CMSC5741 Big Data Tech. & Apps.

Assignment 1

Due Date: 23:59 Oct.27, 2018

Submission Instruction:

For this assignment, please submit electronic version only. We don't accept hard copy. For the programming questions, you need to submit BOTH your codes and your results. Submit codes as zipped tar file and the output of your program in plain-text file. You should place the relevant files (if any) in their separate directory (preferable 1, 2, 3… for each question). For other questions, answer them in **one word** document.

Please compress all your files as one zip file named by your student id, e.g., 10xxxxxxxx.zip, and submit to cmsc5741@cse.cuhk.edu.hk with email title "CMSC5741 Asg#1, Your name, Your student ID".

Policy on Late Submission:

1). Late within 2 days - 30% deduction.

2). Late after 2 days - not allowed.

Part A. MapReduce Programming (25 points)

The following problem is based on lecture 2.

In this problem, you need to write a MapReduce program to find the top-N pairs of similar users from a repository called Movielens. Provided is the dataset of Movielens 20M, which can be found in http://grouplens.org/datasets/movielens/20m/. The original range of the rating scores is in [0.5,5.0]. For simplicity, we ignore the missing values in the ratings. You need to calculate the top-N pairs of similar users based on the ratings. In order to calculate the similarity, we redefine the rating score in [0.5,2.5] as `unlike', and the rating score in (2.5,5.0] as `like'. Thus you can preprocess the dataset as (user, movie, `like') or (user, movie, `unlike'). Given a user i, you can construct a set of movies with `like' (denoted by $L_i$), and also a set of movies with `unlike' (denoted by U_i). Then, for a pair of users of (i, j), you can calculate the similarity between them via the following metric as

$$Jaccard = \frac{(L_i \cap L_j) \cup (U_i \cap U_j)}{(L_i \cup L_j) \cup (U_i \cup U_j)}$$

You output the top-N pairs of similar users based on the values of Jaccard. (Hint: You can leave the final sorting step to a UNIX utility called "sort". Your MapReduce program needs only to calculate the metric of Jaccard.)

Please write a map reduce program (one mapper and one reducer) to list the **top-100 pairs of similar users with similarity (from most similarity to least similarity)**. Each line should have a pair of users' ID and the similarity. A valid example is as follows:

"2" "3" 0.75

For this problem, please submit all your codes as a zipped tar file and name the result files as **A.txt** under the directory **A**.

Part B. Frequent Itemsets (20 points)

The following problem is based on lecture 2.

Suppose there are 10000 items, numbered 1 to 10000, and 10000 baskets, also numbered 1 to 10000. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all the even-numbered baskets, and so on. Basket 12 consists of items {1, 2, 3, 4, 6, 12}, since these are all the integers that divide 12.

Write the A-Priori algorithm to answer the following questions. Each line has a frequent itemset in ascending order, separated by a space. Itemsets with larger size should always be listed after itemsets with smaller size. If two itemsets are of the same size, they should be listed in ascending order as well. For example, the following outputs are in valid format:

1 2 3

1 3 4

1 2 3 4

1 3 4 5

1 5 6 8 9

1. (10 points) If the support threshold is 100, which items are frequent?
2. (10 points) If the support threshold is 20, find the maximal frequent itemsets, i.e., frequent itemsets with the largest size.

For these problems, please submit your code as a zipped tar file and name the result files as A.txt and B.txt for corresponding questions under the directory B.

Part C. Locality Sensitive Hashing (10 points)

The following problem is based on lecture 3.

This problem is related to the concept about shingling, minhashing, and locality sensitive hashing. Suppose you are given the following five sentences:

I. abbcba
II. ccbbca
III. bbaacb
IV. bbacab
V. cbbbac

1. (3 points) Calculate the set of 2-shingles for each sentence and use matrix to represent the sentences, where the element is enumerated from 0.
2. (3 points) Compute the minhash signature for each column if we use the following three hash functions: $h1(x) = 2x + 1 \bmod 6$; $h2(x) = 8x + 2 \bmod 10$; $h3(x) = 6x + 2 \bmod 10$.
3. (2 points) Which of these hash functions are true permutations?
4. (2 points) How close are the estimated Jaccard similarities for the 10 pairs of columns to the true Jaccard similarities?

For this problem, please write your answers in the word file.

Part D. Mining Data Streams (25 points)

The following problem is based on lecture 4.

Usually we use DGIM algorithm to count ones in a bit stream. Now we have a ten-thousand-bit stream (provided in cmsc5741_stream_data.txt). You are required to program to count ones in last one thousand bits for this stream through DGIM algorithm. Output should demonstrate how you set the buckets and your estimate. **You are required to implement the algorithm in $O(\log^2 N)$ space complexity with $O(\log \log N)$ bucket size.**

Updating rules:

(1) If the current bit is 0, no other changes are needed;
(2) If the current bit is 1:
   (a) Create a new bucket of size 1, for just this bit, and end timestamp = current time;
   (b) If there are now three buckets of size 1, combine the oldest two into a bucket of size 2;
   (c) If there are now three buckets of size 2, combine the oldest two into a bucket of size 4;
   (d) And analogize for the rest.

Hints: you cannot directly count ones, which is wrong answer.

For this problem, please submit your code as a zipped tar file under the directory **D** and describe your buckets (the separation of buckets) and your estimation in the **word file**.

Part E. Scalable Clustering (10 points)

The following problem is based on lecture 5.

Practice the k-means algorithm using Euclidean distance.

Suppose there are 15 data points A1(10,4), A2(11,5), A3(10,6), A4(9,6), A5(6,2), A6(6,4), A7(7,2), A8(5,3), A9(6,10), A10(5,10), A11(5,12), A12(1,7), A13(1,8), A14(2,9), A15(2,9). Assume we cluster these 15 data points into 4 clusters. The initial seeds are A2, A7, A9, A12. After running the k-means algorithm for 1 epoch only,

answer the following questions one by one:

1. (3 points) The new clusters (i.e., the examples belonging to each cluster)
2. (3 points) The centers of the new clusters
3. (4 points) How many more iterations are needed to converge? Please write down the final centers.

For this problem, write down your answers for all questions in the word file.

Part F. Dimensionality Reduction (10 points)

The following problem is based on lecture 6.

1. (3 points) Compute the eigenvalues and eigenvectors of matrix $C_1 = C^T C$, where $C = \begin{bmatrix} 0 & 6 \\ 2 & 0 \\ 0 & 8 \end{bmatrix}$.
2. (3 points) Compute LU Decomposition on matrix $C_2 = \begin{bmatrix} 1 & 3 & 4 \\ -2 & -8 & -9 \\ 3 & 9 & 10 \end{bmatrix}$.
3. (4 points) Given a matrix $C_3 = \begin{bmatrix} 2 & 0 & 4 & 0 & 1 & 0 \\ 3 & 1 & 0 & 0 & 0 & 5 \\ 0 & 1 & 0 & 2 & 1 & 0 \\ 0 & 4 & 0 & 0 & 0 & 1 \\ 4 & 0 & 0 & 3 & 3 & 0 \\ 0 & 1 & 0 & 4 & 0 & 0 \end{bmatrix}$ and $r = 2$, please compute the CUR approximation. Here we assume the random selection of rows is 3 and 5, and random selection of columns is 1 and 3.

For this problem, write down your answers for all questions in the word file. Show all your computations step-by-step.