TDT4305 2021 - Assignment 1 solutions

Intro

1) For each of the three Vs, answer the following:
   a) Why is this property showing up in the Big Data era and not previously?
      - *See [ElmasriNavathe] section 25.1.*
   b) What challenges does this property give for traditional RDBMSs?
      - *See [ElmasriNavathe] section 25.1.*
2) Given Big Data from a field of your choice, how could you use it to create value for yourself or others? E.g. Big Data in medicine can be used for diagnosing diseases based on patients' health history.
   - *See [ElmasriNavathe] section 25.1 for inspiration.*
3) What are the challenges in ensuring the trustworthiness of Big Data?
   - *See [ElmasriNavathe] section 25.1.*

MapReduce and HDFS

1) Solve the following exercises from the main course book MMDS:

   2.2.1

   a) We expect there to be significant skew, since there is often a big variation in the lengths of the value lists for different keys, so different reducers take different amounts of time.
   b) We expect the impact of skew to be less if we combine the reducers into 10 Reduce tasks. If keys are sent randomly to Reduce tasks, we can expect there will be some averaging of the total time required by the different Reduce tasks. On the other hand, we may not benefit from using 10,000 Reduce tasks, since there is an overhead associated with each task we create, and in general, the number of Reduce tasks should be lower than the number of keys.

   2.3.1

   a) The Map function simply outputs each number. The key can be ignored. A single Reducer scans all received integers and outputs the largest. Since the *max* function is associative and commutative, a combiner can be used after the Map task.
   b) As above, the Map function simply outputs each number and keys can be ignored. A single Reducer sums and counts all numbers and computes the average. Since the *average* function is not associative and commutative, a Combiner cannot be used.

c) We use the same strategy as word count, but ignore the values, i.e. the number of times a word occurred. The Map function again simply outputs each number. A single Reducer is here an identity function. Since this corresponds to the *unique* function, which is associative and commutative, a combiner can be used after the Map task.

2.3.2

- The algorithm does not need generalization to work on non-square matrices.

2) What is the role of the DFS (GFS or HDFS) in a MapReduce system?
   - *See [DeanGhemawat] section 3.4.*
3) What is the difference between a NameNode and a DataNode in HDFS?
   - *See [HDFS] section 2.*
4) What is a data block in HDFS and how is it replicated across data nodes?
   - *See [HDFS] section 3.*

TDT4305 2021 - Assignment 2 solutions

Spark

1) What are some of the differences between Spark and MapReduce?
    - *See [Zaharia] section 1.*
2) What are *narrow* and *wide dependencies*, and how do they relate to a *stage*?
    - *See [Zaharia] section 5.1.*
3) What are the three ways of storing persistent RDDs and how are they each useful?
    - *See [Zaharia] section 5.3.*
4) What is checkpointing and when is it typically used?
    - *See [Zaharia] section 5.4.*
5) What is the purpose of lazy evaluation?
    - *See [LearningSpark] page 29.*

TDT4305 2021 - Assignment 3

1. k-Shingling and Jaccard

Given four sequences of letters, baaa, accc, abba, and baac, find the *k*-shingles for each sequence, where *k=2*.

a) Fill in the table with the remaining *k*-shingles ordered alphabetically.

b) Fill in which *k*-shingle appears in which sets. Use 0s and 1s.

Characteristics matrix:

id	***k*-shingle**	**baaa**	**accc**	**abba**	**baac**
1	aa	1	0	0	1
2	ab	0	0	1	0
3					
4					
5					
6					

c) Fill in the Jaccard similarity table with the remaining similarities between the sequences using their *k*-shingle sets.

Jaccard similarity:

	baaa	**accc**	**abba**	**baac**
baaa	1	0	1/4	
accc				
abba				
baac				

2. Min-Hashing

Generate signatures using the permutation method presented in MMDS ch. 3 video 2. We want signatures of length *n*=3, so three permutations are needed. They are given below.

Row permutations:

1st	**2nd**	**3rd**	**4th**	**5th**	**6th**
6	4	3	2	1	5
4	1	5	2	3	6
2	3	6	4	1	5

a) Fill in the table with the correct signature values.

Signatures:

s1	***s2***	***s3***	***s4***
2	1	2	2

b) Why is it not always feasible to use permutations and how can the permutations be emulated?

3. LSH

Given signatures for the four sequences, find candidate pairs by hashing the signatures into buckets. To make it more interesting, we now assume n=9. We will use b=3 bands with r=3 rows each. We will assume an identity hash function, i.e. for two signatures ($x1, x2, x3$) and ($y1, y2, y3$) appearing in the same band, the signatures go to the same bucket if and only if $x1=y1$ and $x2=y2$ and $x3=y3$.

a) Fill in the table below with the remaining part of the signatures found in exercise 2a.

Signatures split in bands:

	s1	***s2***	***s3***	***s4***	
b1	2	1	2	2	
b2	5 3 1	4 3 1	4 3 1	5 1 1	
b3	4 1 2	2 2 3	4 5 5	4 1 2	

b) What candidate pairs can we extract from the buckets?

TDT4305 2021 - Assignment 4 solution

Adwords problem

1. Given the following table of advertisers and their bids on queries, compute the advertiser-query pairs using the three algorithms. For all algorithms, tie-break on the index, smallest index first.

Advertiser	Query	Bid
a_1	q_1	0.5
a_1	q_3	1
a_2	q_2	0.5
a_3	q_2	0.5
a_3	q_4	1
a_4	q_1	0.75

a) Assume for the Greedy algorithm that all bids are 1 or 0 and the budget of each advertiser a_i is B_i=2. Fill in the table for the Greedy algorithm.

Time	Query	Candidates	Budget left	Accu. revenue	Notes
1	q_1	**a_1**, a_4	B_1=1	1	Tie-break
2	q_2	**a_2**, a_3	B_2=1	2	Tie-break
3	q_3	**a_1**	B_1=0	3	
4	q_4	**a_3**	B_3=1	4	
5	q_3	a_1	x	4	
6	q_3	a_1	x	4	
7	q_2	**a_2**, a_3	B_2=0	5	Tie-break
8	q_4	**a_3**	B_3=0	6	

Assume the following budgets B_i for advertisers a_i in the next two algorithms:

Advertiser	Budget
a_1	3
a_2	1
a_3	1
a_4	2

b) Fill in the table for the Balance algorithm.

Time	Query	Candidates & bids	Budget left	Accu. revenue	Notes
1	q_1	**(a_1, 0.5)**, (a_4, 0.75)	B_1=2.5	0.5	Largest remaining budget
2	q_2	**(a_2, 0.5)**, (a_3, 0.5)	B_2=0.5	1	Tie-break
3	q_3	**(a_1, 1)**	B_1=1.5	2	
4	q_4	**(a_3, 1)**	B_3=0	3	
5	q_3	**(a_1, 1)**	B_1=0.5	4	
6	q_3	(a_1, 1)	x	4	No budget left
7	q_2	**(a_2, 0.5)**, (a_3, 0.5)	B_2=0	4.5	Only remaining budget
8	q_4	(a_3, 1)	x	4.5	No budget left

c) Fill in the table for the Generalized Balance algorithm.

Time	Query	Candidates & bids	Scores	Budget left	Accu. revenue	Notes
1	q_1	(a_1, 0.5), **(a_4, 0.75)**	$0.5(1-e^{-1})\approx0.31$ $0.75(1-e^{-1})\approx\underline{0.47}$	B_4=1.25	0.75	Highest score
2	q_2	**(a_2, 0.5)**, (a_3, 0.5)	$0.5(1-e^{-1})\approx\underline{0.31}$ $0.5(1-e^{-1})\approx0.31$	B_2=0.5	1.25	Tie-break
3	q_3	**(a_1, 1)**		B_1=2	2.25	
4	q_4	**(a_3, 1)**		B_3=0	3.25	
5	q_3	**(a_1, 1)**		B_1=1	4.25	
6	q_3	**(a_1, 1)**		B_1=0	5.25	
7	q_2	**(a_2, 0.5)**, (a_3, 0.5)		B_2=0	5.75	Only remaining budget
8	q_4	(a_3, 1)		x	5.75	

2. What is the definition of the competitive ratio?
    - *See [MMDS] section 8.2.3.*
3. What is broad matching and why is it useful?
    - *See [MMDS] section 8.4.3.*
4. What is second-price auction and why is it useful?
    - *See [MMDS] section 8.4.3.*

TDT4305 2021 - Assignment 5 solution

1. For a stream of integers, we maintain a Bloom filter answering which integers have been seen so far. Assume the following hash functions: $h_1(x)=2x \bmod 10$, $h_2(x)=3x \bmod 10$.

   a) For the incoming stream values 6, 18, and 3, fill in the table below and update the Bloom filter.

Time	Element	h_1	h_2	Filter
1	6	2	8	0010000010
2	18	6	4	0010101010
3	3	6	9	0010101011

2. For the following bit stream, we want to estimate the number of 1-bits in the last k bits. The left-most bits are the oldest.

   a) Divide the following bit stream into a valid set of buckets using the DGIM algorithm.

   … 1 1 0 1 0 0 1 0 0 1 0 1 1 0 1 0 0 0 0 1 0 1 0 0 1 0 1 0

   b) Update the buckets after the arrival of one more bit.

   … 1 1 0 1 0 0 1 0 0 1 0 1 1 0 1 0 0 0 0 1 0 1 0 0 1 0 1 0 ***1***

   c) Calculate the estimated number of 1 bits in the latest 20 bits using the DGIM algorithm with guaranteed maximum 50% error. What is the error in this case?

   - Estimated number of 1 bits: 7
   - 0.222% error

TDT4305 2021 - Assignment 6 solution

Storm

1. Mention four of the main components in Storm and describe their responsibilities and communication with other components.
    - *Spout and bolt ([Storm] Sect. 2)*
    - *Nimbus, Zookeeper, and Supervisor ([Storm] Sect. 2.2.1)*
    - *Workers and executors ([Storm] Sect. 2.2.3)*
    - ...
2. How is parallelism achieved with tasks and executors?
    - *[Storm] Sect. 2.1*
3. What are the two processing semantics and how are they achieved?
    - *[Storm] Sect. 2.3*

AsterixDB

1. Mention four of the main components in AsterixDB and describe their responsibilities and communication with other components.
    - *Node controller and metadata node controller ([AsterixDB1] Sect. 1)*
    - *Algebricks ([AsterixDB1] Sect. 4.2)*
    - *Feed joint ([AsterixDB1] Sect. 4.5, [AsterixDB2] Sect. 5.1)*
    - *Nodegroup ([AsterixDB2] Sect. 3.2)*
    - *Feed adaptor ([AsterixDB2] Sect. 4.1)*
    - *Manager and worker nodes ([AsterixDB2] Sect. 5.2)*
    - ...
2. How is parallelism achieved with Operators?
    - *[AsterixDB1] Sect. 4.1*
3. What are the different ways AsterixDB can deal with a congested pipeline?
    - *[AsterixDB2] Table 2 and Sect. 5.3*

TDT4305 2021 - Assignment 7 solution

Content-based recommendation

1. What are the steps for recommending items to users using content-based recommendation?
    - *See sections 9.2.4 to 9.2.6.*

Collaborative filtering recommendation

1. Given the following utility matrix, predict user 3's rating of movie 1.

		Users				
		1	**2**	**3**	**4**	**5**
Movies	**1**	1	2	?	2	3
	2		3	3		
	3		1	3	5	
	4	5	3	2		2

a) Calculate the Pearson correlation between movie 1 and the other movies.

1. m_1=(1+2+2+3)/4=2

   r_1=[1-2, 2-2, 0, 2-2, 3-2]=[-1, 0, 0, 0, 1]

   cos(r_1,r_1)=$r_1 \cdot r_1$/||r_1||·||r_1||=2/2=1

2. m_2=(3+3)/2=3

   r_2=[0, 3-3, 3-3, 0, 0]=[0, 0, 0, 0,0]

   cos(r_1,r_2)=$r_1 \cdot r_2$/||r_1||·||r_2||=0/0=undefined

3. m_3=(1+3+5)/3=3

   r_3=[0, 1-3, 3-3, 5-3, 0]=[0, -2, 0, 2, 0]

   cos(r_1,r_3)=$r_1 \cdot r_3$/||r_1||·||r_3||=0/4=0

4. m_4=(5+3+2+2)/4=3

   r_4=[5-3, 3-3, 2-3, 0, 2-3]=[2, 0, -1, 0, -1]

   cos(r_1,r_4)=$r_1 \cdot r_4$/||r_1||·||r_4||=-3/3.464=−0.866

b) Let N be the set of 2 movies most similar to movie 1 that have been rated by user 3. Calculate user 3's predicted rating of movie 1.

- (0*3-0.866*2)/(0-0.866)=2

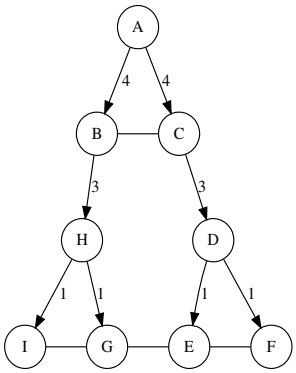
TDT4305 2021 - Assignment 8 solution

Mining Social Network Graphs

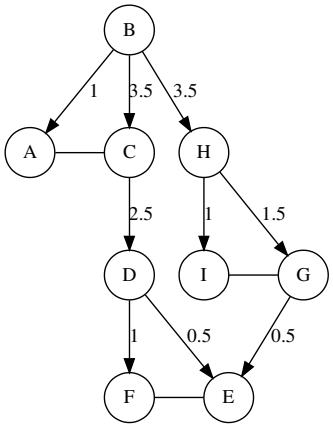
1. Why do the hierarchical and point-assignment clustering methods from [MMDS] chapter 7 not work for clustering social graphs?
   - *See section 10.2.2.*
2. What is the definition of edge betweenness?
   - *See section 10.2.3.*
3. [MMDS] exercise 10.2.1. (a) and (b).

   Figure 10.9 is an example of a social-network graph. Use the Girvan-Newman approach to find the number of shortest paths from each of the following nodes that pass through each of the edges.

   a)



   b)



4. Given the answer to question 3 above, how would you proceed to find the final betweenness scores and a set of communities?
    - “To complete the betweenness calculation, we have to repeat this calculation for every node as the root and sum the contributions. Finally, we must divide by 2 to get the true betweenness, since every shortest path will be discovered twice, once for each of its endpoints” [MMDS p. 366]. To find a set of communities, “[s]tart with the graph and all its edges; then remove edges with the highest betweenness, until the graph has broken into a suitable number of connected components” [MMDS p. 367].