#Course: CSCI 4140

#Name: Huan-Yun Chen

#Date: 1/29/2018

#The following program is program in Python 3 and Python IDLE

A. I picked news reviews religion government and romance, because these are the genres I'm more familiar with these topics
B. First I pick the set of words that I think would match the genres. Like for romance I would pick word like ["love","live","forever","life"], for news I pick ["crime","society","victim","safety"]. Than I would try words set that is more commonly seen in the articles such as ["could","would","can","do","does","should"] the amount of counts increase significantly. I also tried the pronoun set, romance corpus has the most counts for all the pronouns especially he she and I. I was surprise when government has 0 count for using the word "she". When I choice words set I would try to match them with the genres I'm focusing on.
C. Conclusion: Different corpus has really different word sets. Base on their topic the amount of words being use could result differently. My hypothesis was proof wrong at the very beginning. The words I choice I thought they would give me a really high counts because they all match with the topic I choice but the result was totally the opposite. All the pronoun has a pretty good counts, government corpus has high counts on certain words than I thought it would be. Lower case and upper case makes a different too