Oftentimes, you will be exposed to somebody else code, and you will try to understand what it does. Different people might have different coding styles and approach the same problem differently. In this case, your job is to reverse engineer their thought process.

Go through the submissions in the folder named submissions and perform the following tasks:

1. Check if the code run. Does the code break at any point? Fix the code if it breaks.
2. Does the code do what it is supposed to be doing?
3. Is the code readable? Are the variables and functions named properly? Rename functions and variables if needed.
4. Rank the files best to worst

These are actual submissions, however, some of the code and paths have been modified so that the author of the code remains anonymous.

Challenge

The files in the folders 105-extracted-date and 106-extracted-date contains 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. *Explore the os library.*

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. Remove non-alphanumeric characters, replace upper by lower case letters.

4. Use the file *droplist.txt* to check if the words in this list are also included in the stopword list provided by the NLTK library (use nltk_stopwords.txt if you did not manage to install NLTK library). Join both lists in a list called stopwords_final and drop these words/tokens from the text.

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.