Homework 03: Data Mining and Text Mining

Homework Instructions

This assignment covers **Association Rule Mining**, **Cluster Analysis**, focusing on **Iterative**, **Hierarchical, and Density-Based clustering algorithms**. You will need to use Python, specifically the *dmba* library, to load datasets and perform the required analyses for each question. Follow the installation instructions below to get started:

```
pip install dmba
import dmba
```

Each question provides a hint about the dataset you need to load using the *dmba* library.

Question 1: Groceries Recommendations. (20 points)

You will apply association rule mining to analyze patterns in transactional data and generate product recommendations. You will conduct this on Groceries dataset, which contains information on products purchased together by customers in different transactions.

Dataset: You are provided with a CSV file named **groceries.csv**, where each row represents a customer transaction. Each transaction lists the products purchased by a customer, separated by commas.

- *Sample Row: milk, bread, eggs*

You will need to do the following:

A. Load and Preprocess Data:
   - Load the dataset using Pandas.
   - Convert the dataset into a format suitable for association rule mining. You may use the TransactionEncoder from mlxtend.preprocessing to transform the data into a one-hot encoded format.

B. Generate Frequent Itemsets:
   - Using the mlxtend.frequent_patterns.apriori function, identify frequent itemsets with a minimum support threshold of 0.01 (i.e., 1%).
   - Display the top 10 frequent itemsets by support.

C. Generate Association Rules:
   - Using mlxtend.frequent_patterns.association_rules, generate association rules from the frequent itemsets with a minimum confidence threshold of 0.2 (i.e., 20%).
   - For each rule, display :
     - Antecedents (items that lead to the recommendation)
     - Consequents (recommended items)
     - Support, Confidence, and Lift values.

D. Recommend Items:
   - Create a function recommend_items(transaction: list) -> list that takes a list of items purchased by a user and returns a list of recommended items based on the association rules generated.
   - For example, if the input transaction is ["milk", "bread"], the function should return items that frequently co-occur with "milk" and "bread" based on the generated rules.

E. Provide analysis:

- Write a brief analysis on how association rule mining helped in uncovering patterns and generating recommendations. Discuss the potential limitations of this approach in a recommendation system.

To get you started here is a pseudocode for you:

```
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules
from mlxtend.preprocessing import TransactionEncoder

# Step 1: Load and Preprocess Data
# Load the dataset and preprocess it for association rule mining

# Step 2: Generate Frequent Itemsets
# Use the apriori algorithm to find frequent itemsets

# Step 3: Generate Association Rules
# Generate association rules based on the frequent itemsets

# Step 4: Recommend Items
# Define a function to recommend items based on association rules
```

Question 2: University Rankings. (20 points)

The dataset on American College and University Rankings (available from www.dataminingbook.com) contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements, including continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or public school).

Note that many records are missing some measurements. Our first goal is to estimate these missing values from “similar” records. This will be done by clustering the complete records and then finding the closest cluster for each of the partial records. The missing values will be imputed from the information in that cluster.

A. **Data Cleaning**: Remove all records with missing measurements from the dataset.

B. **Hierarchical Clustering**: For all continuous measurements, apply hierarchical clustering using complete linkage and Euclidean distance. Be sure to normalize the data before clustering. Based on the dendrogram, how many clusters seem appropriate for this dataset?

C. **Cluster Characterization**: Compare the summary statistics (e.g., mean or median) for each cluster. Describe the characteristics of each cluster in context (e.g., "Universities with high tuition, low acceptance rates..."). *Hint*: You can use the `pandas groupby(clusterlabel)` method along with aggregation methods like `mean` or `median` to summarize each cluster.

D. **Categorical Analysis**: Use the categorical variables (State and Private/Public) that were not part of the clustering to describe each cluster. Is there any noticeable relationship between the clusters and these categorical variables?

E. **External Information**: What other external information could help explain the characteristics of some or all of the clusters?

F. **Missing Data Imputation**: Consider *Harvard University*, which has missing data. Compute the Euclidean distance between *Harvard* and each of the clusters you identified earlier, using only the available measurements. Which cluster is *Harvard* closest to? Impute the missing values for *Harvard* by taking the average of that cluster's corresponding measurements.

Hint: *df = dmba.load_data('Universities.csv')*

Question 3: Customer Rating of Breakfast Cereals. (**20 points)**

The dataset **Cereals.csv** includes nutritional information, store display, and consumer ratings for 77 breakfast cereals.

Data preprocessing. Remove all cereals with missing values.

A. Apply hierarchical clustering to the data using Euclidean distance to the normalized measurements. Compare the dendrograms from single linkage and complete linkage, and look at cluster centroids. Comment on the structure of the clusters and on their stability. (Hint: To obtain cluster centroids for hierarchical clustering, compute the average values of each cluster members, using groupby() with the cluster centers followed by mean: dataframe.groupby(clusterlabel).mean().)

B. Which method leads to the most insightful or meaningful clusters?

C. Choose one of the methods. How many clusters would you use? What distance is used for this cutoff? (Look at the dendrogram.)

D. The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet. For this goal, you are requested to find a cluster of “healthy cereals.” Should the data be normalized? If not, how should they be used in the cluster analysis?

Hint:

df = dmba.load_data('Cereals.csv')

Question 4: Marketing to Frequent Fliers. (20 points)
The file **EastWestAirlinesCluster.csv** contains information on 3999 passengers who belong to an airline's frequent flier program. For each passenger, the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers.

A. Apply hierarchical clustering with Euclidean distance and Ward's method. Make sure to normalize the data first. How many clusters appear?
B. What would happen if the data were not normalized?
C. Compare the cluster centroid to characterize the different clusters, and try to give each cluster a label.
D. To check the stability of the clusters, remove a random 5% of the data (by taking a random sample of 95% of the records), and repeat the analysis. Does the same picture emerge?
E. Use k-means clustering with the number of clusters that you found above. Does the same picture emerge?
F. Which clusters would you target for offers, and what types of offers would you target to customers in that cluster?

Hint:
df= dmba.load_data('EastWestAirlinesCluster.csv')

Question 5: Discovering Frequent Flyer Groups with DBSCAN (Density-Based Spatial Clustering of Applications with Noise). (20 points)

The dataset **EastWestAirlinesCluster.csv** contains information on 3999 passengers who belong to an airline's frequent flier program. For each passenger, the data include various details on their mileage history and different ways they accrued or spent miles in the last year.

The goal is to use DBSCAN to identify different groups of passengers based on their flight and mileage behaviors. Since DBSCAN doesn't require specifying the number of clusters beforehand, it's well-suited for identifying irregularly shaped clusters and noise points, which may represent outliers or passengers with unusual patterns.

A. Prepare the Data

- Select appropriate numerical features for clustering.
- Normalize the data to ensure that all variables contribute equally to the distance metric.

B. Finding the Optimal Parameters for DBSCAN

- Use the NearestNeighbors approach to determine the optimal epsilon (eps) parameter by examining the elbow plot.
- Set min_samples to 5 (default for DBSCAN) and apply DBSCAN to the normalized data.

C. Analyzing the Results

- How many clusters did DBSCAN identify? How many points were classified as noise (outliers)?
- Examine the cluster centroids to understand the characteristics of each group.
- How does the presence of noise points help in identifying outliers? Discuss the characteristics of these outliers.

D. Cluster Insights and Recommendations

- Based on the identified clusters, which group of frequent flyers would you target for special offers, and why?
- Provide one marketing strategy per cluster, tailored to the flight behavior and spending characteristics of that group.