Course Home Page

Quick Questions 15

COURSE

Announcements

Video Lectures

Video Errata

EXERCISES

Quizzes (Homeworks)

Surveys

Final (Advanced)

Help

The **due date** for this exam is **Mon 1 Dec 2014 12:01 AM PST**.

This is the Advanced Final Exam for the MMDS Course. Your score on this exam counts only for a SoA "With Distinction.". You must submit your work within 2 hours of opening. The exam is open-book; any inanimate source may be used. There is no penalty for a wrong answer (compared with no answer), so feel free to guess.

☐ **In accordance with the Coursera Honor Code, I (Tommi Hovi) certify that the answers here are my own work.**

Question 1

Suppose ABCD is **not** a frequent itemset, while ABC and ACD **are** frequent itemsets. Which of the following is **definitely** true?

○ BC is a frequent itemsel.

- ○ ABD is a frequent itemset.
- ○ ABCE is not a frequent itemset.
- ○ ABCD is in the negative border.

Question 2

Suppose we are representing sets by strings and indexing the strings according to both the symbol and its position within the prefix. We want to find strings within Jaccard distance at most 0.2 (i.e., similarity at least 0.8), and we are given a probe string of length 24. Into how many buckets must we look?

- ○ 15
- ○ 18
- ○ 5
- ○ 21

Question 3

In the following question we consider an example of the implementation of the PCY algorithm. All numbers should be treated as decimal; e.g., "one million" is 1,000,000, NOT 2^{20} = 1,048,576. All integers (item counts and bucket counts) require 4 bytes.

We have one billion bytes of main memory available for the first pass. There are 100,000,000

items, and also 100,000,000 baskets, each of which contains exactly 10 items. Say that PCY is **effective** if the average count of a bucket is at most half the support value. For the given data, what is the smallest support value for which PCY will be effective?

- ○ 6
- ○ 60
- ○ 600
- ○ 6000

Question 4

Suppose we want to represent the multiplication of two 10-by-10 matrices as a "problem" in the sense used for our discussion of the theory of MapReduce algorithms. How many pairs are in the input-output mapping?

- ○ 2000
- ○ 100
- ○ 20
- ○ 1000

Question 5

The "all-triples" problem is described by n inputs, (n choose 3) outputs, and an input-output mapping where each output is connected to a different set of three inputs. Suppose q is the reducer size. Which of the following functions of n and q approximates, to within a constant factor, the lowest possible replication rate for a mapping schema that solves this problem?

- ○ $r = n^2/q^2$
- ○ $r = n^3/q$
- ○ $r = n^2/q$
- ○ $r = n/q^2$

Question 6

Suppose we are running the DGIM algorithm (approximate counting of 1's in a window. At time t, the list of bucket sizes being maintained is 8,4,4,2,1,1. At times t+1, t+2, and t+3, 1's arrive on the input. Assuming no buckets are deleted because they fall outside the window, what are the numbers of buckets after each of the times t+1, t+2, and t+3?

- ○ time t+1: 5; time t+2: 6; time t+3: 5
- ○ time t+1: 7; time t+2: 6; time t+3: 7.
- ○ time t+1: 7; time t+2: 8; time t+3: 9
- ○ time t+1: 6; time t+2: 7; time t+3: 5.

Question 7

Apply the HITS algorithm to a network with four pages (nodes) A, B, C, and D, arranged in a chain:

A-->B-->C-->D

Compute the hubbiness and authority of each of these pages (scale doesn't matter, because you only have to identify pages with the same hubbiness or the same authority). Which of the following is FALSE.

- ○ A and B have the same authority.
- ○ A and B have the same hubbiness.
- ○ B and C have the same hubbiness.
- ○ B and C have the same authority.

Question 8

Let G be the complete graph on five nodes (i.e., there is an edge in G between every pair of distinct nodes). What is the sum of the **squares** of the elements of the Laplacian matrix for G?

- ○ 0
- ○ 100
- ○ 20
- ○ 40

Question 9

Note: This problem is similar to one on the Basic Final, but involves a combiner.

Consider the following MapReduce algorithm. The input is a collection of positive integers. Given integer X, the Map function produces a tuple with key Y and value X for each prime divisor Y of X. For example, if X = 20, there are two key-value pairs: (2,20) and (5,20). The Reduce function, given a key K and list L of values, produces a tuple with key K and value sum(L) i.e., the sum of the values in the list.

Suppose we process the input 9, 15, 16, 23, 25, 27, 28, 56, using a Combiner. There are 4 Map tasks and 1 Reduce task. The first Map task processes the first two inputs, the second the next two, and so on. How many input tuples does the Reduce task receive?

- ○ 8
- ○ 11
- ○ 3
- ○ 6

Question 10

Consider an AdWords scenario with 4 advertisers competing for the same query Q, all with the

same budget of $100 and the same clickthrough rate. The table below shows the bid and the dollars spent by each advertiser until this point. Suppose we use Generalized BALANCE, and show one ad for each query. Which advertiser do we pick the next time query Q comes up?

Advertiser	Bid	Spend
A	$1	$20
B	$2	$40
C	$3	$60
D	$4	$80

- ○ C
- ○ D
- ○ B
- ○ A

Question 11

Suppose we wish to estimate the rating of movie M by user U using item-Item Collaborative Filtering, but there are no movies really similar to movie M. The average of all ratings is 3.5, user U's average rating is 3.1, and movie M's average rating is 4.3. What is our best guess for the rating of movie M by user U using a global baseline estimate?

- ○ 3.9
- ○ 3.5

- ○ 4.3
- ○ 4.7

Question 12

The table below shows data from ten people showing whether they like four different ice cream flavors.

Chocolate	Vanilla	Strawberry	Peanut
Y	N	Y	Y
N	Y	Y	N
N	N	N	N
Y	Y	Y	Y
Y	Y	N	Y
N	N	N	N
Y	Y	Y	Y
N	Y	N	N
Y	N	Y	N
Y	N	Y	Y

Fit a decision tree that predicts whether somebody would like Peanut ice cream based on whether she liked the other three flavors. Use Information gain as the measure to make the splits. What is

the order of splits?

- ○ Only Chocolate
- ○ Strawberry->Chocolate->Vanilla
- ○ Vanilla->Chocolate
- ○ Chocolate->Strawberry->Vanilla
- ○ Chocolate->Vanilla

Question 13

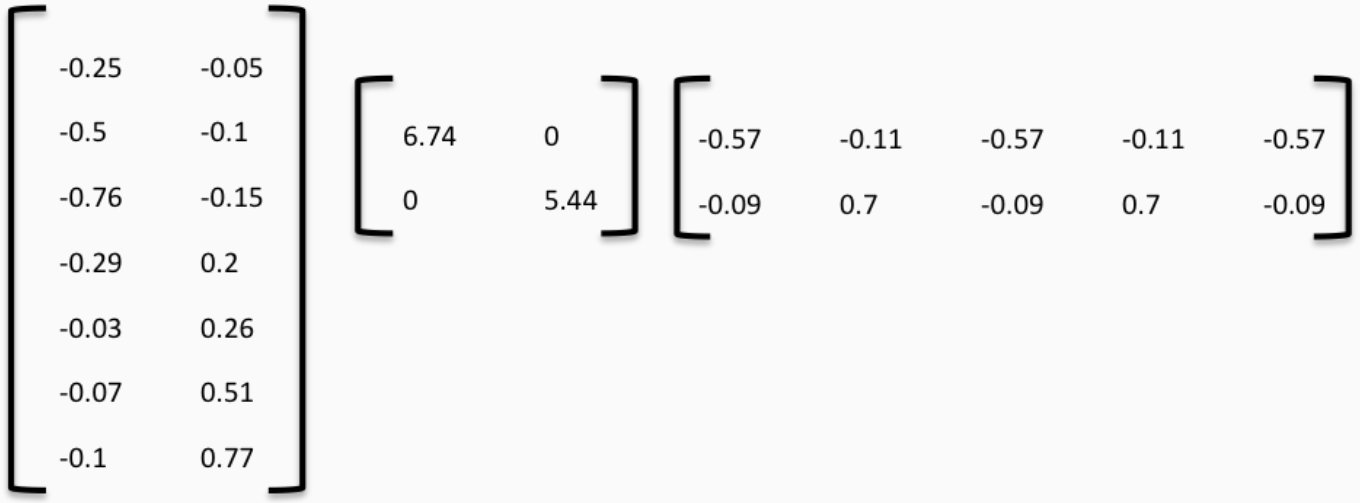
For an unknown graph with 3 nodes, r_1 is the topic sensitive PageRank using teleport set {0, 1} and r_2 is is the topic sensitive PageRank using teleport set {1}. What's the value of the topic sensitive PageRank vector when using teleport set {0}?

- ○ $r_2 - r_1$
- ○ $2r_1 - r_2$
- ○ $r_1 - 2r_2$
- ○ $r_1 - r_2$

Question 14

A is a users times movie-ratings matrix like the one seen in class. Each column in A represents a movie, and there are 5 movies in total. Recall that a Singular Value Decomposition of a matrix is a

multiplication of three matrices: U, Σ and V. is the following is such a decomposition for matrix A:



If we get three new users with the following rating vectors: User 1: [5,0,0,0,0] User 2: [0,5,0,0,0] User 3: [0,0,0,0,4] If for advertising purposes we want to cluster these three customers into two clusters using the movie concepts as features. How would you cluster them? (use cosine distance).

- ○ [1] and [2] and [3]
- ○ [1,2,3]
- ○ [1,3] and [2]
- ○ [1] and [2,3]
- ○ [1,2] and [3]

Question 15

The soft margin SVM optimization problem is:

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$s.t.\ y_i \cdot (x_i \cdot w + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n$$
$$\xi_i \geq 0, \quad \forall i = 1, \ldots, n$$

If for some i we have ξ=0, this indicates that the point x_i is (check the true option):

- ○ Exactly in the decision boundary
- ○ Correctly classified
- ○ Incorrectly classified
- ○ A support vector

☐ **In accordance with the Coursera Honor Code, I (Tommi Hovi) certify that the answers here are my own work.**

Submit Answers | Save Answers

You cannot submit your work until you agree to the Honor Code. Thanks!

Time remaining
1:59:45