Report

The threshold I chose was 262. Since the total number of transactions is 26205, and I would like to find key words item sets that are relatively frequent among these transactions, I thought 1% of all transactions would be a reasonable amount, which is around 262 for the support. Anything under 262 were considered as not frequent enough.

I used optimization in the candidate generation process that I learned in class. When I tried to generate k+1 item sets from k item sets, I sort all the item sets, and then pair each itemset with the rest, if their first k-1 items are the same then merge them to generate the new candidate. Since the generation process is the most time-consuming part of the algorithm, this technique can avoid repeatedly generate same item sets and also make sure at least two subsets of this new item set are frequent, increasing the probability that it is also frequent.

The lessons I learned include that it makes the code more readable and clearer structured if I separate each step of the algorithm into independent functions. Also, data mining tasks often have several coding ways to solve, but it is better for us to find a way that is most time efficient since we need to deal with large dataset.

I used 262 as the minimum support value and finally got 471 item sets that have at least 262 supports. From the result, I found that “flu” and “shot” are the most frequent two words with 26186 and 23167 supports respectively, and their combination is the third most frequent, which match with the topic of these tweets. They are followed by item sets that include either “get”, “getting” or “got” combined with “flu” or “shot”, showing a potential trend of tweets talking about getting flu shot. Item sets with “today” and “arm” are also quite frequent in the list. The most frequent item sets are mostly one- or two-words item set. While with the decrease of support, there appears to be an increase of three- of four-words item sets. There is hardly item sets over 4 words with this minimum support.