**Project 3.B**

1. Project Description

In this project, you will implement a simplified version of the Apriori Algorithm to mine frequent itemsets from the LitCovid corpus. Mining frequent itemsets can gain insights to what topics are commonly covered in the corpus.

Specifically, you are asked to write scripts in Jupyter Notebooks (python 3, template is provided) for the following tasks.

1. **Query Data [Points: 10]**

This project is to mine frequent itemsets from the passage.text field of the LitCovid corpus. Each passage.text corresponds to one paragraph of the paper. You will first connect to LitCovid database you set up for Project3A, and query the database to obtain all 'passages.text'.

1. **Data preprocessing**

Code is provided in template. This step is to prepare transactions for the pattern mining step.

1. **Patten mining [Points: 90]**

You will implement Apriori algorithm to mine all frequent 1-itemsets and frequent 2-itemsets from transactions. You will finish function “def frequentItem(transactions, minSupport=200, k):” in the template.

The input to the function includes: list of transactions, minimum support threshold, and the itemset length k (1 or 2).

The output of the function is a sorted list (sort based on frequency) of all frequent k-itemsets.

The expected result of frequent 1-itemsets is

[('covid-19', 9052),

('patients', 4704),

('study', 3213),

…]

The expected result of frequent 2-itemsets is

[('covid-19,patients', 2837),

('covid-19,pandemic', 1921),

('covid-19,disease', 1661),

…]

**Submission Instruction**

*Rename project 3.B template.ipynb to project 3.B.ipynb. Submit this file to Gradescope.*