**Text Mining based on MEDLINE**

**Objective:**

Given a gene list of 6 human genes and several files of MEDLINE citation, testing the co-occurrence of two genes in the same abstract; build a gene network based on the test results.

**Steps:**

1. Extract all the PMID codes from xml files

There are 3 xml files of MEDLINE citations in the folder. Each PMID code in the files refers to a specific article on NCBI website. Given the PMID code, the program can find out and read the abstract of the relevant articles online. Thus, the first step is to extract all the PMID codes from xml files for further text mining.

To simplify the procedure, the three xml files just contain six PMID codes.

1. Get combinations from gene list

Given N genes, we can have gene pairs.

To simplify the procedure, I just included 6 genes in the list.

3. Check Association

For each gene pair, I tested the co-occurrence of the two genes in the same abstract from the six articles. If co-occurrence exists, association is set as TRUE and FALSE vice versa.

**Results:**

The gene network is saved in Network.txt

**Discussion:**

The python program works for a small number of human genes and xml files of MEDLINE citation. Thus, we can apply it to a larger number of genes (such as 20,000) and xml files. The problem to solve is time consumption.

**Possible Improvements:**

Given six articles to search from and six genes to check association, it took around **three minutes** for python program to finish. If we are to test 20,000 human genes based on millions of articles, improvements on the algorithm to reduce time consumption are necessary. One the other hand, we can try parallel computing.