Dr. Tony Diana
DATA 690 Introduction to NLP
Homework, Week 7

---

**Exercise 1.** Sentiment Analysis

- Using textblob, what is the probability that the sentiment in the Burbank text is going to negative?

**Exercise 2.** Sentiment Analysis

- Using the data from exercise 1 and textblob, what is the overall sentiment and subjectivity?

**Exercise 3.** Key topic using 'Word' from textblob (very simple way to determine the key topics) based on the Burbank text file.

- Import Word from textblob. Identify the key topics by using Word from textblob.

**Exercise 4.** Sentiment analysis with spaCy.

- Load the datasets 'amazon_cells_labelled.txt', 'imdb_labelled.txt', 'yelp_labelled.txt'
- Create 'combined_col' by joining the tables such that combined_col=[data_amazon, data_imdb, data_yelp]
- Check the structure of data_amazon
- Add headers for columns in each dataset: 'Review' and 'Label'
- Create a "Company' column to identify each company 'Amazon', 'imdb', and 'yelp'
- Explore the structure of the new dataset called 'comb_data'
- Use 'comb_data.to_csv' to create the 'Sentiment_Analysis_Dataset'
- Print the columns
- Check for null values
- Import STOP_WORDS from spacy and stopwords from spacy.lang.en.stop_words
- Build a list of stopwords for filtering
- Import string, define 'punctuations' and define a 'parser'
- Tokenize the sentences

- Import ‘CountVectorize’, ‘TfidVectorizer’, ‘accuracy_score’, ‘TransformerMixin’, ‘Pipeline’, and ‘LinearSVC’
- Create a class 'predictors(TransformerMixin)'. Within the class, define 'transform', 'fit', and 'get_params'
- Create a basic function to clean the text
- Vectorize and use LinearSVC as a classifier
- Use TfidfVectorizer
- Split the ‘com_data’ dataset into a train and test (20%) set
- Create a pipeline to clean, tokenize, vectorize, and classify as ‘pipe_countvect’
- Fit the data
- Predict with the test dataset
- Prediction results as ‘1’ for positive reviews, and ‘0’ for negative reviews
- Use print(sample, “Prediction➔”, pred)
- Determine the accuracy for the test dataset, X_test/sample prediction, and train dataset