Homework 7

Overview

In this assignment, you will practice calculating support and confidence, and you will use R to apply association rule mining to a "bag of words" dataset.

For this assignment, you will use the provided template file (provided on blackboard) for your homework assignment.

Objectives

- Calculate support of itemsets given transaction database
- Calculate support of association rules given transaction database
- Calculate confidence of association rules given transaction database
- Use the Apriori algorithm to find frequent itemsets with R
- Use the Apriori algorithm to find association rules with R

Grading

- Uploaded requested files: 5%
- File is properly formatted, using the template provided on blackboard: 5%
- Part A (all questions weighted evenly): 45%
- Part B (all questions weighted evenly): 45%

Deliverables

- .pdf file (generated by knittinng your .Rmd file)
- .Rmd file (used to generate your .pdf file)

Part A

The table below has 10 transactions (1 per row). Each transaction may contain any of the following items: A, B, C, and D. A transaction includes an item if there is a 1 under that item, and a transaction does not have a particular item if it has a 0 under that item. For example, transaction 3 (id = 3) includes the items A, C, and D, and transaction 1 includes items A, B, and C.

id	A	B	C	D
1	1	1	1	0
2	1	1	0	1
3	1	0	1	1
4	0	0	1	0
5	1	0	0	1
6	0	0	1	0
7	1	1	1	0
8	1	0	0	1
9	0	1	1	1
10	1	0	1	1

Calculate the support for the following itemsets (you may calculate these by hand):

1. {A, B}
2. {B, C}
3. {A, B, C}
4. {A, B, D}

Calculate the support and confidence for the following rules:

5. {A} –> {B}
6. {B} –> {A}

Part B

The UCI Machine Learning website has a large collection of data sets that are appropriate for practicing and testing data mining techniques. The files at https://archive.ics.uci.edu/ml/datasets/bag+of+words are sets of text that are well-suited for learning association rules. I have converted one of these datasets into a transaction format for you (`words_kos.dat`), which you can download from blackboard. If you are curious about how I converted the bag of words dataset into a transactions format, you can find the script I used here.

Each row of `words_kos.dat` is the *set* of words found in one blog post from the Daily Kos website. The Daily Kos is a blog that covers politics in the US. The data for this assignment was scraped from the blog sometime prior to the 2004 presidential election. It represents 3430 postings. Each post is an article or opinion piece. The data collected is based on 6906 words, selected based on their usage. Words like "a", "the", and "be" are very common and no so helpful for analysis, so they were excluded from the dataset. The 6906 words that were chosen are used frequently in the blog posts but not too frequently. Find more information about this dataset here. Find more information about the 2004 US presidential election here.

You are to download `words_kos.dat`, load the transactions into R (using the `arules` package), and do questions 7 through 10 (below). For these questions it may be helpful to reference the documentation for the `arules` package.

7. Use the Apriori algorithm (e.g., from the `arules` package) to identify all frequent itemsets with a minimum support of 0.3 (i.e., 30%). Use the `inspect` function to print out the frequent itemsets that you identify.
8. Which itemset has the greatest support out of all the frequent itemsets that you identified in question 7?
9. Use the Apriori algorithm (e.g., from the `arules` package) to find association rules. You may need to adjust the minimum support and confidence thresholds in order to get a reasonable number of rules (Hint: try starting your minimum support threshold between 10% and 20%).
10. Write three interesting rules (in your opinion), and explain why each of the rules you chose is interesting.