95-851 Making Products Count: Data Science for Product Managers
Homework 3: Natural Language Processing
Fall 2022

Due 11:59 PM November 30, 2022

Overview

In this assignment, you will practice using natural language processing techniques to understand tweets about self-driving cars. You will analyze a set of tweets and find the words in the tweets that are most predictive of positive, neutral or negative sentiments about self-driving cars. The submission will be a Jupyter Notebook with the name *DSPM_HW3_<your Andrew id>.ipynb*. The deadline for the homework is **November 30, 2022 11:59 PM EST.**

Tasks

The data file for you to analyze is in the assignment in Canvas (Twitter-sentiment-self_drive-DFE.csv). For the purposes of the questions below, you can ignore the unit_state, sentiment_gold and sentiment_gold_reason fields, as only a few of the observations have been tagged this way (although you may find it instructive to compare the automatically calculated sentiment to the human-generated sentiment_gold and sentiment_gold_reason field where they are available).

1) Perform exploratory data analysis (for example, summary statistics and histograms on the numeric fields) on the data and indicate if you foresee any problems using the data to build a model to predict sentiment. (5 points)

2) Tokenize the Tweets for analysis. Show the most common words/tokens overall, nouns, adjectives for all of the tweets. You should exclude stopwords and the frequent words/phrases like “self-driving cars” that do not help in the prediction task. (Note: if you decide to stem or lemmatize the words for modeling in the following step, which is recommended, you should do that after the part of speech tagging in this step.) (5 points)

4) Create a model to predict what sentiment a Tweet will be given on the 1-5 scale. Note that you do not have to run any sentiment analysis packages (e.g. Vader or TextBlob) to build this model, just use the features in the data (including words in the text) as predictors. What are the 10 top words/tokens for each sentiment rating? (5 points)

5) Evaluate how well the model performs for each rating (1-5). Where does it make errors and how might you improve it? (5 points)

HW3 Resources

- Learning to classify text can be found at http://www.nltk.org/book/ch06.html