**Discussion**

In this assignment, we followed a basic process of text mining to find interesting associations. We tokenized the text from medical records into sentences and selected candidate sentences which contained one or more family members. Selecting sentences with family members helped us in narrowing the word associations to those containing vital information related to family. Stop words were removed from the sentences and normalization process such as: changing the text to lower case, changing plurals to singular further helped in narrowing and highlighting the important word associations.

The word associations found in this project highlighted the following findings:

1. The numbers of word associations with family, diseases, family & no diseases, family & disease, no family & no disease appears to follow a similar distribution pattern in 3-spanned, 5-spanned and 10-spanned wordlists. (fig-1, fig-2 and fig-2)
2. 5- spanned wordlist resulting is more associations with both disease and family when top 500 frequent wordlists were considered.
3. In 5-spanned wordlist the most common disease found in association with family member was “cancer” while in 3-spanned wordlist it was “depression”. The association of cancer with family member was missed when span was reduced to 3 for consideration. The 10-spanned wordlist has similar pattern to 5-spanned wordlist.
4. When considering the associations at family member wise:
   1. Father was associated with words like “died, death “ and diseases like “stroke , mi, heart”
   2. Mother was associated with “died, depression” and diseases “ cancer, depression”
   3. Sister was most commonly amongst other family member associated with “cancer, breast, breast cancer”
5. Analysis was extended to find words associated with male and female family members : Females associations commonly had “cancer” as compared to male word associations.

**Limitations with current process of finding associations and possible suggestion:**

1. While removal of stop words generic English stop words were removed, more specific stop word list should be generated and their removal should be considered. Presence of such words created artifact word association

example: word “sample” –usually family members collect lab reports .

word : “ consent” – very commonly consent is taken from family members.

**Solution**: Thus specific stop word list should be creates

1. Splitting word based on spaced resulting in breaking of some important identities like “cardiovascular disease”, “renal cancer”, “breast Cancer”, “cerebrovascular “

Solution:

* 1. sequence of n consecutive words (2 grams, 3 grams or n grams should be considered in mining associations.
  2. Medical adjective lists should be created. Such words will usually have following word associated with it. Eg : cardiovascular, cardiac, pulmonary , renal etc. Identification of such words could be used to avoid splitting of important word associations.

1. Semi- structural nature of medical reports:

The medical reports usually are semi-structured, with important heading like “Family History” identification of such headings may help in selecting more appropriate candidate sentences. Further exclusion of such headings in word associations is necessary to avoid creation of artifact associations.

Figure : 3-spanned word associations

Figure : 5-spanned word associations

Figure : 10 spanned word associations