Homework 4:
Advanced Data Analysis in Python

Due end of day Friday, May 22, 2020

The purpose of this homework is to familiarize you with feature extraction and build on your knowledge of machine learning tools. In the Gaussian process lecture, we saw how we could use a GP to analyze the sentiment of open-ended survey responses about immigration. In this homework you will build on the code already provided, and extend it.

Specifically, we used word frequency as the extracted features. For this assignment, you must figure out how to extend this to include bigrams, frequency of word pairs. 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.

Once you have the bigram feature matrix, simply rerun the GP and see if the correlation changed between the estimates and the ground-truth. If you would like, you could use a different machine learning algorithm other than GPs, but this is optional.

Just include your code in the submission.