- **(5 points) Problem:**
  The data I have are reviews of whiskey out of 100 and a description to supplement the review https://www.kaggle.com/koki25ando/22000-scotch-whisky-reviews/discussion/62308. The Data has 2,247 observations. Naive Bayes, Support Vector Machines, and Principal Vector Analysis are used to predict analyze the text and predict the review value.
  Example of fields used:

  Data[0]

  review 97
  description Magnificently powerful and intense. Caramels, dried peats,

- **(5 points) Data Prep:**
  To prep the data, BeautifulSoup was used to remove html mark up. Anything that is not a letter was removed from the description and all letters were made lowercase. Then all words were vectorized using word2vec from genism.models. One issue that came up when vectorizing was if a word was not in the English dictionary it was assigned 0 which caused some division by 0 problems. An If statement returning the feature_vec was used to avoid this. The vector data was then split into a training and testing set.

- **(25 points) Experiment:**
  The first model used was Gaussian Naive Bayes. The first run of the first model had below a .03 accuracy. To initially improve the model the min_word_count was changed until settling at min_word_count = 16. The last run of the model gave an accuracy of 0.03111111111111111, accuracy was higher for earlier runs. Although the accuracy was low, the average distance away from the actual value is 2.3733333333333335 which is not too far considering the ratings are out of 100. Then Support Vector Machine was performed and the accuracy was 0.09777777777777778, which is better than the Naïve Bayes. The average distance from actual value is 0.6377777777777778 which is pretty good. Than a Principal Component Analysis variant is used to transform data into principal components. The principal components maximize the variance allowing less information loss. With the transformed data Gaussian Naive Bayes, and Support vector machines were rerun with the transformed data. The model for Gaussian Naïve Bayes ran better and its accuracy increased to 0.09333333333333334. The SVC model performed the same.

- **(15 points) Interpret the results:**
  From analyzing the data, I believe that accuracy may not be the best metric to use. The accuracy was low but the mean distance from the actual review was pretty close. A metric that measures mean percent error might represent the model better. This may be true for many datasets with the high target columns (whiskey data reviews were out of 100). When playing around with model.wv the different comparisons showed many reviews were positive, the most negative word I found was medium which explained the high ratings. The relatively low variability and range of the ratings could have resulted in the small mean difference from the actual value. The

algorithms attempt to predict the actual value it would be interesting to have it predict a small range of values and see how much accuracy changes.