CS 634 DATA MINING MIDTERM PROJECT

KHANG TRAN

MAR 12, 2021

1. **INTRODUCTION:**

This project is the implementation of the Apriori algorithm to extract the association rules from the transactions. The input for the algorithm is the transactions and the output from the codes is the association rules. My project is flexible for new data which must meet some requirements and it’s also flexible for different support and confidence rate. The code can be found by this link: <https://github.com/khangtran2020/CS634_midterm> More details will be described below.

1. **SET UP:**
   1. **REQUIREMENTS:**

* Python 3.6
* Numpy
* Pandas
* Itertools
  1. **SET UP**
* First download the .zip file to a location that you will use and extract files to that directory.
* Or you can clone it: <https://github.com/khangtran2020/CS634_midterm>
* Open a terminal and run command: pip install -r requirements.txt

1. **RUNNING:**
   1. **Data Format:**

The transactions are put into a .csv file with the rows is the transactions and the columns is the transaction\_id and the transaction’s items. The transaction’s items are separated by a comma. In my project, I provided five datasets: amazon, nike, kmart, bestbuy and custom – all of these data are given in the data folder. Figure 1 show the data format of the amazon data.

Graphical user interface, text, application

Description automatically generated

*Figure 1: Data Format – Amazon Data.*

In order to run your own data, set up your transaction and put it into a csv file as figure 1 and remember the second columns contains the transaction. Then you put it into the data directory. And it’s all done.

* 1. **Running process:**

To run the program, open a terminal and go to the directory that contains the codes. Next you just have to put this command into the terminal: python MIDTERM\_PROJECT.py

First of all, the program will show up the data that’s available to use and their option – the input number to choose the data. To choose the data, input the number corresponding to the data as in figure 2 and press ENTER.

Text

Description automatically generated

*Figure 2: List of data and the option which is lie between the parentheses.*

After that the program will let you choose the support rate and the confidence rate. For the input the options and the support/confident rate will be make sure to be in the right form: support rate must lie in [0, 1] and the confidence rate must also be in this range.

Text

Description automatically generated

*Figure 3: optional Support rate and confidence rate.*

After that, the program will start working with the Apriori algorithm. After a while to process the data, will output the association rules with the confidence of that rules

Text

Description automatically generated

*Figure 4: Result*

1. **RESULT:**

Text

Description automatically generated

*Figure 5: Result for amazon dataset*

*Text

Description automatically generated*

*Figure 6: Results for kmart dataset*

*Text

Description automatically generated*

*Figure 7: Results for nike dataset.*