Assignment 1
Data can do amazing things!

Homeworks Guidelines and Policies

- It is expected that the students submit an assignment report (HW1_[student_id].pdf) as well as required source codes (.m or .py) into an archive file (HW1_[student_id].zip). Please combine all your Persian reports just into a single .pdf file by problems order. Code without report has an exact zero point.

  Please do not use implementation tools when it is asked to solve the problem by hand, otherwise you will be penalized and lose some points.
- You are free to solve by-hand problems on a paper and include their pictures in your report. Here, cleanness and readability are of high importance. Images should also have appropriate quality.
- Your work will be evaluated mostly by the quality of your report. Do not forget to explain your answers clearly, and provide enough discussions when needed.
- In each homework, 4 points (out of a possible 100) belong to compactness, expressiveness, and neatness of your report and codes.
- By default, we assume you implement your codes in Python. If you are using MATLAB or R, you have to use the equivalent functions when it is asked to use specific Python functions.
- Your codes must be separated for each question.
- Make sure you have access to Courses, because that is where all assignments as well as course announcements are posted. Homework submissions are only made through Courses.

- Please submit your work **before the end of April 7th**.
- There will be a 10% penalty for every late day.
- You are not allowed to share codes/answers or use works from the past semesters. Violators will receive a zero for that particular problem.
- If there is any question, please do not hesitate to contact us through the Telegram group chat or following email addresses: **m.ebadpour@aut.ac.ir**, **faramarz.aghajani@aut.ac.ir**, and **danial.alizadeh@aut.ac.ir**.

1. Power of Mapping and Reducing (32 Pts.)

In this section, there are two questions, each of which is related to a separate dataset that has been provided to you. Please read the questions carefully and do the requested things for each one; Some parts may require you to give more explanations and please use the language you are more fluent in explaining because the readability of your answers is important.

1. In this question we are going to use a massive dataset that contains information about customer purchases from an e-commerce platform. Each entry in the dataset represents a single purchase and includes details such as customer ID, item purchased, price, and timestamp.
    a. Analyze this dataset to find the total revenue generated by each product category during a given period.
    Hint: To achieve this goal, use the MapReduce programming model.
    b. Find the most popular customers in terms of their purchase volume. (The report of the first 10 customers is enough).
    c. State the key steps you used to implement this solution using Python and MapReduce.
    d. Describe how you designed your mapper and reducer functions to efficiently process this dataset.
2. In this question, we are going to use a massive dataset that contains information about user interactions on a social media platform. Each entry in the dataset represents a user action, such as sending a message, liking a post, or following another user. The dataset contains details such as user IDs, action types, timestamps, and related identifiers (eg post IDs, user IDs).
    a. Analyze this dataset to identify the most influential users based on their engagement (Reporting the top 10 users is sufficient).
    Hint: Define influence as a combination of the number of followers a user has and the amount of interaction their posts receive (likes, comments, shares).
    b. Explain how to approach this problem using the MapReduce programming model.
    c. Identify the key steps involved in implementing this solution, including designing your mapper and reducer functions.
    d. Discuss the challenges you foresee in efficiently processing such a large and complex dataset.

2. Drown deep in realm of frequent itemsets mining (32 Pts.)

1. Suppose items are {1,2,3,4,5}. Here is a collection of fifteen baskets.

Basket_ID	Basket_Items	Basket_ID	Basket_Items
1	1,2,3,5	**9**	2,3,4
2	1,2,3	**10**	1,5
3	2,3,4	**11**	2,3
4	1,4,5	**12**	1,4,5
5	1,2,3	**13**	3,4,5
6	1,3,5	**14**	2,3
7	1,2,5	**15**	1,5
8	1,4,5	-	-

On the first pass of the PCY Algorithm, we use a hash table with 7 buckets, and the set {i, j} is hashed to bucket (i × j) mod 7. The support threshold is 7.

a) compute the support for each item and each pair of items.

b) Which pairs hash to which buckets?

c) Which pairs are counted on the second pass of the PCY Algorithm?

2. Suppose items are $\{A,B,C,D,E,F,G,H\}$. Which of the following association rules has a confidence that is <u>certain</u> to be at least as great as the confidence of $A,B,C \rightarrow D,E,F,G,H$ and no greater than the confidence of $A,B,C,D,E \rightarrow H$?

a) $A,B,C,D \rightarrow E,G,H$

b) $C,D,E \rightarrow A,B,G,H$

c) $A,B,C,D,E \rightarrow G,H$

d) $A,B,C,E \rightarrow D,G,H$

e) $A,B,C \rightarrow F$

3. In this exercise, we intend to analyze the dataset related to shopping baskets using the A-Priori algorithm. The file "store.csv" contains a collection of shopping baskets from a store. Each row in this dataset represents a shopping basket.

a) Report the 10 most frequent items (minimum support for an item is 200).

b) Report the 10 most frequent pairs of items (minimum support for pairs of items is 100).

c) Report the 10 most frequent triples of items (minimum support for triples of items is 75).

d) For all the frequent triples (A,B,C), calculate the confidence scores for the rules $(A,B) \rightarrow C$, $(A,C) \rightarrow B$, $(B,C) \rightarrow A$ and report the top 5 rules based on confidence.

e) Repeat part "d" using the interest criterion.

f) "Lift" is another measure used in association rule mining. What information does it provide, and how is it calculated?

g) Repeat part "d" using the lift criterion.

Note: Using libraries for implementing A-Priori is not allowed.

Corresponding TA: Aghajani

3. Unveiling the Mysteries of Poet Recognition Through Poetic Associations (32 Pts.)

Imagine you are a curator in a museum of literary treasures, tasked with identifying the creators behind anonymous works of art. Your latest challenge involves unraveling the mystery behind a collection of poetic verses from ancient Persia. Each poem is a masterpiece, meticulously crafted by one of five renowned poets: Khayyam, Ferdowsi, Hafez, Saadi, and Rumi. These poets have left behind a legacy of profound verses that have stood the test of time.

Your mission, should you choose to accept it, is to employ the power of data mining techniques to uncover the hidden identity of each poet based solely on the distinct patterns and motifs present in their poetic compositions. Drawing inspiration from the concept of frequent itemset mining, you will embark on a journey to decode the unique fingerprints of these literary giants.

In your possession are five collections of poems, each attributed to a different poet: Khayyam, Ferdowsi, Hafez, Saadi, and Rumi. These poems, written in the Persian language, are stored in separate text files. Your task is to devise a method that can automatically determine the poet behind a given set of verses.

To accomplish this, you will utilize frequent itemset mining algorithms, specifically the A-Priori and PCY algorithms. In the context of this problem, each poem will be treated as a "basket," with the individual words or sequences of words serving as the "stocks" or items within each basket.

a. Split the available poems into training and test datasets. This step ensures that your model can learn from a subset of the data and evaluate its performance on unseen samples. Use 8:2 ratio. Do not forget the Shuffling. Name your other preprocessing steps and challenges.

b. Implement the A-Priori and PCY algorithms to identify frequent itemsets within the training dataset. In this case, each item represents a unique word or sequence (or subset) of words from the poems. You are free to use your own ideas. Just mention them in your report in details. You are not allowed to use built-in and ready-to-use functions and libraries. In your report, discuss about structure of implementation and used parameters like minimum support threshold and etc. How do you reach to optimal hyperparameters? What are their importance?

c. Based on the frequent itemsets discovered, generate association rules that capture the relationships between words or sequences of words within the poems. These rules will serve as the key indicators for identifying the poet behind a given set of verses. For each poet, generate a .txt file involved his related rules.

d. When presented with a new set of verses (test dataset), apply the generated association rules to classify the poet responsible for those verses. By matching the patterns observed in the test data against the learned associations from the training data, you can infer the likely poet with a degree of confidence. Again, you are free to apply your own idea! One possible strategy is checking all rules of each poet and assign and score followed by comparing scores of poets. For this part, calculate accuracy of classification. From each poet, generate a .txt file which shows that each poem of him assigned to whom.