**The Structure of the Output Files**

Assume the output directory , given was C:\Users\Swati\Desktop\Results

1. Output of Step-0:
   * *C:\Users\Swati\Desktop\Results\filename.txt*
     + Contains the absolute path to the data files.
2. Output of Step-1:
   * *C:\Users\Swati\Desktop\Results\Step1out\sentences\_allwords.txt:* This file contains all the candidate sentences (sentences containing family members) and has all words ( no stop words or numbers removed) – this file is later used to check the span of the associations , and the order of the association words.
   * *C:\Users\Swati\Desktop\Results\Step1out\sentences\_nostopwords.txt :* This file contains the candidate sentences after stop word removal. This file is used to create bag of words which is later used by Apriori algorithm to find the word associations.
3. Output of step-3:
   * *C:\Users\Swati\Desktop\result2\Step1out\BagOfWords.csv* : This is csv file generated from step-3 , it contains bag of words used by Apriori.
4. Output of step-4:
   * *C:\Users\Swati\Desktop\result2\Step1out\output.txt* : This is the output of Apriori it gives the word associations and their support .
   * Eg : cancer mother (23) where 23 is the support.
5. Output of step-5:
   * *C:\Users\Swati\Desktop\result2\Step1out\span3.txt*
   * *C:\Users\Swati\Desktop\result2\Step1out\span5.txt*
   * *C:\Users\Swati\Desktop\result2\Step1out\span10.txt*

Contain the word association that satisfy the span text , span3.txt has word associations that lie within 3 words, and span5.txt has word association that lies within 5 words and span10.txt has 10-spanned word associations.

In this step I have eliminated single words from further consideration.

1. Output of step-6:
   * *C:\Users\Swati\Desktop\result2\Step1out\FrequencySets\_3.txt*
   * *C:\Users\Swati\Desktop\result2\Step1out\WordList\_3*

This step generates above two set of files for each 3-spanned , 5-spanned, and 10-spanned inputs ( so there are total 6 files)

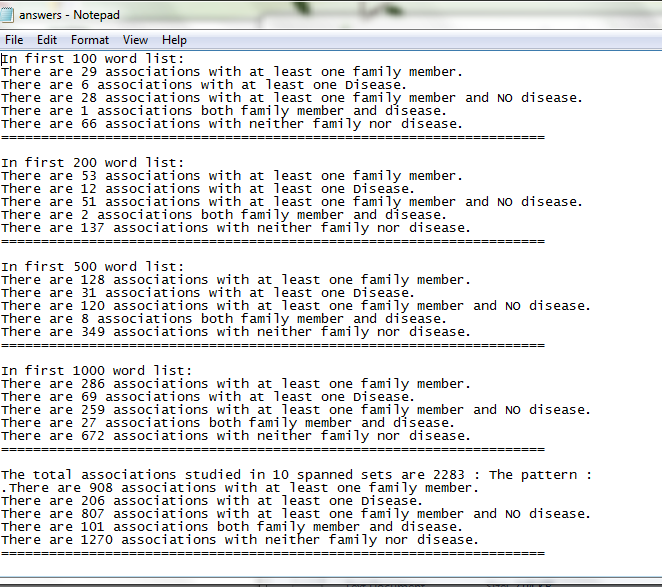
> FrequencySets\_k.txt files have all the permutations and their frequencies.

> WordList\_k.txt has same word associations ordered with respect to their decreasing frequencies. The numbers indicate frequency. The word associations with 0 frequencies are not included

1. Output of Results:
   * *C:\Users\Swati\Desktop\Results\Step1out\Ans\_freqSpan\_3*
   * *C:\Users\Swati\Desktop\Results\Step1out\Ans\_freqSpan\_5*
   * *C:\Users\Swati\Desktop\Results\Step1out\Ans\_freqSpan\_10*

Include the answers to the questions which are summarized in ***answers.txt*** within each of the folders.

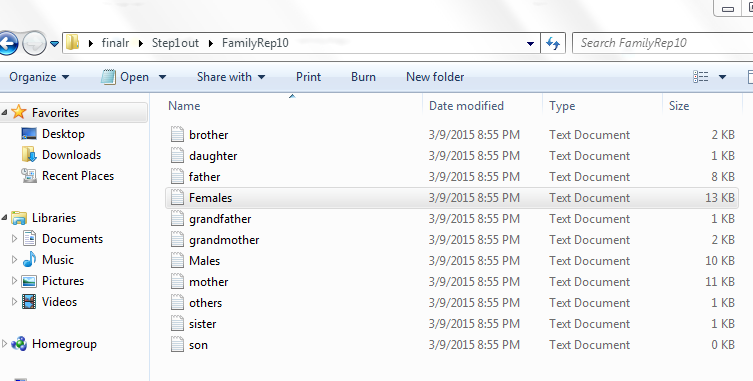
Following is the snap shot of answers from 10-spanned folder:



For the last question : to study the characteristics of family member wise wordlist following folders are generated:

* *C:\Users\Swati\Desktop\result2\Step1out\FamilyRep3*
* *C:\Users\Swati\Desktop\result2\Step1out\FamilyRep5*
* *C:\Users\Swati\Desktop\result2\Step1out\FamilyRep10*

Which contain:



These files have word association separated based on family members.

Further discussion of the results in included in Discussion.doc