CS412: Introduction to Data Mining, Fall 2017, Homework 3

Name: Nestor Alejandro Bermudez Sarmiento (nab6)

Worked individually

Assignment 3 deals with concepts of frequent itemsets, outliers and minimum occurrence windows.

Question 1

For the itemset **{D, E, F}**, report the following statistics.

Length of minimum occurrence window.

For S1: 4
For S2: 4
For S3: 4

Number of outliers in the outlier based minimum occurrence window.

For S1: 1
For S2: 0
For S3: 1

Question 2

The code can be found in the zip file accompanying this report. A few things to note:

1. I used Python 3.6 but I tested the code with Python 2.7 and it works but for some reason if you use Python 3.6 the itemsets are well ordered by size, which doesn't happen with Python 2.7. So, if possible, try using Python 3.6 for better readability.
2. The Python script expects the **data.txt** file to be located at the same level as the main script.
3. The name of the input file is hard-coded to **data.txt**.
4. The name of the output file is hard-coded to **nab6-HW3.txt**

My code implementation makes use of the Apriori algorithm to mine the frequent itemsets with some modifications to handle the outlier based minimum occurrence window.
I first initialize the ranges in which the 1-itemsets appear on each sequence. Note that the ranges have the same start and end positions because they are 1-itemsets. For example, if

{A} appears in the position 2 and 5 in S1 then its ranges will be **(2, 2)** and **(5, 5)**. This redundancy is important to have the same treatment for all itemsets.

Now, when the candidates for the $(k+1)$-itemsets are being generated I take the ranges for the two k-itemsets that are being combined and create new ranges by taking the minimum value between the start position of both itemsets as the start position of the new range and the maximum value between the end position of both itemsets as the end position of the new range.
For example, if **{A, B}** appears in **(3, 5)** and **{B, C}** appears in **(5, 6)** the range of **{A, B, C}** will be **(3, 6)**. If the itemsets have more than one occurrence then all of them are considered.

Later on, the ranges are used to take a portion of the corresponding sequence and find the outliers given an itemset as a reference.

Results for the sample data.txt file

```
1
2
4
5
6
1, 2
1, 4
2, 4
4, 5
5, 6
1, 2, 4
1, 4, 5
4, 5, 6
1, 2, 4, 5
```