# Part 1:

I chose to download all the files and make local copies of all unzipped files. From there all lines in each file were appended together to form a single file. The reason for this, is that it was easier for me to conceptualize on how to put together a single file string from all the combined files.

# Part 2:

Number of Sentences: 579780

# Part 3:

How many Unigram types: 145282

How many unigram tokens: 16730749

twenty most common words: THE TO OF AND IN A THAT TAIWAN SAID 'S FOR ON WILL WITH IS `` AT AS BY ''

**After stop words removal**

Unigram Types : 145127

Unigram Tokens: 10349849

The number of tokens is greatly reduced. About 1/3 of the corpus consisted of the twenty most common words.

Twenty most common words: TAIWAN SAID 'S `` '' CHINA PERCENT GOVERNMENT ALSO CHEN PRESIDENT YEAR TAIPEI NT TWO MAINLAND US PEOPLE NEW CHINESE

Others:

I ran out of time trying to create my functions for things the NLTK library takes care of. I realized this far too late to go back and redo things so I got stuck at this portion. The run time for my program script was slow. Going back now I would have used NLTK library more extensively to help reduce the amount of work.