John Mikos
Homework 3

1. I followed gensim's training a model tutorial to get the code I currently have. I followed along and tried to figure out how the built in lee_background.cor corpora worked. After I had a good understanding of how gensim's word2vec model worked, I created a new class to test on the masc corpora which can be found through the anc site provided in hw3.pdf. Since many of the files are short and spread apart, I created a new text file that combined all the text in the written/jokes directory. The only issue I found with this corpus is that it contained all of the website scripting (html tags and more) which showed as the most common words in the vocab since I did no data cleaning. After this I created model and tested on pairs of words by using model.similarity(word1, word2). When I did this, I would get similarity rating in the high 90%. I felt that these high percentages were because the data set was so small. I did find that words that would seem to belong with each other in jokes had slightly higher percentages than those that did not.
2. I completed this part by following gensim's tutorial. I followed the same example provided and some issues arose due to my download of gensim. Eventually I got the same results presented in the tutorial. Word2Vec considers these words similar because they have similar appearance in sentences. Say king and monarch are very similar words or almost interchangeable, well the only way word2vec will figure this out is by testing the words around it and their similarity.
3. I continued to have issues with gensim download and needed help understanding how the code worked. Eventually I got the code to work. I ended up getting a spearman correlation constant of about 0.74.
4. Some word analogies I thought of and tested are listed below:
    - King – Man + Woman = Queen
    - Bread + Sauce + Cheese = Pizza
    - Wolf – Dog + Cat = Lion
    - Mouse + Large = Rat
    - Christianity – Jesus = Judaism
    - Flour + Egg + Milk = Cake
    - Turkey – European = Asian

    I then implemented them in my code.