Assignment 2 Writeup
Xinyu Zheng

1. **Deliverables**
   1) **3** combined counts for all ten files (use Query String Authentication for one of these files)
   2) **30** summary counts for each individual file
   3) a writeup explaining the details

2. **Libraries I used**
   1) nltk – natural language toolkit which provides easy-to-use interfaces to over 50 corpora and lexical resources. Libraries used in this library include: stem, word-tokenize, stopwords, bigrams, and trigrams.
   2) unicodedata - provides access to the Unicode Character Database which defines character properties for all Unicode characters.
   3) string - contains a number of useful constants and classes, as well as some deprecated legacy functions that are also available as methods on strings.

3. **Process of text processing**
   1) Process an individual document
      ● Tokenize the document into a bunch of unitokens (unigrams)
      ● Remove unigrams that are stop words in stop words set
      ● Stem unigrams
      ● Remove digits and numbers from unigrams
      ● Remove single letters from unigrams
      ● Remove special characters from unigrams
      ● Create bigrams and trigrams for the document
      ● Calculate counts for unigrams, bigrams
   2) Process a combined document
      ● Tokenize each document and combine all unitokens (unigrams) into one document
      ● Repeat the process of individual document processing (shown as above)

4. **Something more**
   Comments are provided with the code.