# Mini Project #1 - Association Pattern_based Family History Identification

**Jie Dong (912837580)**

Please note that in this report, frequent word associations are frequent item sets, and closed word associations are closed item sets.

## A. Source code

I used Python to develop this project.

## B. A brief description of main steps

I modified the sentence splitter posted on iLearn to save the report name for each family history sentence. After applying Apriori, I filter out frequent word associations if they appear in less than 5 reports. I followed the subtasks listed in project description.

The main thing I did different is utilizing closed word associations. This helps to minimize search space by nearly 1/3. The corresponding steps are:

1. From the frequent word associations, create a dictionary to map closed word associations to their corresponding frequent word associations (see script connect_closed_with_frequent_word_associations.py). I understand the closed word associations that I found are not strictly closed; However, they help to reduce search space of 3481 frequent word associations to 1911 closed word associations (see primitive_word_associations.txt and primitive_closed_word_associations.txt).
2. Extract all the sentences containing the closed word associations and store them in a file (see script extract_sentences_for_each_word_association.py and closed_word_association_with_sentences.txt). This helps to centralize the search space. For further analysis, like applying word span and word sequence, the processing are done in two steps:

First, for each closed word association, fetch its corresponding frequent word associations from the dictionary.

Second, for each frequent word association, apply the processing technique required on the containing sentences. There is no need to open other files to find containing sentences any more.

Other less important steps I did:
Utilize the property that frequent word associations that meet lower word span requirement is a proper subset of the word associations that meet higher word span requirement. E.g. Set(frequent word associations for k = 3) is a proper subset of Set(frequent word associations for k = 5). This helps to reduce search space greatly too.
Store diseases and family members as text files. In my program, I load these files to get the information. This helps to make changes in the future easier.
Keep the frequency of each frequent word associations in all steps. This helps to analyze the results.

**C.Instructions on compiling and running my program**

1. Unzip the file.
2. Under the shell, run the following command:

cd <the newly created directory>
./command

**D. Discussion**

1. Quality of word association patterns

Word association patterns are good for discovering frequent occurring patterns. However, they are not perfect at finding meaningful patterns in the natural language field. In text analytics of natural languages, words have syntactic and semantic meanings. Things like the distance between words and the sequence of appearing oder matter a lot. Just breaking up the sentences into words, and mining frequent patterns are not good enough.
Based on the results of applying an association mining algorithm, below are my observations:
First, meaningless patterns appear often, e.g. the support of (she, her) is 70. Clearly, this pattern offers very little insight for the problem. I think word association mining algorithms should be used with consideration of the targeting area. For example, with some analysis of the field, incorporate more stop words to help eliminate meaningless word associations.
Second, combine word association patterns with other techniques to incorporate more syntactic information of the sentence. The postprocessing that we did for this project, filter word association patterns by its appearance in different reports, apply word span and sequence are very necessary.
Overall, word association patterns should be used with other techniques for this problem.

2. Impact of the parameter k

k is used to filter word associations if their word spans are less k. Since word associations don't incorporate the syntactic information of the sentence, applying the word span help remove word associations that probably don't have association.

Below is the table of number of word associations before and after applying word span:

| | Total Number | Percentage |
|---|---|---|
| word associations | 1731 | 100% |
| k = 3 word associations | 219 | 12.7% |
| k = 5 word associations | 491 | 28.4% |
| k = 10 word associations | 1514 | 87.5% |

Form this table, we can clearly see how much the size of word associations shrinks after applying word span. When k = 10, it shrinks very little; while when k = 5, it shrinks to almost to its quarter.

3. WordLists

WordLists are used to incorporate syntactic information of the sentence. Basically, the sequence of word orders matters. By creating wordlists, we increase the size of search space, but lower the corresponding support. Below is the table:

| | Total Number | Percentage | Min Sup | Max Sup |
|---|---|---|---|---|
| k = 3 word associations | 210 | 100% | 5 | 117 |
| k = 3 wordLists | 301 | 143% | 1 | 117 |

| | Total Number | Percentage | Min Sup | Max Sup |
|---|---|---|---|---|
| k = 5 word associations | 491 | 100% | 5 | 117 |
| k = 5 wordLists | 710 | 145% | 1 | 117 |

| | Total Number | Percentage | Min Sup | Max Sup |
|---|---|---|---|---|
| k = 10 word associations | 1514 | 100% | 5 | 117 |
| k = 10 wordLists | 2285 | 151% | 1 | 117 |

Despite increasing the size of candidates, wordLists greatly decrease the minimum support for each candidate. This helps to eliminate less meaningful word associations.

4. Recommended solution

In order to identify family history information of a patient in the format: (family member, disease), we need to have the disease information.

From the beginning, we extract sentences containing both family member and disease. Then we can follow the procedure described in this project, and find the top N wordLists contain both family member and disease. Also mining the characteristics of each family member can be helpful to find the pattern in the format (family member, disease). This approach will greatly shrink the data we are trying to analyze.

However, the disadvantage of this approach is the definition of disease. In order for the data to be beneficial, it needs to include a more detailed description of the disease. There are many diseases, which could branch into more specific diseases. For example, there are many forms of cancer, and many stages. Also, there are survivors of cancer, and people who died from cancer. Also, people may use different words to describe the same kind of disease.

One solution to the problem of disease is to sample a small portion of the data to mining the frequent occurring diseases, and use this as the disease information.