

MIDDLE EAST TECHNICAL UNIVERSITY, NORTHERN CYPRUS CAMPUS

CNG514 Data Mining

NG213 Data Structures – Project 1

**Assignment 3: Association Mining and Clustering**

**Date handed-out: 20 May, Friday 2022**

**Date submission due: 5 June, Sunday 23:55**

**Purpose:**

The main purpose of this assignment is to revise the topics related to Association Rule Mining and also clustering practically. It has the following two parts.

**Part 1: Association Rule Mining**

Eye-tracking is a technique used to record fixations of a user – mainly the points where the user looked at. These points can be processed on a given stimuli and created as sequences. In this first part, you are given sequences of 36 people which shows where they looked at on the home page of Apple (see the file called “Apple\_sequence\_dataset” in the attached zip file). In this data file, each line represents a user’s sequence. Therefore, there are 37 lines. For example, the sequence of the first user is as follows:

[['I', '359'], ['E', '179'], ['C', '140']…..

This means that the user is looked at the segment labelled as I with 359 ms, then E with 179 ms, etc. The segments are the blocks on a web page which is visualized as below.



Figure 1 The home page of Apple page with segmented labelled from A-Z

Your task in this first part of the assignment is to process the sequences for the given Apple home page and find out the frequent itemsets of web page segments with Apriori algorithm. These frequent item sets will show the most popular segments and co-occurrences of these segments that people look at. Conduct the following experiment:

* Use minimum support as %75 and confidence of 75%, and identify frequent itemsets.

Beside this experiment, in your report discuss your frequent itemsets, for example, the length of the frequent itemsets, items (i.e., segments) in your itemsets, etc. and also discuss what happens if you change the support or confidence values.

**Part 2: Clustering**

In this part of the assignment, you are again given the data of 38 people from an eye-tracking experiment with the same web page – the home page Apple page (see “Apple\_fixation\_dataset” folder in the attached zip file).

This time, you are given the raw data for each user (file labelled as P-1 means user 1) with the following columns/attributes/features:

* *FixationIndex*: That shows the unique index of the fixation. It’s incremented by one and whenever the eye-tracking record a fixation, it increments it by one.
* *Timestamp*: That shows an incremental timestamp of the fixation.
* *FixationDuration*: This shows the duration of the fixation.
* *MappedFixationPointX*: This shows the X coordinate of the fixation. Please note that top-left coordinate is (0,0) and this value increases when you go horizontally.
* *MappedFixationPointY*: This shows the Y coordinate of the fixation. Please note that top-left coordinate is (0,0) and this value increases when you go vertically
* *StimuliName*: This shows the URL of the web page.

Your task in this assignment is to use the DBSCAN clustering algorithm and cluster these data points and then visualize it on the given web page image. When people’s eyetracking data was recorded they were asked to complete a searching task on the Apple web page that required the access of web page segments: E and B in that order (see Figure 1). Therefore, your task here is to experiment with the parameters of DBSCAN to identify the best fitting clustering for these segments and report your parameters. You can find the image of this page in the data folder called “Apple.png”. These clusters will show the parts of the page that are commonly looked at.

**Submission Instructions:**

Write a report that includes the following two parts:

* Part one will include the details of your experiment, the results from the first experiment and also the discussion of the results.
* Part two will include the details of your experiments with the DBSCAN and in particular the results which are visualized on the web page image. Discuss whether it would be possible to identify the segments that people looked at with this technique.

Submit a single ZIP file that includes all the **scripts** (source code), and also the **report**.

**Resources:**

You are strongly advised to use Phyton. Please check out these modules:

* <https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing>
* <https://www.scipy.org/>
* <https://numpy.org/>
* <https://matplotlib.org/>

However, if you would like to use Weka, Matlab or R, you can speak to the instructor.

**Grading Schema:**

|  |  |  |
| --- | --- | --- |
| **Item** | **Percentage (Total 100)** | **For Full mark..** |
| Apriori experiments and the results | 50 | Experiments are fully described with clear results and also they are discussed. |
| DBSCAN experiments, results and also discussion of the results | 50 | Experiments are fully described with clear results and also they are discussed. |