CMPSC 497: Frequent Itemset Mining

In this assignment, we will look for relationships between movies that customers viewed. The file, freqMovies.txt, contains a line with movie id’s that have been watched by a customer. The movies.txt file has a mapping of these id’s to movie titles. Your task is to generate a list of the most interesting association rules from this dataset.

Here are your tasks:

a. The dataset has already been sampled in that we include only the customers that have viewed a large number of movies. You will further sample the dataset by using every 20th customer (i.e. use the “baskets” on first line of the freqMovies.txt file, the 21st line, the 41st line, etc.)

b. Using a minimum support threshold of 0.3, find the following quantities:
   a. The number of frequent itemsets of each size
   b. The number of maximal frequent itemsets of each size

c. Using a minimum confidence level of 0.85, find all rules which have 1 movie as an antecedent and 1 movie as a consequent.
   a. How many rules did you find?
   b. For each rule, give the association rule and the interest value. Use the movies.txt file to list the titles, rather than the movie id’s, in the rule.

d. Which rule is the most interesting and which rule is the least interesting?

Deliverables

A report (as a pdf, doc, or docx file) with your answers to parts *b*, *c*, and *d,* and a py file with your code.