Fall Semester 2018

KAIST EE412

Foundation of Big Data Analytics

Final Exam

Name: _____

Student ID: _____

I agree to comply with the School of Electrical Engineering Honor Code.

Signature: _____

This exam is open book and notes. You may also use portable electronic devices, but only for viewing PDF files and not for Internet connection or calculation. Read the questions carefully and focus your answers on what has been asked. You are allowed to ask the instructor/TAs for help only in understanding the questions, in case you find them not completely clear. Be concise and precise in your answers and state clearly any assumption you may have made. You have 165 minutes (9:00 AM – 11:45 AM) to complete your exam. Be wise in managing your time. Good luck.

Question	Score
1	/10
2	/10
3	/15
4	/10
5	/10
6	/15
7	/15
8	/15
9	/15
10	/10
11	/10
Total	/135

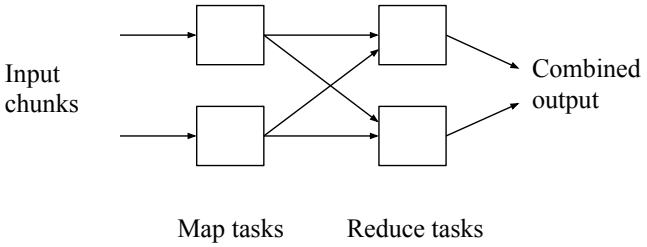
1 (10 points) Map Reduce

The $DISTINCT(X)$ operator is used to return only distinct (unique) values for an attribute X in a relation. For example, given the relation below, $DISTINCT(URL)$ returns (url 1, url 2, url 3).

URL
url 1
url 1
url 2
url 3

(a) (4 points) Implement the $DISTINCT(X)$ operator using the Map Reduce framework. Provide the algorithm pseudocode for the Map and Reduce functions. One pass over the data should be sufficient.

(b) (3 points) Suppose the input relation has one attribute called URL and consists of 100M random rows where each row can be one of 10 urls (url 1, ..., url 10) with equal probability. Also, suppose there are 2 Map tasks and 2 Reduce tasks running your code in (a) as shown below and that each tuple emitted by a Map task is 1B in size. What is the expected communication cost (i.e., total amount of data) sent from the Map tasks to the Reduce tasks for running the operation $DISTINCT(URL)$ on this relation?



(c) (1 point) What is a MapReduce functionality that can be used to further save the above communication cost?

(d) (2 points) If we use the functionality in (c), what is the resulting expected communication cost?

2 (10 points) Spark

(a) (5 points) Implement a PySpark program that counts the number of distinct strings in a text file. Assume the text file contains one URL per line. Do not use the built-in `distinct().count()`. Please write real code, but we will also give partial credit for pseudo code. You can get full credit even if your code does not compile. Fill in the blank lines in the code below.

```
> count = sc.textFile(file) \
```

(b) (1 point) **True or False**: Spark is better than MapReduce for performing interactive analysis.

(c) (1 point) **True or False**: Spark only runs on memory and does not use disk.

(d) (1 point) **True or False**: Spark has a more complicated recovery mechanism than MapReduce, but compensates by running faster.

(e) (1 point) **True or False**: Spark is suitable for implementing Web applications that require frequent updates.

(f) (1 point) **True or False**: Using a FlatMap function is necessary for counting words in documents.

3 (15 points) Frequent Itemsets

Suppose we have a collection of six baskets (the first two are duplicates). Each contains three of six items 1 through 6.

$$\begin{array}{lll} \{1, 2, 3\} & \{1, 2, 3\} & \{1, 3, 5\} \\ \{1, 4, 5\} & \{2, 3, 5\} & \{2, 3, 6\} \end{array}$$

Suppose the support threshold is 3, and we use the following three hash functions for mapping an item pair $\{i, j\}$ to a bucket:

$h_1(i, j) = i \times j \bmod 10$
$h_2(i, j) = i + j \bmod 5$
$h_3(i, j) = i + 2j \bmod 5$

When evaluating h_3, order the items so that $i < j$ to ensure that h_3 is symmetric.

(a) (6 points) If we run the PCY algorithm using h_1 and 10 buckets, what is the set of candidate pairs C_2?

(b) (6 points) Suppose we run the Multistage algorithm using h_2 and then h_3 with 5 buckets each.

How many pairs of items are hashed to the buckets of the second hash table?

What is the set of candidate pairs C_2?

(c) (3 points) Find all the association rules with confidence at least 0.8 and support at least 3.

4 (10 points) Finding similar items

(a) (4 points) Suppose we generate minhash signatures for documents where we divide the signature matrix into 4 bands and 3 rows to use locality-sensitive hashing. If the probability that the two documents are candidates (i.e., they hash to the same bucket at least once) is 0.5, what is their Jaccard similarity? A correct expression is sufficient, and you do not have to give the actual number.

(b) (6 points) Compute sketches using the random hyperplane method.
Consider the following vectors in a 4-dimensional space:

$a = [2, 3, 4, 5]$
$b = [-2, 3, -4, 5]$
$c = [-2, -3, -4, -5]$

Suppose we use the following random hyperplanes:

$v_1 = [+1, +1, +1, -1]$
$v_2 = [+1, +1, -1, +1]$
$v_3 = [+1, -1, +1, +1]$

Compute the sketches of the three vectors.

a:_____

b:_____

c:_____

Estimate the angles between the vectors using their sketches.

a, b:_____

b, c:_____

a, c:_____

5 (10 points) Clustering

(a) (6 points) Perform hierarchical clustering on a one-dimensional set of integers 1, 2, ..., 8. In case there are ties on which clusters to merge first, we merge the two clusters that contain the lowest integer. That is, if the distance between clusters A and B is the same as the distance between clusters C and D, we merge A and B if the minimum integer in A and B is smaller than the minimum integer in C and D. Show the dendogram (i.e., tree) of the complete grouping of the points when the distance between two clusters is defined as follows:

The minimum of the distances between any two points, one from each cluster:

The maximum of the distances between any two points, one from each cluster:

The average of the distances between any two points, one from each cluster:

(b) (1 point) **True or False**: The k-means algorithm is not affected by the curse of dimensionality.

(c) (1 point) **True or False**: For the BFR algorithm, the space required for storing clusters increases linearly to the number of dimensions.

(d) (1 point) **True or False**: The CURE algorithm can run on data that does not fit in memory.

(e) (1 point) **True or False**: When running the GRGPF algorithm, assuming the cluster tree is initialized, the points on disk are read with a single scan.

6 (15 points) Recommendation Systems

In class, we learned three techniques for making recommendations: content-based recommendation, user-user collaborative filtering, and item-item collaborative filtering. Suppose you are working at a company and are building a movie recommendation service.

(a) (3 points) Initially, there will be very few customers of your system. Which technique above would you use for the best performance? Briefly explain why.

(b) (4 points) When using collaborative filtering, a utility matrix like below is used to predict the ratings of users on movies. Suppose we have a utility matrix where there are N customers and M movies.

	HP1	HP2	TW	SW1	SW2	...
A	4	5		5	1	...
B		3	4	3	1	...
C	2		1	3		...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

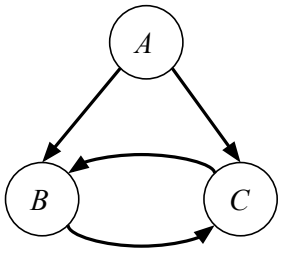
In terms of N and M, what is the time complexity (i.e., use the big-O notation like $O(M^2)$ mentioned in class) of predicting the ratings of a customer using user-user collaborative filtering? Briefly explain why.

(c) (4 points) In terms of N and M, what is the time complexity of predicting the ratings of a customer using item-item collaborative filtering? Briefly explain why.

(d) (4 points) Continuing from (b), suppose the utility matrix is very sparse where most customers rated a small number of movies, and only a handful of customers rated most of the movies. In terms of N and M, what is the time complexity for user-user collaborative filtering? Briefly explain why.

7 (15 points) Link Analysis

Consider the following Web graph of three pages A, B, and C:



Do the following link analyses on the pages. For all the questions, a correct expression is sufficient, and you do not have to give the actual number.

(a) (5 points) Compute the PageRank of each page where the taxation parameter is $\beta = 0.9$.

PageRank of A:_____

PageRank of B:_____

PageRank of C:_____

(b) (5 points) Compute the TrustRank of each page where the taxation parameter is $\beta = 0.9$, and the teleport set of trustworthy pages is $\{B\}$.

TrustRank of A:_____

TrustRank of B:_____

TrustRank of C:_____

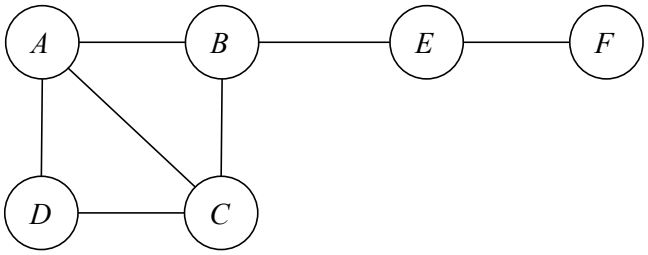
(c) (5 points) Compute the hubbiness and authority scores **h** and **a** for the pages. The largest components of **h** and **a** must be 1.

Hubbiness scores **h**:_____

Authority scores **a**:_____

8 (15 points) Mining Social-Network Graphs

Consider the following social-network graph.



(a) (5 points) Compute the betweenness scores for the following edges.

Score for $A - B$:_____

Score for $A - C$:_____

Score for $A - D$:_____

Score for $B - E$:_____

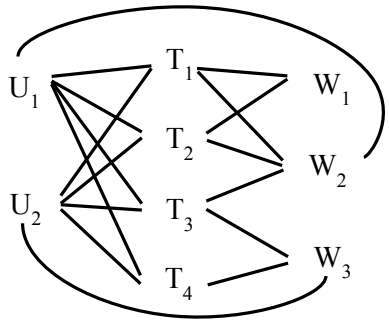
Score for $E - F$:_____

(b) (4 points) Find the best normalized cut in the same graph and divide the graph into two communities. What are the set of edges in the cut and the normalized cut value?

Set of edges in cut:_____

Normalized cut value:_____

(c) (3 points) In the following graph, count the number of heavy-hitter triangles.



Number of heavy-hitter triangles:

(d) (3 points) Recall the algorithm for finding triangles covered in class counted the heavy-hitter triangles and the other triangles separately. Why does this strategy make the algorithm efficient? Briefly explain.

9 (15 points) Large-scale Machine Learning

(a) (6 points) Suppose the following four examples constitute a training set:

([2, 2], +1)	([3, 1], +1)
([0, 0], -1)	([0, 1], -1)

Find the hyperplane $\mathbf{w}.\mathbf{x} + b = 0$ that separates the positive and negative examples with the maximum margin where all points are outside the margin (i.e., for each training example ($\mathbf{x}$, y), you have $y(\mathbf{w}.\mathbf{x} + b) \geq +1$).

w :_____

b :_____

Margin:_____

(b) (3 points) Suppose we are constructing an SVM with approximate separators on a linearly separable training set as in (a). If the regularization parameter C is set to a small value, is it possible for some of the points to fall within the margin hyperplanes (i.e., between $\mathbf{w}.\mathbf{x} + b = -1$ and $\mathbf{w}.\mathbf{x} + b = +1$)? Briefly explain why.

(c) (6 points) Suppose we have the following table where we would like to construct a decision tree by splitting the set of training examples below based on the numeric Population feature. For the impurity measure, use accuracy, which is defined as $1 - \max(p_1, p_2)$ where p_1 is the fraction of countries that like Soccer, and p_2 is the fraction of countries that like Baseball.

Country	Population	Sport
A	10	Soccer
B	500	Soccer
C	50	Baseball
D	100	Baseball
E	1000	Soccer

Select any test that minimizes the weighted-average impurity, which is the average of the impurities of the left and right children where each child's impurity is weighted by the fraction of countries in that child. What are the test, children, and weighted-average impurity?

Test at root node:_____

Countries in left child:_____

Countries in right child:_____

Weighted-average impurity:_____

10 (10 points) Mining Data Streams

(a) (5 points) Suppose there is a stream of positive integers that can be stored in 3 bits, and we use the extended version of the DGIM method to estimate the sum of the last k integers for any $1 \leq k \leq N$ where N is the window size. Recall the extension involves counting bits in each position and combining the counts into the estimated sum of integers. Suppose the window size is $N = 8$, and the stream is 12175313. What is the estimated sum for the last $k = 4$ positions? (The actual sum is 12.)

(b) (5 points) Suppose we extend the AMS algorithm to estimate the third moment of the following stream of length $n = 10$.

Stream: $a, b, c, d, a, b, c, a, b, a$

What is the actual 3rd moment of the stream?

For the AMS algorithm, suppose we pick the 5th and 6th positions to define the two variables X_1 and X_2. What is the average of the 3rd moment estimates?

11 (10 points) Advertising on the Web

Suppose there are three advertisers A, B, and C. There are three queries x, y, and z. Each advertiser has a budget of 2. Advertiser A bids on x, y, and z; B bids on x and y; C bids on x only. Suppose the query sequence is $xxxyyz$ (i.e., three x's followed by two y's followed by one z). Also when there are ties to be broken, use alphabetical order (e.g., if A and B are tied, then the query is assigned to A) instead of arbitrarily order.

(a) (2 points) What is the revenue of the greedy algorithm?

(b) (2 points) What is the revenue of the BALANCE algorithm?

(c) (2 points) Based on the result from (b), the competitive ratio of the BALANCE algorithm is at most:

(d) (4 points) Find a query sequence of length 6 where the BALANCE algorithm has a competitive ratio that is lower than that of (c).

Query sequence:_____

Competitive ratio:_____

[This can be used for scratch paper.]