CS 634 DATA MINING MIDTERM PROJECT

KHANG TRAN

MAR 12, 2021

I. 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.

II. SET UP:

a. REQUIREMENTS:

- Python 3.6
- Numpy
- Pandas
- Itertools

b. 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`

III. CODE USAGE

The main part of my code lies in the figure 1 and figure 2 below. Figure 1 illustrates the codes that extract the frequent patterns from the transactions. First, it counts the number of existence of each item and compare to the `min_support` value. This process will scale down the number of permutations for the Apriori. Secondly, it generates all the possible

permutation and combination, then compare them to the `min_support`. Finally it outputs the list of frequent patterns by the variable `list_frequent`

```
count_dict = {}
for item in list_of_items:
    if np.sum(df[item]) >= min_support:
        count_dict[item] = np.sum(df[item])


# In[10]:


list_frequent = []
for i in range(2, len(count_dict) + 1):
    ls = list(itertools.combinations(count_dict.keys(), i))
    for per in ls:
        if (np.sum(np.sum(df[list(per)], axis = 1) == i) >= int(min_support)):
            list_frequent.append(per)
#             print(len(list_frequent), list_frequent, per)
```

Figure 1: Code for extract frequent patterns

The codes in figure 2, takes the list of frequent as the input and then generate the association rules with confidence higher the `confidence_rate`.

```
print("Procesing ...")
print("Result: ")
for fred in list_frequent:
#     print("For frequent list," fred, "we have rules:")
    sp2 = np.sum(np.sum(df[list(fred)], axis = 1) == len(list(fred)))
#     print("Support: ",sp2)
    ls = list(itertools.permutations(fred, len(fred)))
    for per in ls:
        passed = False
        sp1 = np.sum(np.sum(df[list(per)[:-1]], axis = 1) == len(list(list(per)[:-1])))
        if (sp2/sp1 >= confidence_rate):
            passed = True
        if(passed):
            print(" * ",list(per)[:-1], "=>",'"'+str(list(per)[-1])+'"',"with confidence:", sp2/sp1)
```

Figure 2: Code for extract the rules

IV. RUNNING:

a. 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.

A	B
TransactionII	Transaction
Trans1	A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans2	A Beginner's Guide, Java: The Complete Reference, Java For Dummies
Trans3	A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Trans4	Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition , Beginning Programming with Java,
Trans5	Android Programming: The Big Nerd Ranch, Beginning Programming with Java, Java 8 Pocket Guide
Trans6	A Beginner's Guide, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Trans7	A Beginner's Guide, Head First Java 2nd Edition , Beginning Programming with Java
Trans8	Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch,
Trans9	Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition , Beginning Programming with Java,
Trans10	Beginning Programming with Java, Java 8 Pocket Guide, C++ Programming in Easy Steps
Trans11	A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans12	A Beginner's Guide, Java: The Complete Reference, Java For Dummies, HTML and CSS: Design and Build Websites
Trans13	A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Java 8 Pocket Guide, HTML and CSS: Design and Build Websites
Trans14	Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Trans15	Java For Dummies, Android Programming: The Big Nerd Ranch
Trans16	A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans17	A Beginner's Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans18	Head First Java 2nd Edition , Beginning Programming with Java, Java 8 Pocket Guide
Trans19	Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition

Figure 3: Data Format – Amazon Data.

In order to run your own data, set up your transaction and put it into a `csv` file as figure 3 and remember the second columns contains the transaction. Then you put it into the `data` directory. And it's all done.

b. 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 4 and press `ENTER`.

```
Welcome to the simulation of the apriori algorithms.
The available data is listed bellow, input the choice you want
amazon.csv - to choose this data input (1)
kmart.csv - to choose this data input (2)
nike.csv - to choose this data input (3)
bestbuy.csv - to choose this data input (4)
custom.csv - to choose this data input (5)
```

Figure 4: 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.

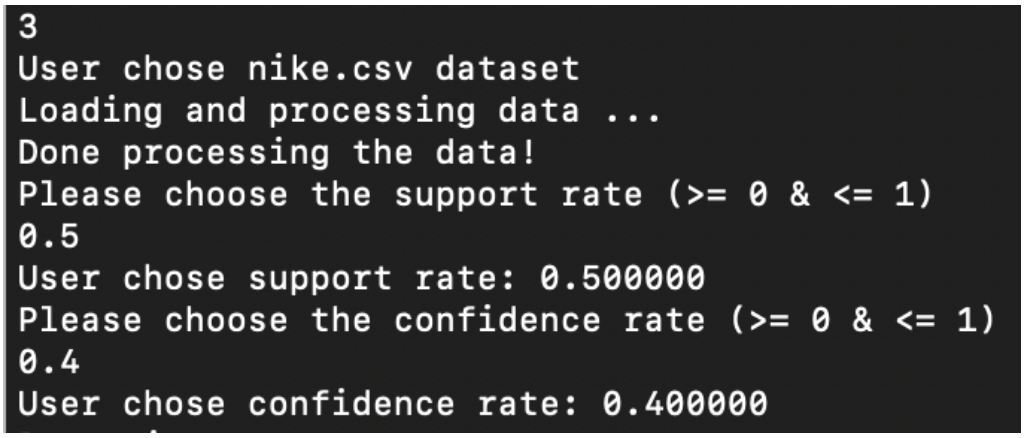

Figure 5: optional Support rate and confidence rate.

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

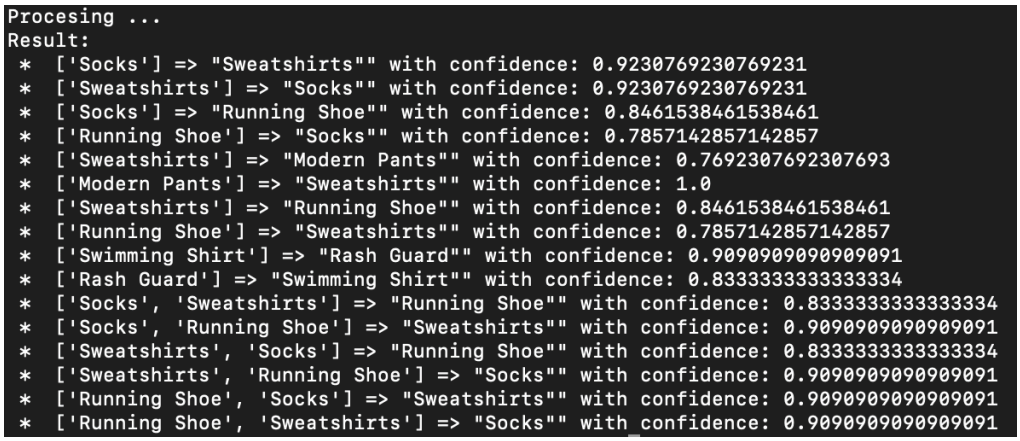

Figure 6: Result

V. RESULT:

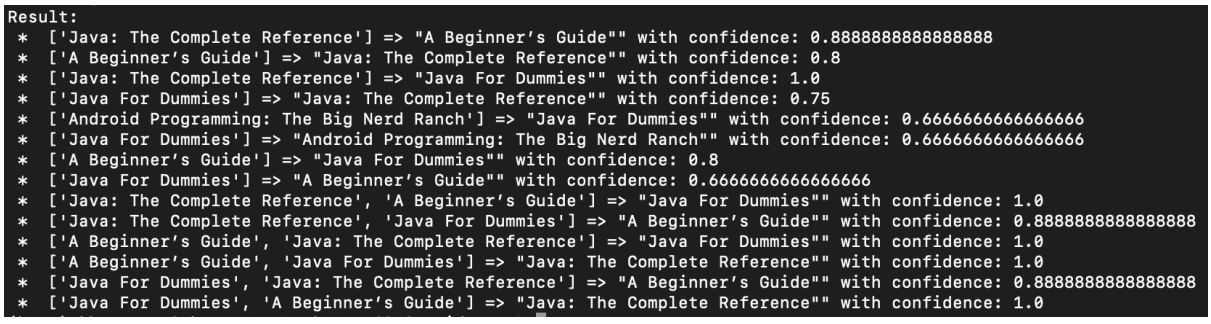

Figure 7: Result for amazon dataset, support 0.4, confidence 0.4

```
Result:
 *  ['Java: The Complete Reference'] => "Java For Dummies"" with confidence: 1.0
 *  ['Java For Dummies'] => "Java: The Complete Reference"" with confidence: 0.75
 *  ['Java: The Complete Reference'] => "A Beginner's Guide"" with confidence: 0.8888888888888888
 *  ['A Beginner's Guide'] => "Java: The Complete Reference"" with confidence: 0.8
 *  ['A Beginner's Guide'] => "Java For Dummies"" with confidence: 0.8
 *  ['Java: The Complete Reference', 'Java For Dummies'] => "A Beginner's Guide"" with confidence: 0.888888
 *  ['Java: The Complete Reference', 'A Beginner's Guide'] => "Java For Dummies"" with confidence: 1.0
 *  ['Java For Dummies', 'Java: The Complete Reference'] => "A Beginner's Guide"" with confidence: 0.888888
 *  ['Java For Dummies', 'A Beginner's Guide'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['A Beginner's Guide', 'Java: The Complete Reference'] => "Java For Dummies"" with confidence: 1.0
 *  ['A Beginner's Guide', 'Java For Dummies'] => "Java: The Complete Reference"" with confidence: 1.0
```

Figure 8: Result for amazon dataset, support 0.4, confidence 0.7

```
Result:
 *  ['A Beginner's Guide'] => "Java For Dummies"" with confidence: 0.8
 *  ['Java: The Complete Reference'] => "Java For Dummies"" with confidence: 1.0
 *  ['A Beginner's Guide'] => "Java: The Complete Reference"" with confidence: 0.8
 *  ['Java: The Complete Reference'] => "A Beginner's Guide"" with confidence: 0.8888888888888888
 *  ['Android Programming: The Big Nerd Ranch', 'A Beginner's Guide'] => "Java For Dummies"" with confidence: 0.8333333333333334
 *  ['A Beginner's Guide', 'Android Programming: The Big Nerd Ranch'] => "Java For Dummies"" with confidence: 0.8333333333333334
 *  ['Android Programming: The Big Nerd Ranch', 'Java: The Complete Reference'] => "Java For Dummies"" with confidence: 1.0
 *  ['Java: The Complete Reference', 'Android Programming: The Big Nerd Ranch'] => "Java For Dummies"" with confidence: 1.0
 *  ['Android Programming: The Big Nerd Ranch', 'A Beginner's Guide'] => "Java: The Complete Reference"" with confidence: 0.8333333333333334
 *  ['Android Programming: The Big Nerd Ranch', 'Java: The Complete Reference'] => "A Beginner's Guide"" with confidence: 1.0
 *  ['A Beginner's Guide', 'Android Programming: The Big Nerd Ranch'] => "Java: The Complete Reference"" with confidence: 0.8333333333333334
 *  ['Java: The Complete Reference', 'Android Programming: The Big Nerd Ranch'] => "A Beginner's Guide"" with confidence: 1.0
 *  ['Java For Dummies', 'A Beginner's Guide'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['Java For Dummies', 'Java: The Complete Reference'] => "A Beginner's Guide"" with confidence: 0.8888888888888888
 *  ['A Beginner's Guide', 'Java For Dummies'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['A Beginner's Guide', 'Java: The Complete Reference'] => "Java For Dummies"" with confidence: 1.0
 *  ['Java: The Complete Reference', 'Java For Dummies'] => "A Beginner's Guide"" with confidence: 0.8888888888888888
 *  ['Java: The Complete Reference', 'A Beginner's Guide'] => "Java For Dummies"" with confidence: 1.0
 *  ['Android Programming: The Big Nerd Ranch', 'Java For Dummies', 'A Beginner's Guide'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['Android Programming: The Big Nerd Ranch', 'Java For Dummies', 'Java: The Complete Reference'] => "A Beginner's Guide"" with confidence: 1.0
 *  ['Android Programming: The Big Nerd Ranch', 'A Beginner's Guide', 'Java For Dummies'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['Android Programming: The Big Nerd Ranch', 'A Beginner's Guide', 'Java: The Complete Reference'] => "Java For Dummies"" with confidence: 1.0
 *  ['Android Programming: The Big Nerd Ranch', 'Java: The Complete Reference', 'Java For Dummies'] => "A Beginner's Guide"" with confidence: 1.0
 *  ['Android Programming: The Big Nerd Ranch', 'Java: The Complete Reference', 'A Beginner's Guide'] => "Java For Dummies"" with confidence: 1.0
 *  ['Java For Dummies', 'Android Programming: The Big Nerd Ranch', 'A Beginner's Guide'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['Java For Dummies', 'Android Programming: The Big Nerd Ranch', 'Java: The Complete Reference'] => "A Beginner's Guide"" with confidence: 1.0
 *  ['Java For Dummies', 'A Beginner's Guide', 'Android Programming: The Big Nerd Ranch'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['Java For Dummies', 'Java: The Complete Reference', 'Android Programming: The Big Nerd Ranch'] => "A Beginner's Guide"" with confidence: 1.0
 *  ['A Beginner's Guide', 'Android Programming: The Big Nerd Ranch', 'Java For Dummies'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['A Beginner's Guide', 'Android Programming: The Big Nerd Ranch', 'Java: The Complete Reference'] => "Java For Dummies"" with confidence: 1.0
 *  ['A Beginner's Guide', 'Java For Dummies', 'Android Programming: The Big Nerd Ranch'] => "Java: The Complete Reference"" with confidence: 1.0
 *  ['A Beginner's Guide', 'Java: The Complete Reference', 'Android Programming: The Big Nerd Ranch'] => "Java For Dummies"" with confidence: 1.0
 *  ['Java: The Complete Reference', 'Android Programming: The Big Nerd Ranch', 'Java For Dummies'] => "A Beginner's Guide"" with confidence: 1.0
 *  ['Java: The Complete Reference', 'Android Programming: The Big Nerd Ranch', 'A Beginner's Guide'] => "Java For Dummies"" with confidence: 1.0
 *  ['Java: The Complete Reference', 'Java For Dummies', 'Android Programming: The Big Nerd Ranch'] => "A Beginner's Guide"" with confidence: 1.0
 *  ['Java: The Complete Reference', 'A Beginner's Guide', 'Android Programming: The Big Nerd Ranch'] => "Java For Dummies"" with confidence: 1.0
```

Figure 9: Result for amazon dataset, support 0.3, confidence 0.8

```
Result:
 *  ['Kids Bedding'] => "Bed Skirts"" with confidence: 0.8333333333333334
 *  ['Bed Skirts'] => "Kids Bedding"" with confidence: 0.9090909090909091
 *  ['Kids Bedding'] => "Shams"" with confidence: 0.75
 *  ['Shams'] => "Kids Bedding"" with confidence: 0.8181818181818182
 *  ['Kids Bedding'] => "Sheets"" with confidence: 0.8333333333333334
 *  ['Sheets'] => "Kids Bedding"" with confidence: 1.0
 *  ['Bed Skirts'] => "Shams"" with confidence: 0.8181818181818182
 *  ['Shams'] => "Bed Skirts"" with confidence: 0.8181818181818182
 *  ['Bed Skirts'] => "Sheets"" with confidence: 0.8181818181818182
 *  ['Sheets'] => "Bed Skirts"" with confidence: 0.9
 *  ['Kids Bedding', 'Bed Skirts'] => "Shams"" with confidence: 0.8
 *  ['Kids Bedding', 'Shams'] => "Bed Skirts"" with confidence: 0.8888888888888888
 *  ['Bed Skirts', 'Kids Bedding'] => "Shams"" with confidence: 0.8
 *  ['Bed Skirts', 'Shams'] => "Kids Bedding"" with confidence: 0.8888888888888888
 *  ['Shams', 'Kids Bedding'] => "Bed Skirts"" with confidence: 0.8888888888888888
 *  ['Shams', 'Bed Skirts'] => "Kids Bedding"" with confidence: 0.8888888888888888
 *  ['Kids Bedding', 'Bed Skirts'] => "Sheets"" with confidence: 0.9
 *  ['Kids Bedding', 'Sheets'] => "Bed Skirts"" with confidence: 0.9
 *  ['Bed Skirts', 'Kids Bedding'] => "Sheets"" with confidence: 0.9
 *  ['Bed Skirts', 'Sheets'] => "Kids Bedding"" with confidence: 1.0
 *  ['Sheets', 'Kids Bedding'] => "Bed Skirts"" with confidence: 0.9
 *  ['Sheets', 'Bed Skirts'] => "Kids Bedding"" with confidence: 1.0
```

Figure 10: Results for kmart dataset, support 0.4, confidence 0.4

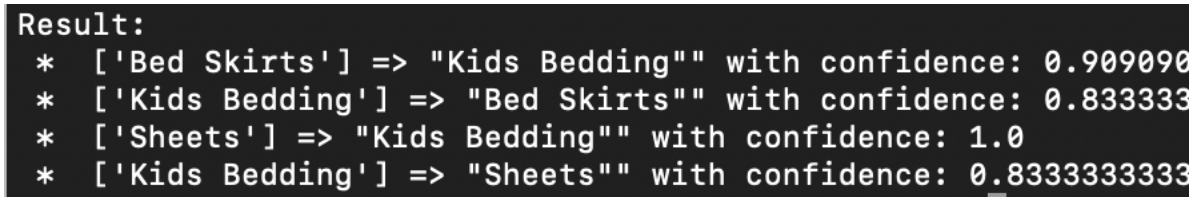

Figure 11: Results for kmart dataset, support 0.5, confidence 0.5

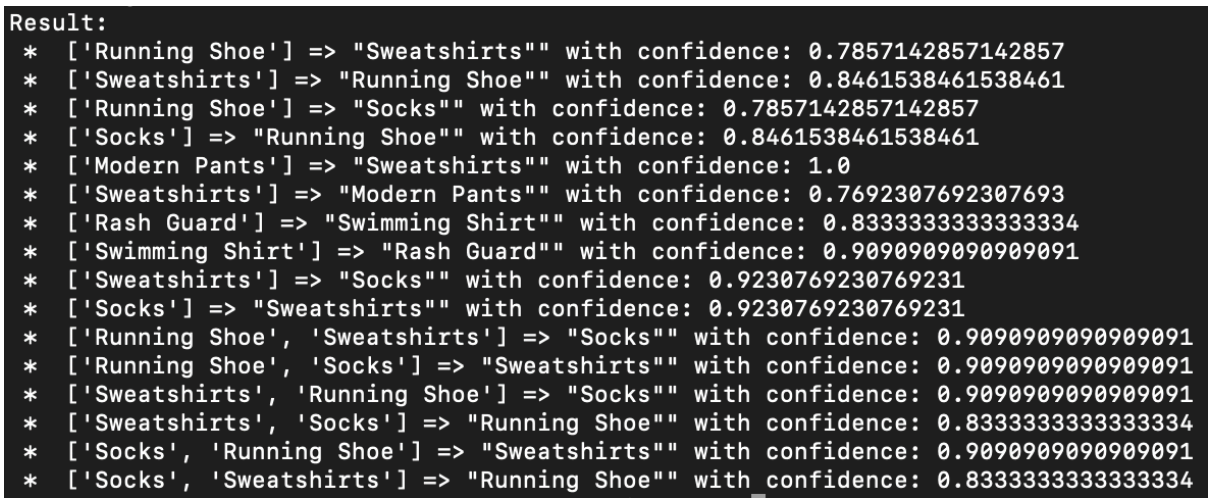

Figure 12: Results for nike dataset, support 0.5, confidence 0.6

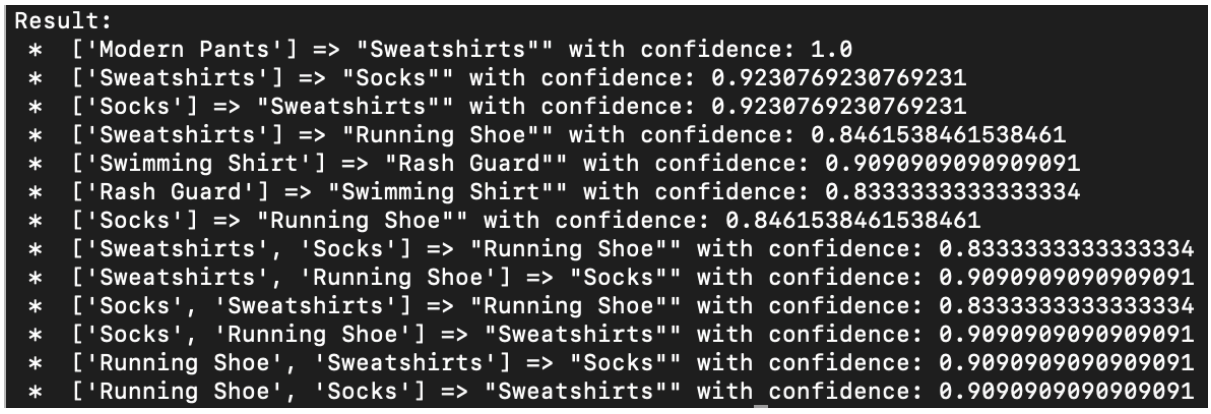

Figure 13: Results for nike dataset, support 0.5, confidence 0.8