Mini Project 1

Description of the main steps implemented:

(All the steps and related files to those steps are provided in the precise manner in their respective folders.)

A README.txt has been provided in each folder which describes how the program can be compiled and executed along with the input and output files.

Step-1: Identifying the sentences that contain family members.

- Using Neil Lewis's sentence-splitter program, I was able to extract the family words out of the 1241 reports. The number of sentences came out to be 872.

Step-2: Removing stop words and other noise.

- I implemented a python script to remove all the stop words using the nltk (Natural Language Tool Kit). I also removed other noise characters such as ‘,’ ‘.’ ‘:’ ‘;’ ‘’’ ‘”’

Step-3: Identifying all word associations.

- Using the FPGrowth.exe provided on the website of Christian Borgelt, I was able to implement the algorithm on my transformed sentences and was able to extract the word associations. The total associations came out to be 2155.

Step-4: Identifying word associations for span k = 3, 5, 10

- I implemented a python script that takes the associations formed from the step-3 along with the original family sentences as input and produce the associations according to k, where k is a parameter. I experimented with values of k as 3, 5, 10.

Step-5: Making ordered patterns (wordLists)

- I implemented another python script that takes the Span associations formed in the last step individually, along with the original family sentences to produce ordered word lists, ie, permutated word associations with their frequencies. I ignored the permutations with frequency = 0.

Anshul Vyas
SFSUID: 915584987