CSE 482: Big Data Analysis (Fall 2022) Homework 3
Due date: Monday, December 5, 2022

Submit the homework via D2L.

1. Consider the transaction data shown in Table 1 to answer the following questions. Assume that the minimum support threshold for frequent itemsets is 40%.

Table 1: Transaction data for Question 1.

Transaction ID	Items Purchased
1	Bread, Coffee, Sugar
2	Bread, Milk
3	Bread, Butter, Milk
4	Coffee, Milk
5	Bread, Butter, Cookies
6	Coffee, Milk, Sugar
7	Bread, Eggs, Milk, Sugar
8	Bread, Butter, Cookies
9	Bread, Butter, Eggs, Milk
10	Coffee, Eggs, Milk

   (a) List all the frequent itemsets extracted from the data along with their support values.

   (b) Based on your answer in part (a), extract all association rules whose confidence is more than 50% and support is at least 40%. The rule must contain at least 1 item on both its left and right hand side (i.e., ignore rules with empty antecedent or consequent). Note: do not use the Apriori implementation in Exercise 6 (if released) since the software applies the minimum support threshold to the left-hand side of the rule.

2. Download the sample Wikipedia dataset `articles.txt` from the class website. The dataset contains the list of articles edited by a sample of Wikipedia users in 2005. Each line in the file corresponds to a "transaction" (user) and each "item" corresponds to an article edited by the user. In this exercise, you will apply the apriori software from `http://www.borgelt.net/apriori.html` to the articles.txt file. Generate the set of association rules that contain 2 items or more from the given dataset. Set the minimum support threshold to be 0.03% and minimum confidence threshold to be 70%. Store the results in a file named rules.txt and submit it with your homework solution. Note that, in this Apriori implementation, given a rule $X \rightarrow Y$, the support threshold is applied to X only instead of $X \cup Y$.

3. Consider the following set of one-dimensional points: {0.090, 0.172, 0.310, 0.335, 0.429, 0.640, 0.642, 0.851}. All the points are located in the range between [0,1].

   (a) Suppose we apply kmeans clustering to obtain three clusters, A, B, and C. If the initial centroids of the three clusters are located at {0.15, 0.25, 0.90}, respectively, show the cluster assignments and locations of the centroids after the first two iterations by filling out the following table. Did the algorithm converges (i.e., did the centroid locations change) after two iterations?

	Cluster assignment of data points (enter A, B, or C)								Centroid Locations		
Iter	0.090	0.172	0.310	0.335	0.429	0.640	0.642	0.851	A	B	C
0	-	-	-	-	-	-	-	-	0.15	0.25	0.90
1											
2											

   (b) Repeat part (a) using {0.10, 0.45, 0.90} as the initial centroids. Show the cluster assignments and locations of the centroids after the first two iterations by filling out the following table. Did the algorithm converges after two iterations?

	Cluster assignment of data points (enter A, B, or C)								Centroid Locations		
Iter	0.090	0.172	0.310	0.335	0.429	0.640	0.642	0.851	A	B	C
0	-	-	-	-	-	-	-	-	0.10	0.45	0.90
1											
2											

   (c) Compute the sum-of-squared errors (SSE) for the clustering solution in part (a).

   (d) Compute the sum-of-squared errors (SSE) for the clustering solution in part (b). Which solution is better in terms of their SSE?

4. Consider the following set of one-dimensional points: {0.10, 0.20, 0.35, 0.55}. All the points are located in the range between [0,1].

   (a) Compute the pairwise Euclidean distance between the data points. Show the resulting 4×4 distance matrix.

   (b) Apply the single link (MIN) algorithm to cluster the data points. Show the resulting dendrogram. Make sure you indicate the distance (when the clusters were merged) on the y-axis of the dendrogram. Compute the sum-of-squared errors (SSE) when the number of clusters is equal to 2.

   (c) Apply the complete link (MAX) algorithm to cluster the data points. Show the resulting dendrogram. Make sure you indicate the distance (when the clusters were merged) on the y-axis of the dendrogram.

Compute the sum-of-squared errors (SSE) when the number of clusters is equal to 2.

(d) Apply k-means with Euclidean distance on the data set with $k = 2$. Assume the two initial centroids are located at: 0.15 and 0.5, respectively. Compute the SSE of the resulting clusters. Compare the clustering result against MIN and MAX (with $k = 2$). Which method(s) produce similar clustering results? Which method(s) produce a solution with lowest SSE?

5. Consider the confusion matrices shown below, where the rows correspond to the clusters and the columns correspond to the ground truth (ideal) solution. The matrices summarize the performance of two clustering algorithms, A and B, on a data set that contains 50 instances from class 1 and 50 instances from class 2.

Algorithm A	Ground truth class	
	Class 1	Class 2
Cluster 1	10	35
Cluster 2	40	15

Algorithm B	Ground truth class	
	Class 1	Class 2
Cluster 1	35	40
Cluster 2	15	10

Each entry in the matrix corresponds to the number of data points from a given class assigned to a cluster. Calculate the adjusted rand index (ARI) for both algorithms A and B. Which clustering solution is better in terms of their ARI?

6. Write the corresponding HDFS commands to perform the following tasks. Each of these tasks must be accomplished with a single HDFS command. Hint: type `hadoop fs -help` for the list of commands available. Note the difference between the local (Linux) file directory and HDFS directory. To double-check your answers, you're encouraged to test the commands (on your own HDFS directory) to make sure they work correctly.

   (a) Make a directory on HDFS named `/user/hadoop`.

   (b) Upload a file named `data.txt` from the local filesystem to HDFS.

   (c) Rename the uploaded file on HDFS from `data.txt` to `temp.txt`

   (d) Move the file from its current location at `/user/hadoop/temp.txt` to `/user/hduser/temp.txt`.

   (e) Copy the file from `/user/hduser/temp.txt` to `/user/hadoop/data.txt` (note that the filename is also changed).

   (f) Display the content of the file `/user/hduser/temp.txt`.

   (g) Delete the file `/user/hadoop/data.txt`.

   (h) Download the file named `/user/hduser/temp.txt` to the local filesystem directory `/user/cse482`.

(i) Merge all the files under the HDFS directory `/user/hadoop/results` into a single file `output.txt` to be stored in your current working directory.

(j) Changing permission of the HDFS directory `/user/hduser` so that only the user hduser can read and write files stored in the directory (but not other users).

7. Write a Hadoop program that computes the correlation between pairs of stocks from the stock market dataset `stocks.csv` provided on the class webpage. Every line in the data file has the following information:

   `date`, `AAPL`, `BAC`, `C`, `F`, `GOOG`, `HMC`, `MSFT`, `TYM`, `WFC`

   where the first column is the trading date, followed by the closing prices for 9 different stocks (same data was used in homework 3). An example of the output reducer file may look as follows:

   (`AAPL`, `BAC`) 0.42344

   Deliverable: Your hadoop source code (*.java), the archived (jar) file, and the output result file (part-r-00000) that contains the pairwise correlation values.