Assigned: 03/30/2021
Due: Sat. 04/24/2021, 11:59pm

Instructions: This project will cover some questions related to association analysis and recommender systems.
This assignment is split into two main parts.

Part A:
You will be asked to walk through methods **manually**, that is not using the available functions to perform the analysis analysis. You can use R, Matlab, or Python for plotting of the data when asked and for support in calculations (compute distances, averages, etc.). But, the process should be done manually with your answer presented in **typed form**.

You may then prepare your solution to Part A as a separate document using Google Docs, Word, LaTeX, with computer generated plots inserted. Note, if you prepare your solutions in Rmd, Sweave, or Python notebooks, your solution to Part A can also be included in that document using Markdown or LaTeX to enter your answers (not a separate document).

Part B:
You will again have a final document submission that includes text responses to questions, tables, R/Matlab/Python code used to calculate answers, and figures.

Formatting of submissions will follow the same approaches as those used in prior assignments. Namely: Matlab + markup, Matlab LiveScript, Rmd, Sweave, Jupyter notebook, Colab notebook. Any other packages or tools, outside those listed in the assignments, examples, or those for reproducible research should be cleared by Dr. Brown before use in your submission.

For this project you are allowed to work in **groups of up to 3**. All students must sign-up on Canvas into a group (e.g., P6-Rho, P6-Tau, etc.).

Name your main submission files as *P6_GroupName*, create a zip-file called *Project6_GroupName.zip* and submit on Canvas. For example, if I was using R, and member of GroupChi, I would submit either:

- *P6_GroupChi.Rmd*, *P6_GroupChi.pdf*, or
- *P6_GroupChi.Snw*, *P6_GroupChi.pdf*

Questions:

PART A - WRITTEN QUESTIONS

1. (1 point) List your project GroupName and all group members names.

2. Association Analysis I
   Given a database of transactions and a min-support = 2,

Trans.	Items
T1	I1, I2, I3, I4, I5, I6, I7, I8, I9, I10
T2	I1, I2, I3, I4, I5, I6, I7, I8
T3	I1, I2, I3, I4, I5
T4	I6, I7, I8
T5	I100, I101, I102, I103

(a) (3 points) Find an example of an association rule that matches the following pattern with min-support = 2 and min-conf = 70

$$(I1, I2, I3, I4, IX \rightarrow IY)$$

(b) (8 points) For the association rule $I1 -> I6$, compute the support, confidence, lift, and conviction

TID	Items
T100	M, O, N, K, E, Y
T200	D, O, N, K, E, Y
T300	M, A, K, E
T400	M, U, C, K, Y
T500	C, O, O, K, I, E

Table 1: Transaction Data

3. (24 points) Association Analysis II
Consider the data listed in Table 1. Show the major steps through the Apriori algorithm. Present L_i and C_i for each level i considered. Also, report at the end the frequent itemsets identified.

P6 : PART B - PROGRAMMMING QUESTIONS

4. (12 points) Confirm the results above of the Apriori algorithm. For R, the `arules` package is available. Matlab has the Association Rules package available from File Exchange[1]. Python has the 'mlxtend' library.

5. (16 points (bonus)) You will analyze a portion of the Instacart Online Grocery Shopping Dataset 2017[2]. The 2 data sets you are given contains just 20K or 500K items purchased, while the original data set has 3 million orders.

   You will only need to focuse on the following files: 'order_products__train_small.csv', 'order_products__train_med.csv', and 'products.csv' for this analysis. You can link the product number in the "order_products" file to the name of the product in the "products.csv" file.

   (a) Create a histogram showing the number of products per order for both the 'order_products__train_small.csv' and 'order_products__train_med.csv' data sets. Indicate with a vertical line where the mean number of products per order lands.

   (b) For the 'order_products__train_small.csv' data, create an top 15 item frequency plot, that is plot the top 15 most frequently purchased items. This should be a bar plot with items vs. frequency (relative support).

   (c) For the 'order_products__train_small.csv' data, use Apriori to find association rules with a minimum support of 0.003 and confidence of 0.5. Report in a table the top 10 rules (sorted by lift) with the product names, the support, confidence and lift.

   (d) Rerun Apriori on the same data set with a minimum support of 0.0025 and confidence of 0.5. Create a scatterplot of the rules, plotting support vs. confidence colored by lift value.

[1] http://www.mathworks.com/matlabcentral/fileexchange/42541-association-rules
[2] The full data set is available here: https://www.instacart.com/datasets/grocery-shopping-2017