CS412 Frequent Patterns Report

Name: Huajie Shao NID:hshao5

1. Briefly explain what algorithms you use in Step4~Step6.
   In step 4, I use the Apriori Algorithm to do frequent mining. In step 5 and Step 6, I write the codes to calculate the closed and max patterns based on the output of step 4.

2. Question to ponder A: How you choose min_sup for this task? Note that we prefer min_sup to be the consistent percentage (e.g. 0.05 / 5%) w.r.t. number of lines in topic files to cope with various-length topic files. Explain how you choose the min_sup in your report. Any reasonable choice will be fine.
   The min_sup is defined as 0.01 w.r.t number of lines of each topic. I choose the min_sup based on the frequency of patterns and the corresponding computational complexity.

3. Question to ponder B: Can you figure out which topic corresponds to which domain based on patterns you mine? Write your observations in the report.
   - Topic 0: Information retrieval.
   - Topic 1: Data Mining
   - Topic 2: Machine learning
   - Topic 3: Database
   - Topic 4: Theory

4. Question to ponder C: Compare the result of frequent patterns, maximal patterns and closed patterns, is the result satisfying? Write down your analysis.
   In general, the results are satisfying.
   - For Topic-0.txt, we can see from the corresponding max pattern and closed pattern files that the support of “information retrieval” is 324. So it is easy for us to judge Topic 0 is related to information retrieval.
   - For Topic-1.txt, we can see from the pattern-1.txt.phrase and closed-1.txt.phrase and max-1.txt.phrase files that the support of the frequent pattern “mining” is 983 and also the support of the phrase of “mining pattern” and “mining association” is 201 and 158. So we can predict that Topic-1 is related to data mining.
   - For Topic-2.txt, we can see from pattern-2.txt.phrase and closed-2.txt.phrase files that the support of “learning” is 2046 and also the support of the phrase of “support vector machine” and “machine learning” is 123 and 115, respectively. So we can predict that Topic-2 is related to machine learning.
   - For Topic-3.txt, we can see from pattern-3.txt.phrase and

closed-3.txt.phrase files that the support of frequent pattern “database” is 1497 and also the support of the phrase of “database system” is 424. So we can predict that Topic-3 is related to database.

- For Topic-4.txt, we can see from pattern-4.txt.phrase and closed-4.txt.phrase files that the support of “algorithm” is 962 and also the support of the phrase of “algorithm approximation” is 113. So we can predict that Topic-4 is related to theory.

5. List your source file names and their corresponding Steps.
   The following is the source files for the corresponding steps to deal with the data:

Source files	Steps
Preprocess_paper.py	Step 2: Preprocessing
Reorganize_topic.py	Step 3: Partitioning(2)
FP_Apriori.py	Step 4: Mining Frequent Patterns
Max_closed.py	Step 5: Mining Maximal/Closed Patterns
Purity.py	Step 6: Re-rank by Purity of Patterns