Wine Data

Hook

his project starts, revolves around (and ultimately hopes to build upon) word choice. Is it possible to really know what someone is thinking based on the words they use? Can you teach a computer to understand sentiment? What if personal vocabulary labels your speech as an outlier? Is there such a thing? This project hopes to explore some of these questions using wine reviews.

These reviews are always one sentence in length, describing the subtleties of wine flavors. The vocabulary is such in the dataset, that a word like 'brimstone' would be most closely associated with earthy tastes like; 'peat', 'chamomile', 'beeswax', and 'granite'. These reviews also give some clue as to how the wine tasters will score the wine with words like 'excellent' to describe a high scoring wine or universally bad descriptors like 'abrasive' for the under performing wines.

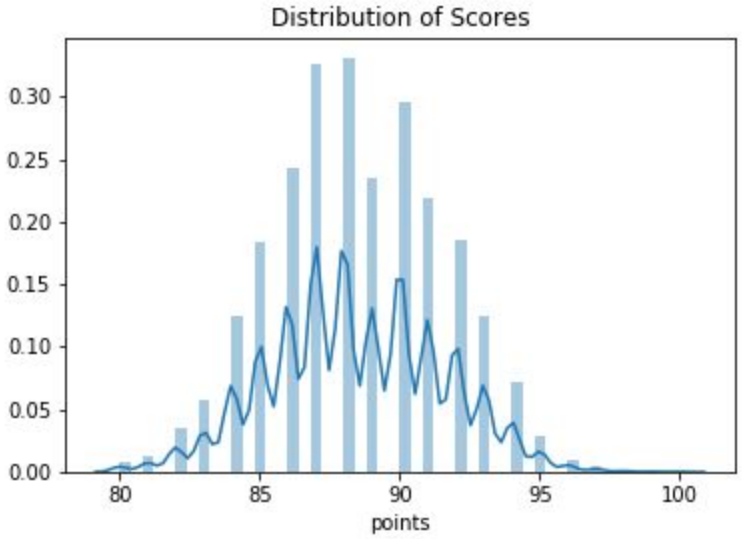
```
[23]: w2v_model.wv.most_similar('coffee')
```

```
[23]: [('espresso', 0.8969775438308716),
       ('carob', 0.7564544677734375),
       ('mocha', 0.7402285933494568),
       ('licorice', 0.7393960952758789),
       ('cocoa', 0.7026122212409973),
       ('char', 0.6930364966392517),
       ('cassis', 0.692969799041748),
       ('molasses', 0.6914410591125488),
       ('coconut', 0.673872709274292),
       ('woodspice', 0.6711657047271729)]
```

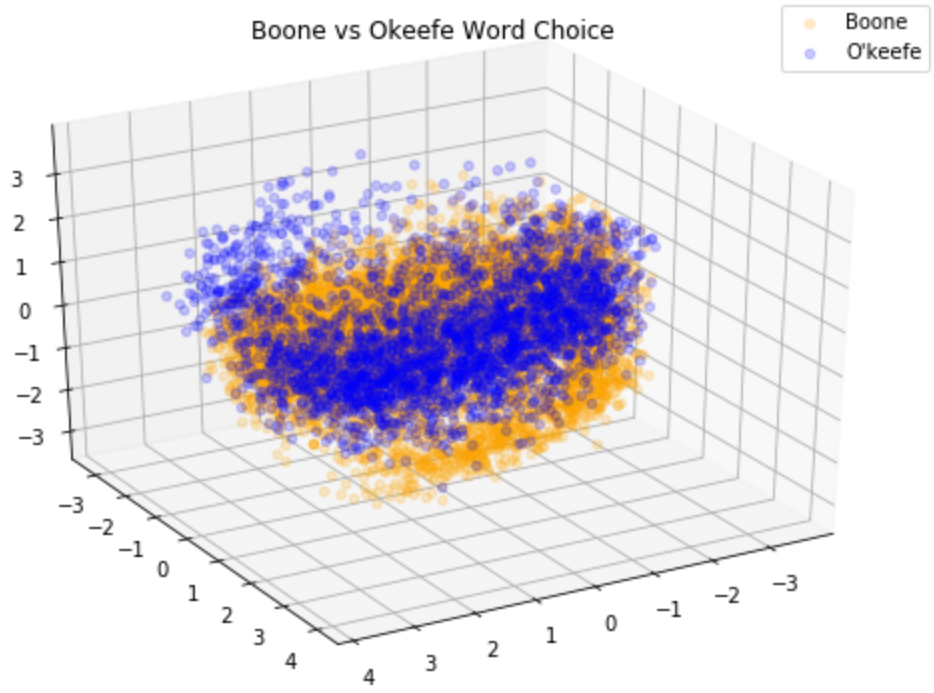
```
[24]: w2v_model.wv.most_similar('brimstone')
```

```
[24]: [('peat', 0.7747483253479004),
       ('flint', 0.7677335739135742),
       ('candle', 0.7435697913169861),
       ('broom', 0.7328553199768066),
       ('bee', 0.7281134128570557),
       ('chamomile', 0.7206580638885498),
       ('wax', 0.719684898853302),
       ('beeswax', 0.7003276944160461),
       ('granite', 0.6992579102516174),
       ('pollen', 0.6967644691467285)]
```

Wine critics are an interesting case study to me, because they have to power to polarize people towards or against something using only their words. These critics are of the fairly toothless variety. They might give a scathing review but still offer a score of 80/100 for the wine. Wine Enthusiast is, after all, selling ad space in the magazine.



We can (and do) go several directions with information like this. First will be seeing if we can train a neural network to accurately predict the score our critics give after their description of the wine. It is an interesting side note that breaking the data apart by reviewer leads variations in word choice by reviewer and sparks the imagination for plotting unique word choice and variance in meaning from word to word and person to person.



For an initial run into NLP I rearranged the score outcome into a bivariate label, either the wine is good, or it is bad. Common sense baseline for this type of model would be the probability of guessing right; it would be 0.5 MAE or looking at it another way it would be 0.5 ROC/AUC.

The random forest model scored a precision of 65%
The most successful neural network scored an accuracy of 90%, it looked like this:

“””

```
model_four = Sequential()
model_four.add(Embedding(input_dim=max_words, output_dim=embedding_dim,
input_length=maxlen,
                         weights=[embedding_matrix], trainable=False ))
model_four.add(LSTM(50, return_sequences=True))
model_four.add(GlobalMaxPool1D())
model_four.add(Dropout(0.5))
model_four.add(Dense(32, kernel_regularizer=regularizers.l2(0.005), activation='relu'))))
model_four.add(Dropout(0.50))
model_four.add(Dense(1, activation='sigmoid'))
model_four.summary()
```

“””