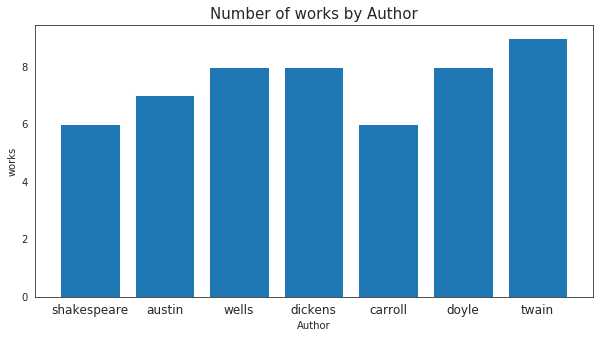
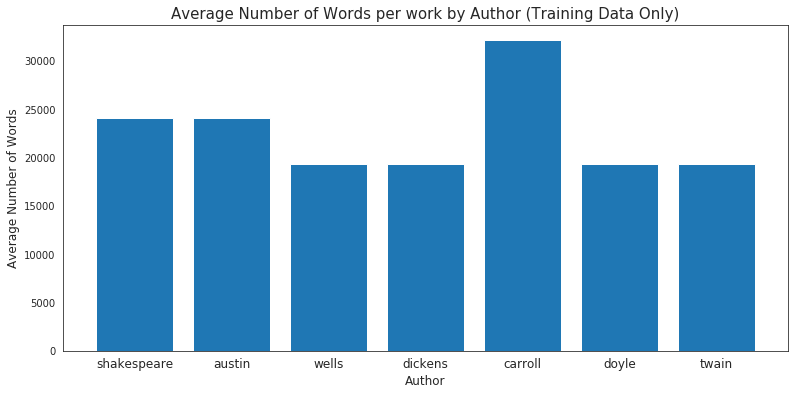
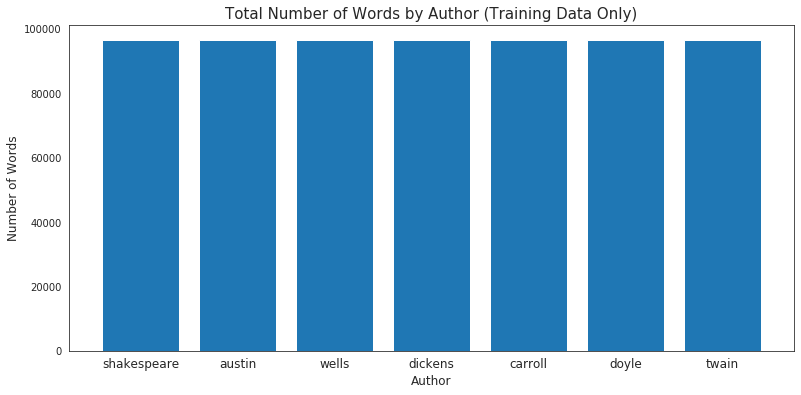
To create the model we used 52 texts from 7 different authors. I tried to stick to the more classical authors found in the Gutenberg Corpus. I decided to try and stick to the top 100 most popular texts and authors and also tried to find authors that had significant number of texts to increase the overall population of the data hopefully helping our model improve accuracy and clusters. I also capped the number of characters possible to 200,000 to address the memory allocation error during the conversion to spacey documents. With the 200,000 characters per text cap our text files are relatively the same size. For the training-test split I weighted the split based on author. The training data had 31 texts and the test data had 21 texts.





We originally attempted more authors but the amount of computing power, memory allocation and time limitations resulted in fewer authors and capping the texts to less than 200,000 characters.

The texts were cleaned and then combined into a corpus by author and then converting the corpus’ to spacey docs. Once this was complete the docs were combined into a dataframe via an array and list.

The dataframe was divided into a training and a test set using a 60-40 split and stratifying it by author to make sure it was divided equally across authors.

Clustering was not very successful at separating the texts by author. The best scores were accomplished by using a n\_clusters of 8. The scores reflected that the clusters were not homogenous and authors were not comprised of a single cluster.

*Using 500 Most Common Words, Scaling Each Sample, and LSA-9*

*ARI score: 0.297 +/- 0.000*

*completeness score: 0.613 +/- 0.000*

*homogeneity score: 0.605 +/- 0.000*

*v-measure score: 0.609 +/- 0.000*

When the clustering is applied to the hold out group the results actually improved. Based on this I think that the clustering model is not the most reliable method of classifying this specific group of texts. I do think that, like many situations, combining efforts can help to stabilize the accuracy of the model.

*Using 500 Most Common Words, Scaling Each Sample, and LSA-8*

*ARI score: 0.365 +/- 0.000*

*completeness score: 0.725 +/- 0.000*

*homogeneity score: 0.733 +/- 0.000*

*v-measure score: 0.729 +/- 0.000*

Final results

The optimum model for predicting the authors was accomplished by using the 500 most common words (some overlapped with other texts) from each text, using Kmeans to cluster the texts into 8 clusters, scaling the data by each sample and using LSA with 9 components. We then ran it through random forest with 300 estimators. Using this model we are able to get an accuracy score of 71.4%.

Our Confusion Matrix:

(The rows are the true labels. The columns are our model's

predictions)

|  | **shakespeare** | **austin** | **wells** | **dickens** | **carroll** | **doyle** | **twain** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **shakespeare** | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| **austin** | 0 | 1 | 2 | 0 | 0 | 0 | 0 |
| **wells** | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| **dickens** | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| **carroll** | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| **doyle** | 0 | 0 | 0 | 0 | 1 | 3 | 0 |
| **twain** | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

We tried using Tf-idf but it didn’t out perform the clustering model. We also utilized LSA and tested multiple component values. Then we ran through a series of combinations and fine tuning the hyper parameters to get the most from each model to determine what the best combination of features and parameters/hyperparameters

The model has issues differentiating between Jane Austin and H.G. Wells and fails to classify even one of the three H.G. Wells correctly. It also failed to classify one out of four texts written by Sir Arthur Connan Doyle correctly and falsely classified one of the Wells texts as being a Doyle. To improve the model we would need to expand the number of texts by each author and include 100% of the texts. With the limitations of my computational power and memory allocation I am not able to accomplish those improvements. I feel confident that by increasing the sample size will allow for even better outcomes.