset.

f) Based on ”data3”, create an RDD that contains number of distinct terms for each country. Example result:

```
[('Argentina', 3), ('Norway', 3), ('France', 3), ('Canada', 5), ('Ja
pan', 4)]
```

Problem 3 – NoSQL – 5 %

Explain the CAP-theorem.

Problem 4 – MinHashing – 10 % (all having same weight)

lx	Element	S1	S2	S3	S4	lx2
0	p	0	0	0	1	3
1	a	0	1	1	1	2
2	g	1	1	0	0	1
3	i	1	1	1	1	0
4	k	0	1	0	0	4

a) Explain ”shingling” and the purpose of ”shingling”.

b) The figure above shows the occurrence matrix for the elements p/a/g/i/k in the sets S1/S2/S3/S4. What are the MinHash signatures given the permutations lx and lx2?

Problem 5 – Streaming data – 30 % (all having same weight)

Suppose you work for a company analysing interest trends related to the new Sony A9 camera that Amazon just started to sel. Your company has decided to do this based on Social Media data, including Twitter and Facebook.

a) Assume that to begin with, you decide to find the number of messages mentioning “Sony A9” by using so-called Standing Query. Explain the difference(s) between ”standing query” and ”ad-hoc query”. How can standing query solve your task?

b) Since we can consider social media data as streaming data, they have also other characteristics and/or challenges than static data. Explain what these characteristics and/or challenges are.

c) As an alternative approach to ”standing query”, you choose ”bit counting” to solve parts of the message counting task. Explain how this task can be translated to bit counting.

d) Your analysis will see the trends the last 2 weeks and you consider number of “clicks” and “purchase” of products at Amazon directly as part of your analysis. Imagine that you want to calculate the fraction of the number of “click” and “purchase” on “Sony A9”. For this purpose we choose to use the sliding window principle. Assume that our sliding window has a size of 500 clicks and purchase in total (i.e., combined). Show how you can compute this fraction.

e) Speaking about clicks, explain how bloom filters could be useful when trying to find the number of clicks. Make any assumptions you find necessary.

f) Use the bloom filter principle to fill out the table below:

Strømelement	**Hash-funksjon - h_1**	**Hash-funksjon - h_2**	**Filtrere Innhold**
			00000000000
56 = 11 1000			
428 = 1 1010 1100			
875 = 11 0110 1011			

Hint: Use *h1(x)=y1 mod 11* and *h2(x)=y2 mod 11* as hash functions*,* where the values of *y1* and *y2* are generated from odd-numbered bits in *x* and even-numbered bits in *x,* respectively.

Problems 6 – Recommender systems and social graphs – 20 % (all having same weight)

The company you just started to work for is expert on recommender systems. Your own task is to develop good recommendation algorithms and methods.

a) Focusing on pros and cons, compare ”collaborative filtering” against ”content-based recommendation”.

b) As part of the analyses of user preferences, you use social media. Specifically you are interested in finding relations between users in terms of “friends” relations on Facebook and “followers” relations on Twitter. Graphs can be used to show these relations. Use this as the starting point to answer the following questions:

i. Explain the differences between ”*overlapping*” and ”*non-overlapping communities”* in graphs. How can we in a simple way find out that two specific graphs overlap?

ii. What are the main purposes with community detection? Explain.

c) Girvan-Newman method to compute ”betweenness” is a common methods within graph mining theory. Show how you go from Figure 1 to Figure 2 below using the Girvan-Newman method to compute ”*betweenness”* in the graph depicted in Figure 1.

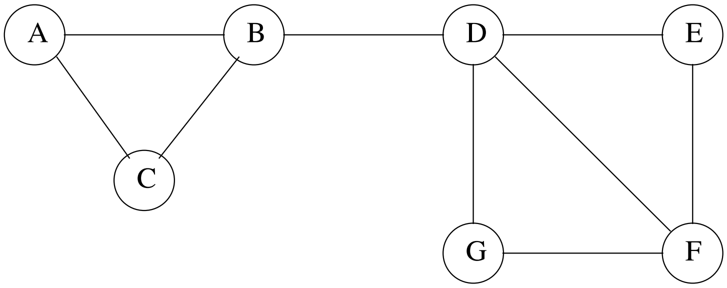


Figure 1: Start graph

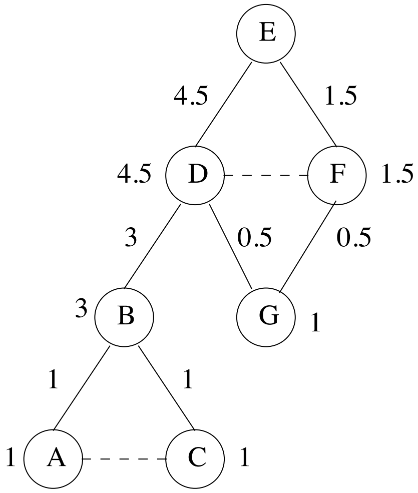
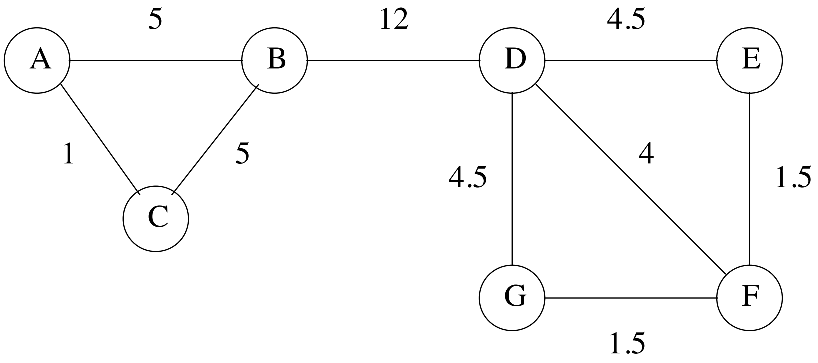


Figure 2: Step 3 in first iteration.

d) How can the (final) result figure below used to detect “communities”?


5
12
4.5
A
B
D
E
1
5
4
4.5
1.5
C
G
F
1.5