Homework 4:
Advanced Data Analysis in Python

Due end of day Wednesday, June 2, 2021

The purpose of this homework is to familiarize you with feature extraction and build on your knowledge of machine learning tools. Specifically, you are to analyze the sentiment of the provided open-ended survey responses on immigration. The provided csv has a column of text and a column of measured sentiment.

First, you are to use word frequency as the extracted features. Split the data into a 50/50 train/test split. Train an algorithm using the features and the sentiment as a target. Use a support vector regressor and a random forest regressor. Optimize both the SVR and the RF separarely (but using the same train/test data). Compare the accuracy of the two algorithms for these data.

Next, use TF-IDF as the features. Re-optimize the models, and note any discrepencies. Again, compare the accuracy of the two models. Finally, use bigrams instead of unigrams for the features. You will find the following link very helpful: `https://scikit-learn.org/stable/modules/feature_extraction.html`. Simply search for bigrams to find the relevant part. Optimize and compare. Do not change your train/test data throughout the comparisons to make the results more comparable. If you wish, you can also add a third algorithm, Gaussian process regression, for each of the steps, but this is optional.

Just include your code in the submission.