Program 5

CS 101

Harrison Lara

Algorithm

• Import math and string

• Set each speech variable to an empty string

• Greet User

• Create function for...

* strip all punctuation

o remove stop words, open file and use punctuation function

o word count

o word commonality

* Relative Frequency
  + Only look at single words, not whole pieces

• Open files and use clean function to remove excess punct. and stop words

• Find the word commonality (only print the one with the highest match)

o Count how often a word appears (100 times, 51 times, etc)

o Use for loop to count +1 for common words

o take length of speech + mystery speech - common words

o round off to four decimal places with formatting

• Find he highest word frequency (only print the highest one)

o Create empty dictionary

o use for loop to with mystery then if statement with speech

o append word from mystery to speech if present in both speeches

o make sum = 0

o variable1 = take speech of dict. word and divide by speech length

o variable2 = take mystery of dict. word and divide by mystery length

o use sum and add variable 1 and 2

o take total = sum divide by length of distinct words

* use square root
* Use mean square root
  + Find all words that the 2 documents have in common (that is, words that appear in both).
  + For each such word: Find the difference in relative frequency between the two documents
  + Square that difference
  + Keep a running total of the sum of the squares
  + When the sum of the squares for all words is known: Divide the sum by the number of distinct words the documents have in common
  + Find the square root of that quotient
* open files for reading and apply functions to strip them of excess
* apply each word common and frequency function to two speeches (do this for all 16 results)
* see all results and print max and min for the common and freq, then print the ones that match best