A Project Report on

Association Pattern Mining

Using association patterns to extract family history from medical transcripts

Individual Project Submitted by

Purva Zinjarde

Instructions on running the program:

1. The program requires the path of the data folder set in the os.chdir('./data') command.
2. Once you set the data path, please comment the “os.chdir” line of code.
3. The program will run on its own and give an output at every step. Detailed step by step description will be provided ahead in the document.
4. The program takes as an input the span of the token sentence. Values 5, 7 or 10 can be entered.
5. A few steps may take some time to process (~40 seconds), especially the step that calculates the sentence span.
6. If the program misbehaves or gives an incorrect output, please try refreshing the browser.

Project Description:

The mini-project entailed working on a collection of transcribed medical reports and by using association patterns to extract family history from medical transcripts.

The working strategy I followed is as given below in a step-wise format.

Step 1: Import all the txt files from the folder - import, decode, and convert from bytes to string.

Step 2: Tokenize the paragraphs and put data in an array of strings**.**

Step 3: Remove all the stop words, white spaces (\n and \t), all numbers and any symbols.

Step 4: Check if the keywords (here, the list of family members provided) are present in the sentences.

Step 5: Only select the lines containing the key words (family members) and append that into an array.

Step 6: Calculate the minimum support threshold value. Since we want to have identify all the word associations that occur in at least 5 sentences, we calculate the support threshold using the formula:

$$\frac{No.\,of\ min\ sentences}{Total\ sentences} = \frac{5}{318} \approx 0.02$$

Step 7: Convert array into dataset, columns would be the individual tokens and rows would be a boolean value as to whether or not they appear in the individual token sentence.

Step 8: Using mlxtend, apply the apriori function and get the association rules.

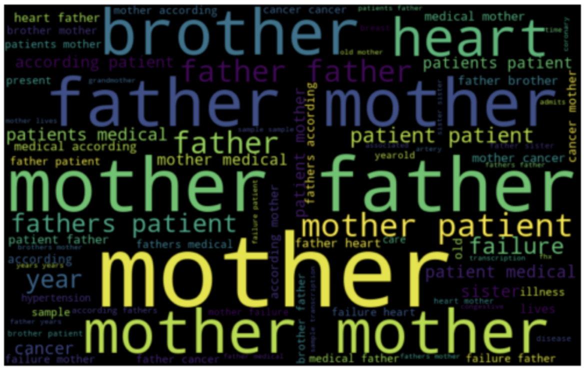
Step 9: Sort the rules in descending order of lift.

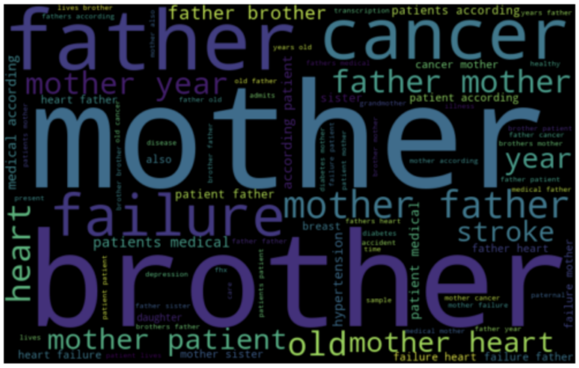
Step 10: Data visualization using the matplotlib library to see a relation between Support, Confidence and Lift.

Step 11: Filter the rules with respect to lift and confidence. Remove unwanted columns from the dataset.

Step 12: Convert the columns containing the antecedents and consequents into a list.

Step 13: For additional data visualization, I have converted the list of antecedents and consequents into a word cloud, to get a better idea of the dominant words.





Step 14: Get the indexes of the antecedents and consequents from token data. Get the span of the sentences from the user. The user may enter values 5, 7 or 10.

Step 15: Calculate the sentence span for each sentence in the original dataset (not tokenized). If the sentence span is more than k, put it into a different dictionary.

Step 16: Calculate if 60% of the sentences have a sentence span of more than k, then remove the association pattern from further consideration. Create a main list of all correct association patterns.

Step 17: Check for relative arrangement of the association patterns. Combine all the relative association patterns into a WordList.

Step 18: Sort the WordList in an ascending order of frequency.

Step 19: Perform analysis on the WordList containing the association patterns.

Results were found for the following scenarios:
1. How many contain at least one family member?
2. How many contain at least one of the following diseases?
3. How many contain one family member but no disease?
4. How many contain both a family member and a disease?
5. How many contain neither a family nor a disease?
6. With respect to each family member listed in 3.1 above, what common characteristics can you observe in its top 20 frequent wordLists?
7. (Extra) WordList that contains atleast one disease but no family member.

NOTE: The results can be found in the python notebook.

Detailed Analysis Report and Conclusions:

Detailed discussion of the quality of the word association patterns, wordLists, the impact of the parameter k, and the recommended solution in Step 3.8.

Quality of the word association patterns: The word association patterns derived from the individual tokens was heavily based on the frequency of the keywords. Hence Sparse Representations were not considered. As a result, the diseases or family members that did not have a heavy representation were not really considered in the final list of association patterns.

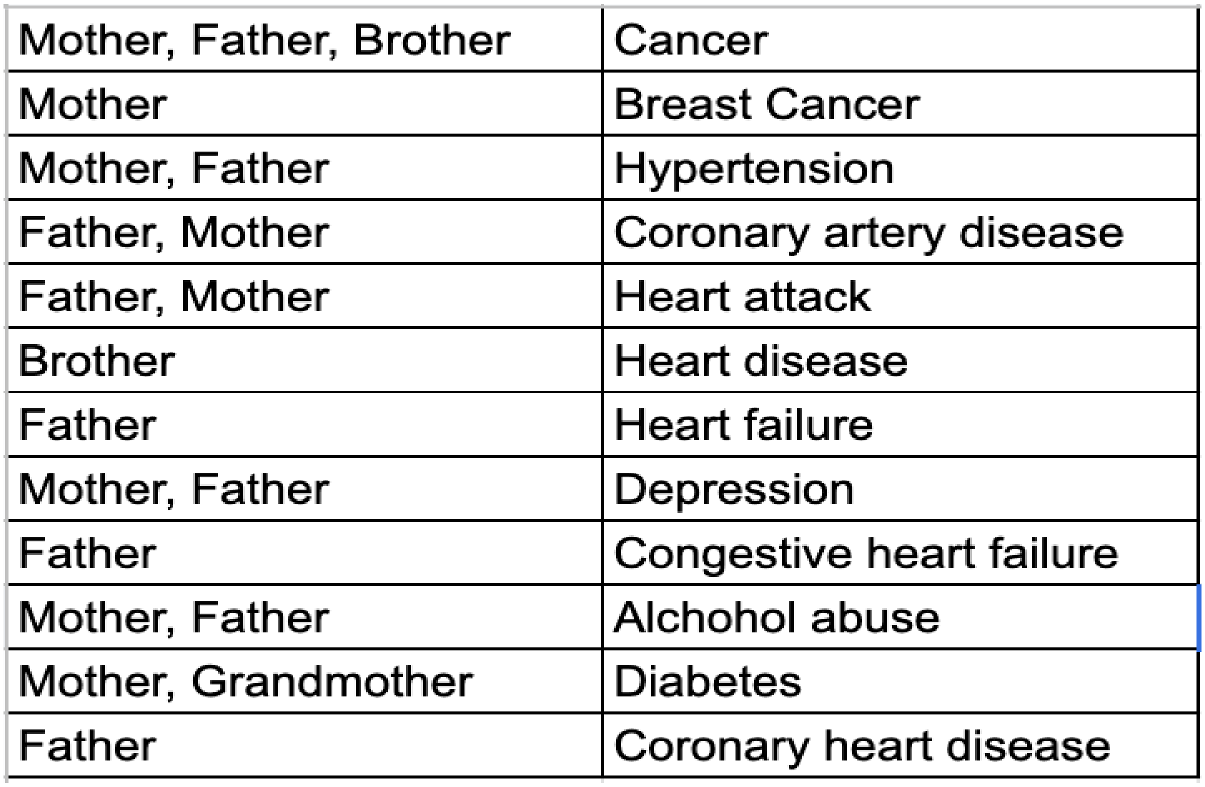
Quality of the wordLists: The wordLists generated had a fair combination of the frequently occuring diseases and family members. I was able to find a lot of pattern associations and was able to derive a conclusive result by processing the word lists and finding the associations. However, since there were some sparse associations, they did not get fair representations and were not included in the WordList.

The impact of the parameter k: The program took a parameter k as an input, which decided the span of the sentences, with respect to the association. I found that the value 5 for span k gives a concise result with lesser, but more accurate result set. There wasn't much difference between a span of 5 and 7, but 7 had a few more items in the WordList than span 5. An input of 10 had significantly more items in the result set WordList, but the data may not have been more accurate.

Conclusive remarks and Potential Solution:

From the wordList, following solution can be proposed.

- It seems that cancer is the most indicative of a heriditary nature.
- There seems to be the maximum correlation between cancer and father, mother and brother. Breast cancer, especially, is related to mother.
- Hypertension is also correlated with mother and father.
- Diabeters seems to be more related to mother and grandmother.
- Heart diseases, coronary heart disease, coronary artery disease seem to be related to father.
- Alchohol abuse and alchoholism along with mental illnesses seem to be inherited from mother.
- Mother seems to be the most frequent in all the family members, followed by father, brother and sister .Least common family member seem to be grandmother, son and daughter.
- This means that most of the diseases seem to be inherited from mother and father.



Mother, Father, Brother	Cancer
Mother	Breast Cancer
Mother, Father	Hypertension
Father, Mother	Coronary artery disease
Father, Mother	Heart attack
Brother	Heart disease
Father	Heart failure
Mother, Father	Depression
Father	Congestive heart failure
Mother, Father	Alchohol abuse
Mother, Grandmother	Diabetes
Father	Coronary heart disease

Pros: There is clear correlation between aforementioned diseases and the family members mentioned.

Cons: The correlation can only be found for the most commonly found diseases. There isn't much data available for all the diseases to find a clear correlation. More data is needed for finding association patterns for all diseases.