Dr. Tony Diana
DATA 690 Introduction to NLP
Homework, Week 10

Use the imdb-reviews_2.csv file.

- Load the data
- Count the number of positive and negative sentiment
- Clean the text 'review' and create a column called 'review_processed'
- Replace short words as

```
df['review_processed'] = df['review_processed'].apply(lambda x:
' '.join([w for w in x.split() if len(w)>2]))
```

- Make entire text lowercase
- Remove stopwords
- Make custom list of words to be removed including 'movie', 'film', 'one', 'make', and 'even'
- Add to the list of words
- Lemmatize the text
- Convert NLTK tags into 'wordnet' tags
- Find the part of speech tag
- Lemmatize sentences using POS. Tokenize the sentence and find POS tag for each token
- Define 'wordnet_tagged.' If there is no available tag, append the token as is. Else, use the tag to lemmatize the token
- Plot the most frequent words from positive reviews using bar chart. Subset positive review dataset, extract words into list and count frequency. Subset top 30 words by frequency in a horizontal bar chart
- Create a word cloud
- Import CountVectorizer and create a sparse matrix of 2,500 tokens. Split the data set into train and test (20%) set
- Use the GaussianNB to train the model on the training data. Provide the accuracy of the model
- Test the model on the test set
- Create the confusion matrix and classification report
- Use Seaborn heatmap to show TP, FP, TN, FN values
- Use a logistic regression model and check the accuracy of the model for C=0.01, 0.05, 0.5, 0.5, and 1.
- Provide the ROC curve. What is the area under the curve?