A close up of a logo

Description automatically generated

Faculty of Science

**Course**: CSCI 4030U Big Data Analytics

**Lab:** #1-5

**Topic:** Mining Frequent Itemsets

**Description**

Your task it to conduct market-basket analysis by developing and comparing experimentally various frequent itemset algorithms, including, Apriori, PCY, sampling and SON. The goal is to find frequent pairs and triples of elements and analyze the results.

**Resources**

See Lecture 2 and Chapter 6 from MMS book posted on Blackboard (Association Rules Data Mining).

**Programming Language**

You can choose your favorite programming language (e.g., C++, Java, C#, Go and Python)

**Tasks and Schedule**

1. Setting up environment, implementation of **Apriori** algorithm for finding frequent pairs of elements (Lab 1).
2. Finishing implementation of Apriori algorithm and conducting scalability experiments on given datasets (Lab 2). Starting implementation of **PCY algorithm**.
3. Finishing implementation of **PCY Algorithm** for finding frequent pairs of elements. Compare experimentally performance of PCY vs Apriori (Lab 3).
4. Extend your implementation to the **Random Sampling** and **SON** version of the algorithm (Lab 4). What is the efficiency of the method in comparison to Algorithms developed in Points 1-3? Quantify the number of false positives.
5. Implement **Multistage or Multihash version** of the PCY algorithm. Polish the implementation and write the final report (Lab 5).

**Dataset**

1. The retail dataset contains the (anonymized) retail market basket data (~88200 baskets) from an anonymous Belgian retail store. The data are provided ’as is’.

Note that since the dataset was anonymized the preprocessing step to map text labels into integers is done for you. (Working with integers is more efficient than textual data as it is saves the main memory.)

Retails dataset is available at:

http://fimi.uantwerpen.be/data/retail.dat

Use Notepad++ or other software rather than Notepad to open the file for the correct formatting.

1. Netflix dataset

Conduct experiments over the Netflix dataset.

Test the limitations of your implementations. (The number of tuples that you can handle.) Netflix data set is available at:

<https://drive.google.com/open?id=1EX_2Pkid6EC4H-4KN0kP_S_89GKaTnXo>

**Experiments**

Perform the scalability study for finding frequent pairs of elements by dividing the data into the chunks: 20%, 40%, …, 100% and measuring the time performance. Provide the diagrams.

Use various thresholds: 1%, 2% and 5%.

Quantify the number of false positives for the sampling methods.

**Report**

After Lab 5 you will have to submit report which will be marked towards your project. Attach separately your code.

**Competition**

The most efficient implementation will be determined and given an extra 1 bonus point.