COMP 4710 – Assignment 2

Association Mining

Total Marks: [20]

Due Date: **Wednesday, October 29th** at 11:59 pm

In this assignment you are provided with data that consists of segments of text from scientific papers. The objective of this assignment is to discover relationships and rules associated with words appearing in the segments of text (frequent itemsets and association rules).

The data you will be processing is within the file medical.arff.

The arff (attribute-relation file format) is a specific file format, and the arff python package may be a useful import to make reading this easier.

Specific tasks and marks associated with each task follow:

1. Read the data file [3]

   Read the data file (medical.arff) into a suitable data structure. load_dataset () within the apriorialg.py file provided should give you a sense of what the expected format looks like. Each word should be an item, and each sentence a transaction.

2. Generate frequent item sets [1]

   Using the functionality provided by the apriorialg.py file, generate frequent item sets associated with the data.

3. Generate association rules [1]

   Using the functionality provided by the apriorialg.py file, generate an initial set of association rules.

4. Examine the rules that are produced, and comment on them [2]

   Discuss the nature of the rules that you observe (1-paragraph). What effect do minimum support and confidence have on the rules that are produced?

5. Find “interesting” rules [7]

   In class, we discussed the need to determine which rules are most interesting. You should implement the following methods to filter out rules that are less interesting:

i. Lift, ii. Interest, iii. PS, iv. Phi (you may refer to the notes)
You should also print out the rules in a format that is nice to read. (1 Mark is dedicated to the formatting)

6. Comment on the "interesting" rules [2]

   Discuss (1 paragraph) how things change when you retain only those rules that are interesting (e.g. compared with the output from step 4.)

7. Single threshold with data filtering [4]

   Recall that in class we discussed that a benefit of multiple thresholds is to allow for less common associations to be discovered even when there are many items that occur with a very high frequency. An alternative to this is to use a single threshold, and filter out words that are very frequent.

   Filter the data by removing frequent words. Do this until you have rules that are less related to language constructs, and more related to actual concepts. Give 5 examples of rules generated that you consider to be especially interesting.

To hand in:

Your **myAssociationMining.py** file.

This should include your implementation for questions 1,2,3,5,7.

You should also hand in a pdf file named **yourlastname_yourfirstname.pdf**.

The pdf file should contain:

i. Your 1 paragraph response to question 4.
ii. Your 1 paragraph response to question 6.
iii. Your examples from question 7. These should use the format produced in question 5.

Tips:

You'll have to experiment with different thresholds for support, confidence and measures of what is interesting. Your intuition should help to guide you in making this choices.