1950’s and 1960’s Science Fiction Stories
Unit 4 - NLP Capstone

Crystal Bevis, Data Scientist

March 2018

https://github.com/cbevis/U4_Capstone

This is an analysis of 1950’s and 1960’s science fiction stories from Project Gutenberg. The stories are from science fiction magazines such as *Galaxy Science Fiction* and *Worlds of Tomorrow*. The links to the documents can be found in the references section at the end of this document. The files were named with the author name and then a number indicating whether it was the first, second, third, or fourth story downloaded by that author. The naming used in for the text files is included in the references section.

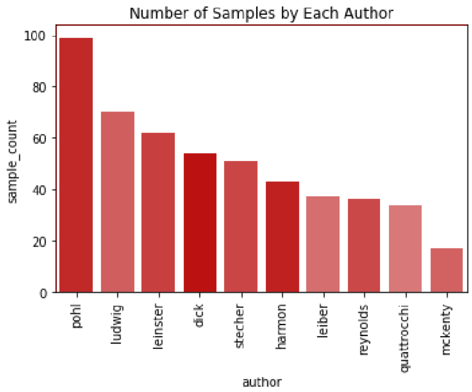
File Cleaning

Each file was downloaded as a text UTF-8 document. The utility function used for cleaning the files funs slowly because it is checking each file for a number of different header and footer options that need to be removed. A different version of the file cleaner could have been written for each type of format, but I decided to sacrifice speed for ease of use and a more universal file cleaning function. In addition, they only need to be cleaned once as the files are written to a new folder that can then be used for the remaining project.

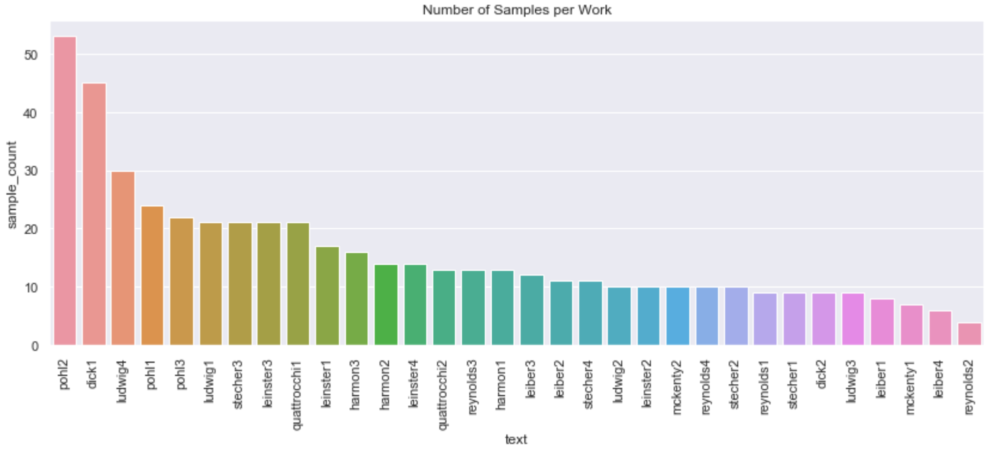
The stories are split at thought breaks that are marked by several astericks or hyphens. These are usually several paragraphs long. Once split into thought breaks, there are over 500 samples of science fiction writing in the data set.

Data Exploration

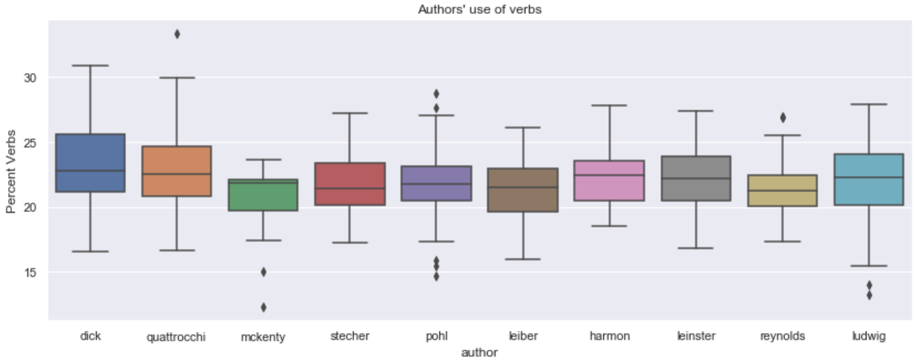
First I looked to see how many of the samples are by each author. There are nearly 100 samples by Pohl and only about a fifth of that by McKenty. There were only two short stories available by McKenty, so I was unable to better balance the data by including more works by him.

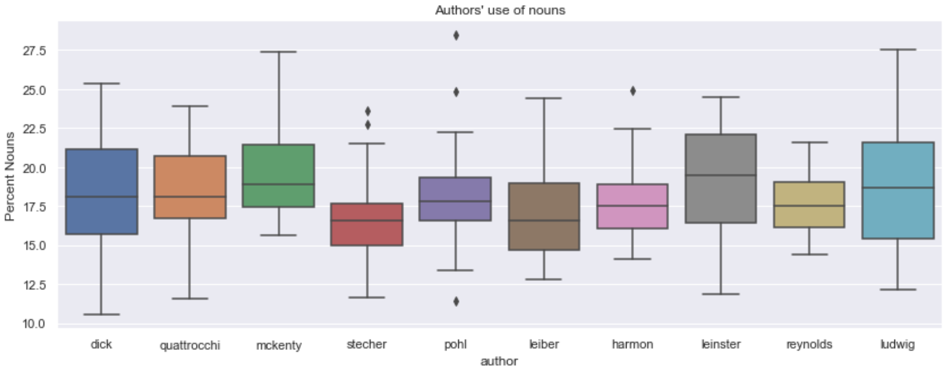


Below is a plot of the number of thought breaks per work. Pohl's "Plague of Pythons" has over 50 samples in the one work. It is more common to have around 10-20 thought breaks per story. One possible consequence of this is when trying to predict the author, words from "Plague of Pythons" may be more more heavily weighted than words from Pohl's other works.



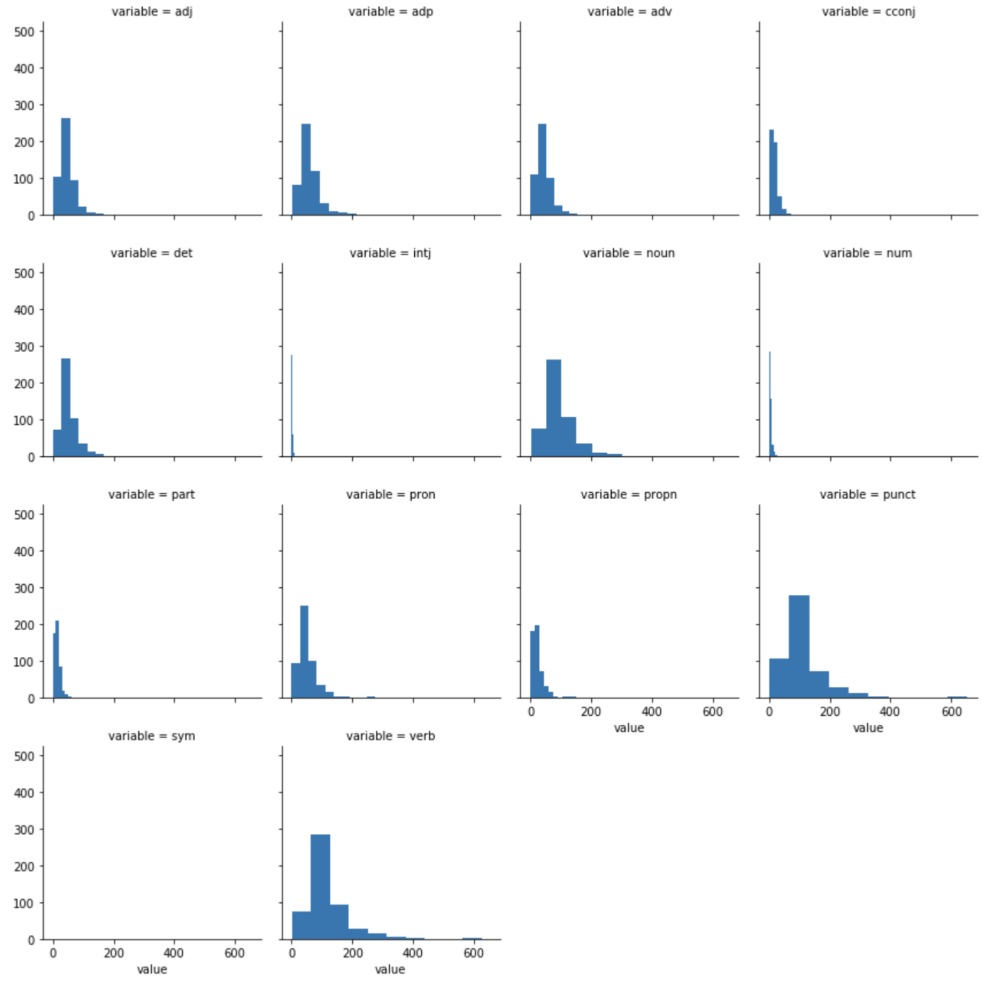
Parts of Speech - Nouns and verbs:





Overall, verbs appear to be used more frequently than nouns. The rate of noun and verb usage appear to vary within each author. This makes sense because different thought breaks or topics could require the use of more more or less of these depending on the point the author is trying to get across. Verbs, nouns, and punctuation are the most frequently used parts of speech as shown in the following facet grid plot drawn using Seaborn.

Parts of Speech Distribution

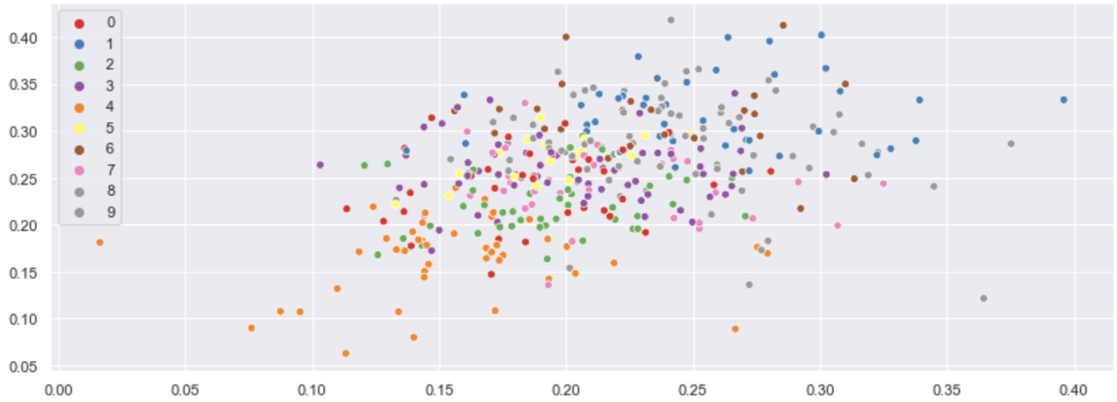


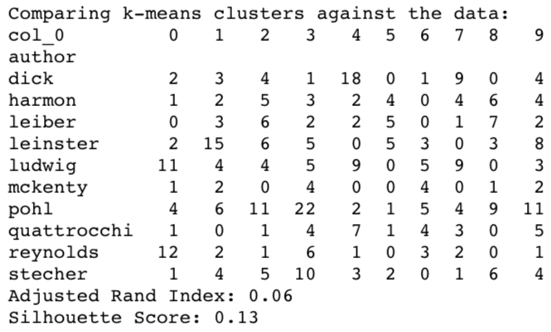
TF-IDF

Tf-idf or term frequency-inverse document frequency was used to analyze the the important words and phrases in the science fiction stories. After tf-idf there are over 2700 features. Truncated SVD was then used to reduce the number of features to 200 before clustering.

Clustering

Several different approaches were tried to cluster the data. When clustered by parts of speech, the adjusted Rand Index score was 0.06 (not much different than random) for 10 clusters. 10 clusters was tried because I had hoped that the authors used different enough parts of speech that we could identify each author as a cluster. This was not very successful.





```
Comparing k-means clusters against the data:
col_0         0   1   2   3   4  5  6  7  8   9
author
dick          2   3   4   1  18  0  1  9  0   4
harmon        1   2   5   3   2  4  0  4  6   4
leiber        0   3   6   2   2  5  0  1  7   2
leinster      2  15   6   5   0  5  3  0  3   8
ludwig       11   4   4   5   9  0  5  9  0   3
mckenty       1   2   0   4   0  0  4  0  1   2
pohl          4   6  11  22   2  1  5  4  9  11
quattrocchi   1   0   1   4   7  1  4  3  0   5
reynolds     12   2   1   6   1  0  3  2  0   1
stecher       1   4   5  10   3  2  0  1  6   4
Adjusted Rand Index: 0.06
Silhouette Score: 0.13
```

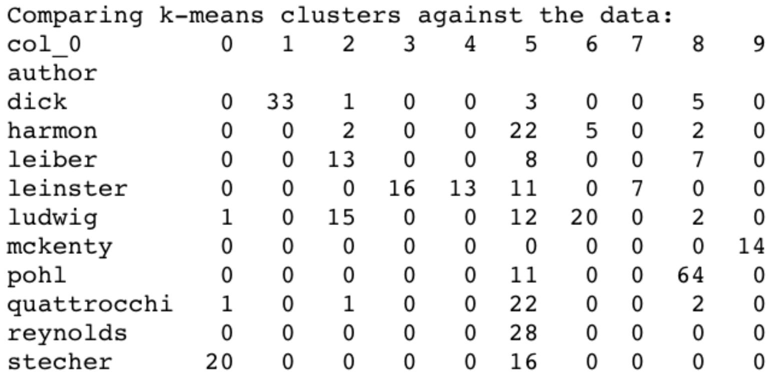
A better approach to clustering was to use the TF-IDF terms. Before clustering with them, I removed all of the names identified by Spacy. Since I only had 2-4 works per author, leaving in the names caused clusters to be formed by character names. This gave a higher Adjusted Rand Index score, but for me was a less interesting result since it is usual for works to have different character names. I was more curious about the other words in the works.

The highest adjusted rand index score for clustering with TF-IDF terms without names is 0.3. The top 12 words the centroids of the clusters are based on are shown below. Several of the words are made up words. This is not surprising because science fiction is, by its very nature, imaginative and two authors are unlikely to make up the same word. This analysis does not measure the frequency at which the authors create words, but that would be an interesting study. Cluster 5 has samples from nearly all of the authors. The words associated with this cluster are “man know time space like look want mr think tell good come.” These are all fairly generic words.

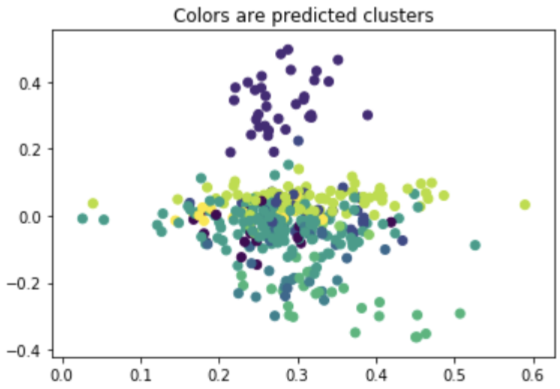
Cluster Centroids

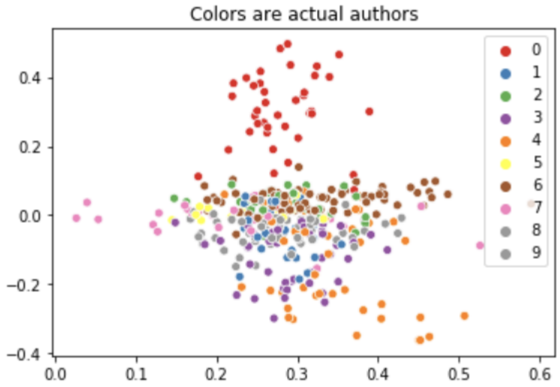
Cluster 0: answer invader universe elephant alpha stitch oracle computer time beta question ship
Cluster 1: tasso klaus bunker ash rudi come claws variety major fire gun soldier
Cluster 2: pa like face blanket eye think hand helmet air earth big martian
Cluster 3: calhoun cattle car ground med maya fence power man city highway ship
Cluster 4: queen passenger ship doctor altaira chlorophage nordenfeld star air lift skipper plant
Cluster 5: man know time space like look want mr think tell good come
Cluster 6: captain lieutenant like sirians planet earth man fox year quade ship remember
Cluster 7: moklin human moklins post brooks trading kid look like good trade sell
Cluster 8: man look exec know mooney like door koitska body burckhardt girl think
Cluster 9: dr senator firework observatory bed work rocket night technician department skyrocket come

Author and Clusters Confusion Matrix



```
Comparing k-means clusters against the data:
col_0         0   1   2   3   4   5   6  7   8   9
author
dick          0  33   1   0   0   3   0  0   5   0
harmon        0   0   2   0   0  22   5  0   2   0
leiber        0   0  13   0   0   8   0  0   7   0
leinster      0   0   0  16  13  11   0  7   0   0
ludwig        1   0  15   0   0  12  20  0   2   0
mckenty       0   0   0   0   0   0   0  0   0  14
pohl          0   0   0   0   0  11   0  0  64   0
quattrocchi   1   0   1   0   0  22   0  0   2   0
reynolds      0   0   0   0   0  28   0  0   0   0
stecher      20   0   0   0   0  16   0  0   0   0
```



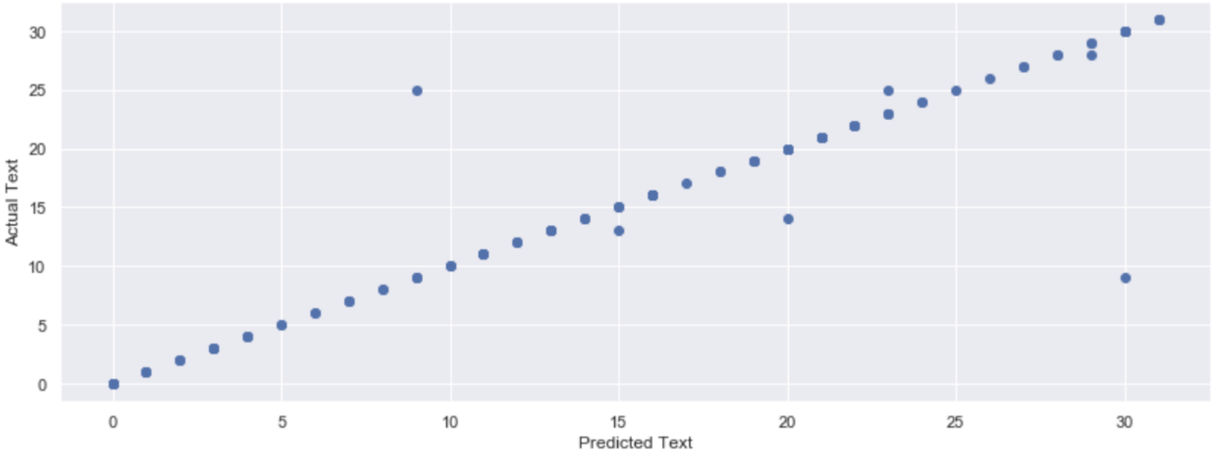


Predicting Author and Title

The TF-IDF dataframe, including names, was used to predict the author. Logistic regression on this gets a perfect training set score and nearly perfect test set score of 0.99. This high level of accuracy is because the passages are nearly all long enough to include main character names that are not shared between works. When logistic regression is used instead to predict which work the passage came from, the results are still pretty accurate. The training set score is 100% and the test set accuracy score is 95%. This reflects some overfitting. This model could potentially be improved by first predicting the author and then predicting the work since only 2 of these misclassifications were between works of the same author and the model predicting authors was more accurate. As it stands, this model treats stecher 2 and stecher 1 as completely different entities, not different works by the same author.

Logistic Regression on TF-IDF Predicting Title

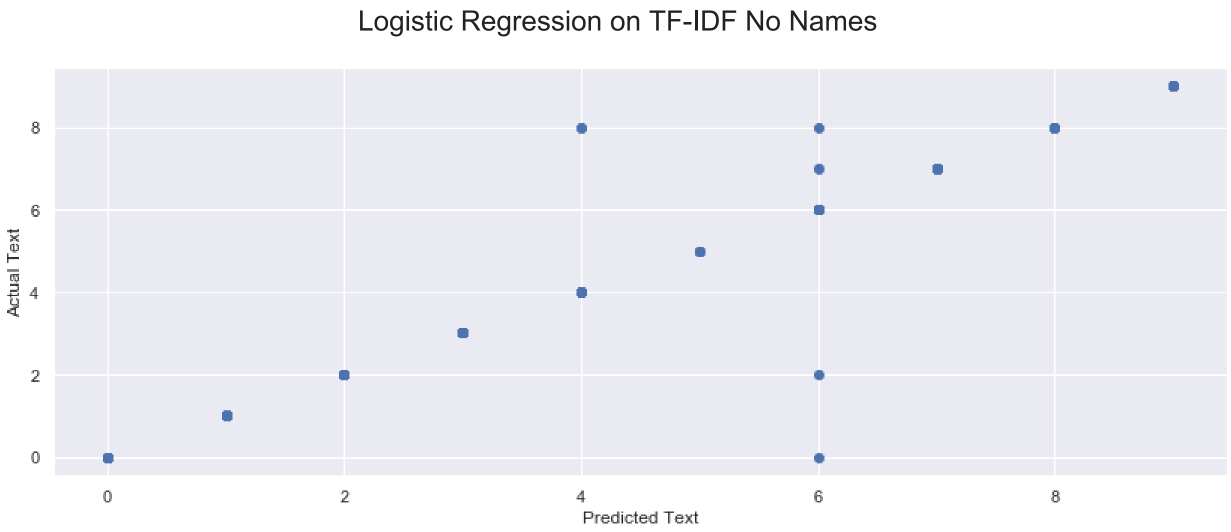
```
predicted text, actual text
stecher3 - leinster1
quattrocchi2 - reynolds2
pohl2 - ludwig2
ludwig3 - ludwig1
leinster1 - reynolds2
stecher2 - stecher1
```



Modeling purely on parts of speech did not perform well. This was expected from how the clustering performed and from the exploratory data analysis at the beginning of the study.

I repeated the modeling using the TF-IDF data set without names. Again, it overfit. The model over-predicts selections to be by Pohl. Going back to the exploratory data analysis, he authored the most selections. Selections by Dick, Leiber, or McKenty are not misclassified. Looking back at the clustering, these three authors were relatively under-represented in cluster 5, a catchall cluster that seems not to differentiate well between authors.

Training set score: 1.0
Test set score: 0.95

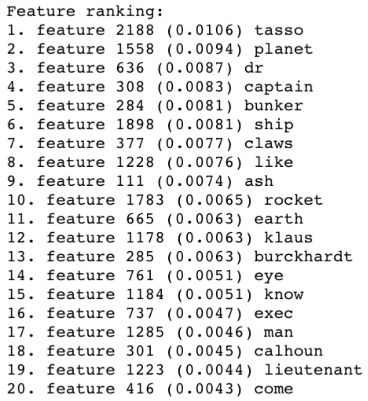

Logistic Regression on TF-IDF No Names

predicted text	actual text
ludwig	reynolds
pohl	reynolds
pohl	dick
ludwig	reynolds
pohl	quattrocchi
pohl	leiber

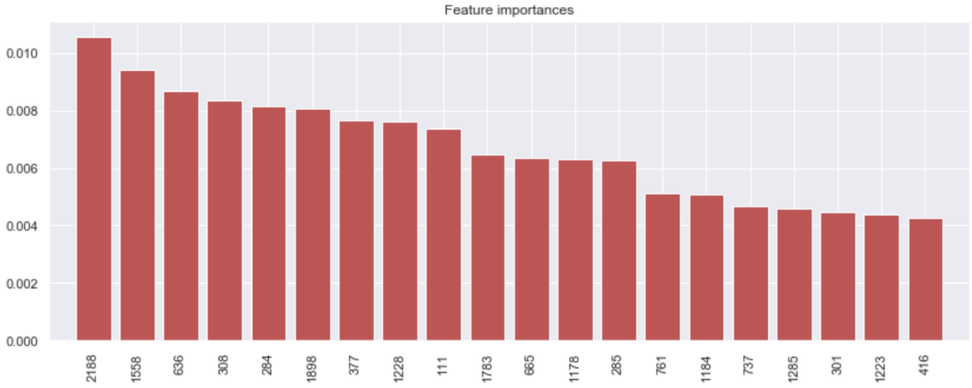
I also used a random forest classifier on the dataset. With 500 estimators, the RFC had a perfect training set score of 1.0 and a test set score of 0.84. This is worse over-fitting than the logistic regression. This could be improved through feature optimization and additional cross validation. However, the logistic regression already performs reasonably well, so it was selected as the preferred model.

Random forest was also performed on the tf-idf no-name data set without performing latent semantic analysis (LSA or truncated SVD). This was to look at what words it would use as the important words and compare them to them to the words used for clustering. Tasso, planet, and dr are the top 3 words used as features and they are also all used for clustering.

Top 20
Features in Random Forest Model



```
Feature ranking:
1. feature 2188 (0.0106) tasso
2. feature 1558 (0.0094) planet
3. feature 636 (0.0087) dr
4. feature 308 (0.0083) captain
5. feature 284 (0.0081) bunker
6. feature 1898 (0.0081) ship
7. feature 377 (0.0077) claws
8. feature 1228 (0.0076) like
9. feature 111 (0.0074) ash
10. feature 1783 (0.0065) rocket
11. feature 665 (0.0063) earth
12. feature 1178 (0.0063) klaus
13. feature 285 (0.0063) burckhardt
14. feature 761 (0.0051) eye
15. feature 1184 (0.0051) know
16. feature 737 (0.0047) exec
17. feature 1285 (0.0046) man
18. feature 301 (0.0045) calhoun
19. feature 1223 (0.0044) lieutenant
20. feature 416 (0.0043) come
```



Future Work

It would be interesting to compare these 1950's and 1960's short stories to modern science fiction short stories. Breaking them down by Parts of Speech did not help with the modeling within this dataset, but perhaps when compared to modern literature it would show a difference. Another possibility would be to compare them to literature from a different genre. Could the science fiction works be correctly classified as science fiction?

Another direction this could be taken would be to break down the works in a different way. For this analysis I used thought breaks. They could also be modeled by sentence or as entire works. By sentence there would be a lot more samples to select from, but less in each sample and so potentially harder to accurately predict the authors.

This model could also be refined by customizing Spacy to detect a few missed character names. For example, Klaus and Calhoun still appear in the important features of the random forest model. These are likely instances of missed names.

References:

Project Gutenberg: https://www.gutenberg.org/
Clustsering KMeans Centroids:
https://scikit-learn.org/stable/auto_examples/text/plot_document_clustering.html#sphx-glr-auto-examples-text-plot-document-clustering-py

Stories:

1. Dick, Philip, “Second Variety,” *Space Science Fiction,* May 1953, https://www.gutenberg.org/ebooks/32032, dick1.txt
2. Dick, Philip, “The Hanging Stranger,” *Science Fiction Adventures,* Science Fiction Adventures, December 1953, https://www.gutenberg.org/ebooks/41562, dick2.txt
3. Harmon, Jim, “The Spicy Sound of Success,” *Galaxy Magazine*, August 1959, https://www.gutenberg.org/ebooks/51351, harmon1.txt
4. Harmon, Jim, “Confidence Game,” *Galaxy Science Fiction*, June 1957, https://www.gutenberg.org/ebooks/51305, harmon2.txt
5. Harmon, Jim, “Break a Leg,” *Galaxy Science Fiction*, November 1957, https://www.gutenberg.org/ebooks/51320, harmon3.txt
6. Leiber, Fritz, “A Hitch in Space,” *Worlds of Tomorrow,* August 1963, https://www.gutenberg.org/ebooks/53042, leiber1.txt
7. Leiber, Fritz, “A Pail of Air,” *Galaxy Science Fiction,* Dec 1951, https://www.gutenberg.org/ebooks/51461, leiber2.txt
8. Leiber, Fritz, “Coming Attraction,” *Galaxy Science Fiction,* November 1950, https://www.gutenberg.org/ebooks/51082, leiber3.txt
9. Leiber, Fritz, “A Bad Day for Sales,” *Galaxy Science Fiction,* July 1953, https://www.gutenberg.org/ebooks/50819, leiber4.txt
10. Leinster, Murray, “Doctor,” *Galaxy Science Fiction*, February 1961, https://www.gutenberg.org/ebooks/51782, leinster1.txt
11. Leinster, Murray, “If You Was a MOKLIN,” *Galaxy Science Fiction*, September 1951, https://www.gutenberg.org/ebooks/51752, leinster2.txt
12. Leinster, Murray, “Med Ship Man,” *Galaxy Science Fiction,* October 1963, https://www.gutenberg.org/ebooks/50999, leinster3.txt
13. Leinster, Murray, “Third Planet,” *Worlds of Tomorrow,* April 1963, https://www.gutenberg.org/ebooks/52574, leinster4.txt
14. Ludwig, Edward, “To Save Earth,” *Worlds of Tomorrow,* October 1963, https://www.gutenberg.org/ebooks/53059, ludwig1.txt
15. Ludwig, Edward, “Spacemen Die at Home,” *Galaxy Science Fiction*, October 1951, https://www.gutenberg.org/ebooks/51249, ludwig2.txt
16. Ludwig, Edward, “The Lonely Ones,” *Galaxy Science Fiction,* https://www.gutenberg.org/ebooks/38302, July 1953, ludwig3.txt
17. Ludwig, Edward, “A Coffin for Jacob,” *Galaxy Science Fiction,* May 1956, https://www.gutenberg.org/ebooks/51203, ludwig4.txt
18. McKenty, Jack, “$1,000 A Plate,” *Galaxy Science Fiction,* https://www.gutenberg.org/ebooks/50921, October 1954, mckenty1.txt

19. McKenty, Jack, “Wait for Weight,” *Galaxy Science Fiction,* October 1952, https://www.gutenberg.org/ebooks/32717, , mckenty2.txt
20. Pohl, Frederik, “The Tunnel Under the World,” *Galaxy Science Fiction,* January 1955, https://www.gutenberg.org/ebooks/31979, pohl1.txt
21. Pohl, Frederik, “Plague of Pythons,” *Galaxy Science Fiction,* https://www.gutenberg.org/ebooks/51804, October and December 1962, pohl2.txt
22. Pohl, Frederik, “Survival Kit,” *Galaxy Science Fiction,* https://www.gutenberg.org/ebooks/51809, May 1957, pohl3.txt
23. Quattrocchi, Frank, “Sea Legs,” *Galaxy Science Fiction,* https://www.gutenberg.org/ebooks/51407, November 1951, quattrochi1.txt
24. Quattrocchi, Frank, “The Sword,” *IF Worlds of Science Fiction,* March 1953, https://www.gutenberg.org/ebooks/32697, quattrochi2.txt
25. Reynolds, Mack, “Spaceman on a Spree,” *Worlds of Tomorrow,* June 1963, https://www.gutenberg.org/ebooks/52995, reynolds1.txt
26. Reynolds, Mack, “*Potential Enemy,”Orbit volume 1 number 2*, https://www.gutenberg.org/ebooks/40954, 1953, reynolds2.txt
27. Reynolds, Mack, “Farmer,” *Galaxy Science Fiction,* June 1961, https://www.gutenberg.org/ebooks/51799, reynolds3.txt
28. Reynolds, Mack, “Unborn Tomorrow,” *Astounding Science Fiction,* https://www.gutenberg.org/ebooks/23942, June 1959, reynods4.txt
29. Stecher, L.J., Jr., “When You Giffle,” *Worlds of Tomorrow,* December 1963, https://www.gutenberg.org/ebooks/53035, stecher1.txt
30. Stecher, L.J., Jr., “An Elephant for the Prinkip,” *Galaxy Science Fiction,* August 1960, https://www.gutenberg.org/ebooks/51434, stecher2.txt
31. Stecher, L.J., Jr., “Man in a Sewing Machine,” *Galaxy Science Fiction,* February 1956, https://www.gutenberg.org/ebooks/50936, stecher3.txt
32. Stecher, L.J., Jr., “Perfect Answer,” *Galaxy Science Fiction,* June 1958, https://www.gutenberg.org/ebooks/51482, stecher4.txt