**Assignment 2**

The files in the folders 105‐extracted‐date and 106‐extracted‐date contain all speeches by U.S. senators in the 105th and 106th Congress (1997‐2000). The name of each file shows the congress‐name‐state abbreviation. For example, the file "105‐akaka‐hi.txt" contains all speeches by Senator Akaka from Hawaii in the 105th Congress (1997‐1998).

The task is to count the frequency of words used by each senator in each Congress. For

example, the number of times in the 105th Congress that senator Akaka mentioned the

word "gun".

1. Write a program that loops over directories and all files in the directories and prints the

full file name.

2. In the loop, read the speech files and split the text based on blanks (space) into words

that are put in a list.

3. Loop over words in the list and remove non‐alphanumeric characters, replace upper by

lower case letters.

4. Use the document droplist.txt and check if the words in this list are also included in the stopword list provided by the NLTK library (you can find this in the inputs folder stopwords.txt). Join both lists in a list called stopwords\_final and drop these words from each list in 3.

5. Count the frequency of each remaining word and save the result in a comma‐separated file in "long" format where the first row contains the variable names: "file", "word", "frequency" and the following rows contain the corresponding values and save the file in the Output folder.