Data Mining
Information Systems Department
Faculty of Computers and Artificial Intelligence
Cairo University

Assignment 2
Clustering

Instructions:

1. Assignment should be done individually , copies or any other method of cheating will be graded to - 5.
2. Each student can solve one problem or both, and in the second case we will consider the higher mark.
3. Total grade is 5 marks.
4. The deadline for this assignment will be on 20-April 11:55 on classroom , no late submissions are allowed.
5. Discussion will be during the office hours of Eng. Doaa Ghaleb I'll send you the discussion table on 20-April .
6. Your program should include a graphical user interface.
7. The interface should enable user to select the percentage of the data needed to be read from the input file e.g. if the file contains 100 records, and the user needs to read 70% of the file then the analysis should be done on 70 records only.
8. The program should enable the user to select the file needed to be clustered,
9. Initial centroid should be choosing randomly.
10. You should detect the outliers (if exists).
11. Using the programming language, you prefer, write a program with the following specifications:
    a. Inputs:
        i. A file with a set of transactions (e.g., Excel, text, or CSV file) – (The file attached).
        ii. The percentage of the data needed to be read from the input file.
        iii. Number of clusters - K - will be provided from the user as an input.
    b. Outputs
        i. The final output of your program should show the content of each clusters and show outlier records.

Problem 1

Description:

- Consider this dataset which is scraped from IMDB's official website in the imdb_top_2000_movies.csv file. It contains the ratings of top 2000 movies between 1921 and 2010.
- Write a program in any programming language you prefer to group the movies based on the similarity of user IMDB ratings.
- You should use k-means algorithm to cluster the movies to k clusters
- Number of clusters (k) will be provided from the user as an input.
- Initial centroid should be choosing randomly.
- You should use Euclidean distance as your distance function.
- You should detect outlier data (if exists).
- The final output of your program should show k lists of users and show outlier user's records.

Problem 2

Description:

- Consider Facebook live interactions dataset in Facebook_live.csv file, it contains statistics regarding different interactions (e.g. comments, shares, like, love, wow, ...etc.) in different type of posts represented by status type column.
- You are required to implement k-means clustering in any programming language you prefer to find intrinsic groups within the dataset that display the same status_type behavior.
- Notice that the status_type behavior variable consists of posts of a different nature (video, photos, statuses and links).
- Number of clusters (k) will be provided from the user as an input.
- Initial centroid should be choosing randomly.
- You should use Manhattan distance as your distance function.
- You should detect outlier data (if exists).
- The final output of your program should show k lists of states and show outlier product's records.