Data Mining, CSCI 347, Fall 2019
Homework 3, Jupyter Notebook and Association Rules, due Oct. 7

Explore association rules using Python's MLxtend library. MLxtend contains the apriori algorithm which extracts frequent itemsets and the association_rules algorithm which builds the rules from the itemsets.

The steps in this assignment come from "Introduction to market Basket Analysis in Python" by Chris Moffitt
https://pbpython.com/market-basket-analysis.html

Working with the OnlineRetail.xlsx dataset from the Machine Learning Repository at University of California, Irvine (UCI) https://archive.ics.uci.edu/ml/datasets.php

The dataset contains transactions occurring between Jan. 12, 2010 and Nov. 12, 2011 for a UK-based online retail store.

Records: 541,909
Attributes: 8
InvoiceNo - nominal, 6-digit number uniquely assigned to each transaction. If code starts with letter 'c', it indicates a cancellation.
StockCode – nominal, 5-digits, unique to each product
Description – nominal
Quantity – numeric (integer)
InvoiceData – numeric (date in the format dd/mm/yyyy hh:mm)
UnitPrice – numeric (decimal with 2 decimal points for cost)
CustomerID – nominal, 5-digits unique to each customer
Country – nominal, the country where the customer resides {United Kingdom, France, Australia

Do the following.

1. Install pandas and MLxtend libraries onto your machine.
   Can use MSPowerShell
   > pip install pandas // Data analysis library that also includes data structures
   // such as Python's dataframes.
   > pip install xlxtend // Library with aprioir module

   (If pip "Python Installed Package" is not yet installed, get it first.)

2. Import the pandas library, and let 'pd' refer to it. Import the modules aprioir and association_rules from the mlxtend library.

```
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
```

3. Use the pandas function read_excel(filename) to put the file into a Python data frame. The following is a good reference for working with data frames: https://www.geeksforgeeks.org/python-pandas-dataframe/.

```
df = pd.read_excel('OnlineRetail.xlsx')
```

4. Try the following.

```
print(df)       // Display the beginning of the dataframe
len(df)         // Display the number of records in the dataframe
df.head()       // Display the head of the dataframe
```

5. Translate the values in the 'InvoiceNo' attribute to strings and remove the cancelled transactions (those transactions containing 'C'). Use

```
df['InvoiceNo'] = df['InvoiceNo'].astype('str')
df = df[~df['InvoiceNo'].str.contains('C')]
```

6. Consolidate the items into 1 transaction per row with each product 1 hot encoded. Limit this to just the sales in France.

```
basket = (df[df['Country'] =="France"]
          .groupby(['InvoiceNo', 'Description'])['Quantity']
          .sum().unstack().reset_index().fillna(0)
          .set_index('InvoiceNo'))
```

7. For 1-hot encoding, only want 0s and 1s in the fields. Make all positive values 1 and others 0.

```
def encode_units(x):
    if x <= 0:
        return 0
    if x >= 1:
        return 1

basket_sets = basket.applymap(encode_units)
basket_sets.drop('POSTAGE', inplace=True, axis=1)
```

8. Generate frequent item sets that have a support of at least 7%. That is, at least 7% of the transactions in the dataset contain this itemset.

```
frequent_itemsets = apriori(basket_sets, min_support=0.07,
                                                use_colnames=True)
```

9. Generate the rules with their corresponding support, confidence and lift.

```
rules = association_rules(frequent_itemsets, metric="lift",
min_threshold=1)
rules.head()
```

10. There are quite a few rules with a high lift value, which means that the itemset occurs more frequently than would be expected. There are also cases where the confidence is high. Filter the dataframe using standard pandas code. Look for a lift value of 6 or more, and a confidence of 80%.

```
rules[ (rules['lift'] >= 6) &  (rules['confidence'] >= 0.8) ]
```

11. It appears that green and red alarm clocks are purchased together and the red paper cups, napkins and plates are purchased together in a manner that is higher than the overall probability would suggest.

    To see how many green alarm clocks are sold

```
basket['ALARM CLOCK BAKELIKE GREEN'].sum()
```

    and how many red alarm clocks are sold

```
basket['ALARM CLOCK BAKELIKE RED'].sum()
```

12. There are 38 countries represented in this dataset. Use the apriori and association_rules algorithms to find rules that apply in other countries. Explore rules for these other countries in an attempt to find something interesting.

13. Place comments into the Jupyter Notebook file to describe your exploration and the results. Show something interesting, and describe it in the comments.

Email the Jupyter Notebook file to me.