# Assignment 3

**CMPT 459 Spring 2021**
**Data Mining**

**Martin Ester**
**TAs: Arash Khoeini** and **Madana Krishnan Vadakandara Krishnan**

**Deadline: 23:59 pm PST on April 12th**.

[Total Marks:100]

The aim of this assignment is to implement a post-processing method to determine all closed and maximal frequent itemsets, given that all frequent itemsets have been obtained by running the Apriori algorithm. See slides 399-401 and 403-405 of the lecture notes.

## BMSWebView2 Dataset

The dataset that you are provided with contains 77512 transactions of click-stream data from an e-commerce store. The dataset includes 3340 unique items in total.

## Tasks

a. Closed Itemsets [35 marks]
Assume that you already have all frequent itemsets in a dictionary, where the value for key i is a dictionary itself, containing a tuple of itemsets of size i as keys and its support as value. Here you can see an example of this data structure:

{1: {('apple',): 3, ('orange',): 3 },
2: {('apple','orange'): 2 } }

Implement a method that takes these itemsets and returns a list of all **closed itemsets**. The return type should be a dictionary, with the same structure as the input dictionary (like the example above).

b. Maximal Itemsets [35 marks]
Again assume that you already have the frequent itemsets in a dictionary of the above format. Implement a method that takes these itemsets and returns a

list of all **maximal itemsets**. The return type should be a dictionary, with the same structure as the input dictionary.

c. Statistics [20 marks]
Use the following implementation of the Apriori algorithm: https://pypi.org/project/efficient-apriori/
to get the set of frequent itemsets from the BMSWebView2 dataset. Run your methods for part a and part b on the result of the *apriori* implementation from that package, which is a dictionary of the format explained in part a. Use the following thresholds: min_support= 0.005 min_confidence= 0.7
Provide the numbers of all, of the closed, and of the maximal frequent itemsets.

d. Insight [10 marks]
Compare the numbers of the different types of frequent itemsets and discuss them. This is a dataset from an e-commerce website, where each transaction contains all the items viewed by a user in a session. What do you think would be the use of finding frequent itemsets in this setting?

**Hints:**

- You should remove all -1s and -2s from the dataset.

**[IMPORTANT]** You should submit two files:

- A report file: *[studentID].pdf*
- A Python file: *[studentID].py*

**Deadline: 23:59 pm PST on April 12th**.
You will lose 10% of the marks for submissions after this deadline, as long as it's not more than 24 hours late. You will lose all the marks for submissions after that.

**Libraries:** You can use libraries including math, numpy, random, etc.

You MUST provide YOUR OWN code for finding closed and maximal itemsets and for all the tasks specified in this assignment. These MUST be implemented from scratch i.e. not using scikit-learn or other libraries. You will be marked on the correctness of your implementation.