

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/comparing various efficient frequent itemsets algorithms, including, Apriori and PCY. The goal is to find frequent pairs and triples of elements. (You can stop there.)

**Resources**

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

**Programming Language**

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

**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 the given dataset (Lab 2).~~
3. ~~Implementation of~~ **~~PCY Algorithm~~** ~~for finding frequent pairs of elements (Lab 3).~~
4. ~~Finishing implementation of PCY and conducting scalability experiments (Lab 4).~~
5. Polishing and writing the report for the project (Lab 5).

**Dataset**

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.)

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

**Dataset link:**

http://fimi.ua.ac.be/data/retail.dat

**Experiments**

Perform the scalability study for finding frequent pairs of elements by dividing the data into the chunks and measuring the time performance. Provide the figure.

Play with different threshold s, e.g. 1% etc.

**Extensions**

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

**Optional (Bonus Points)**

Implement Multistage (3 Passes) or Multihash version of PCY or Sampling Based or SON version of mining frequent itemset algorithm.

**Report**

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