Book Classification

Unsupervised Machine Learning

Evan Hintz

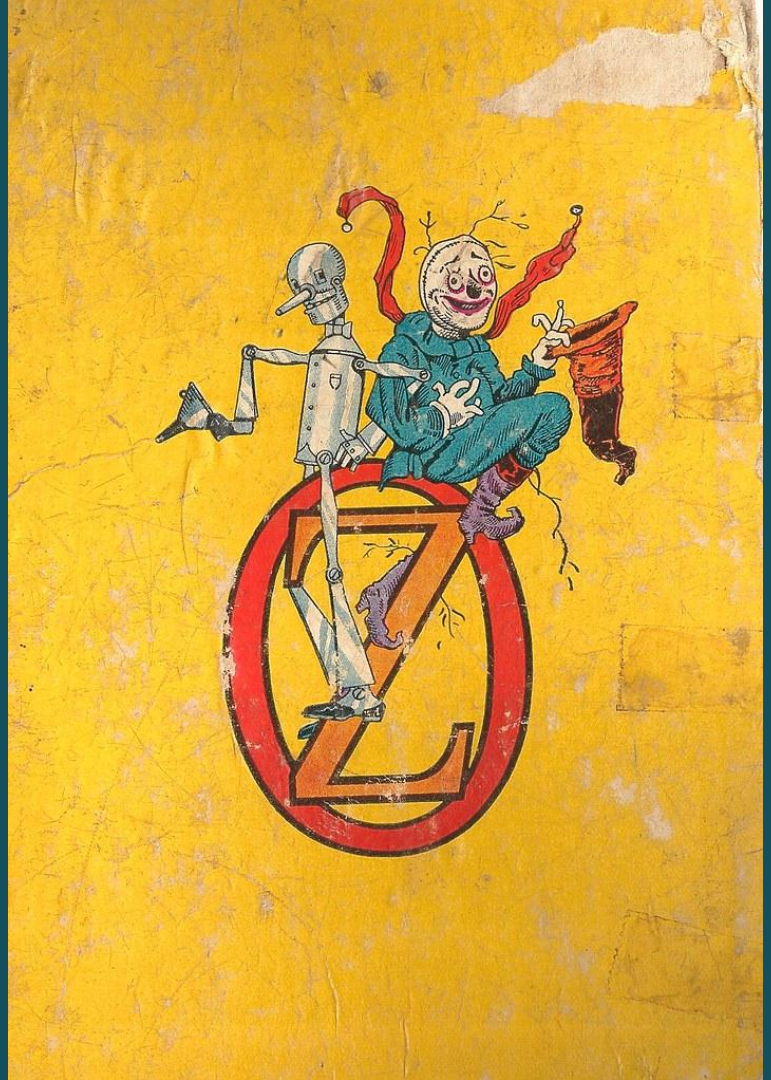
Contents:

Motivation

Genres are a great way to classify books with themes we can understand. But is there a better way? This project unfortunately can't answer that (yet) but is a look into how an unsupervised machine can be used to find patterns and classify books in an attempt to improve future recommendation systems.

All data is open-sourced from Project Gutenberg

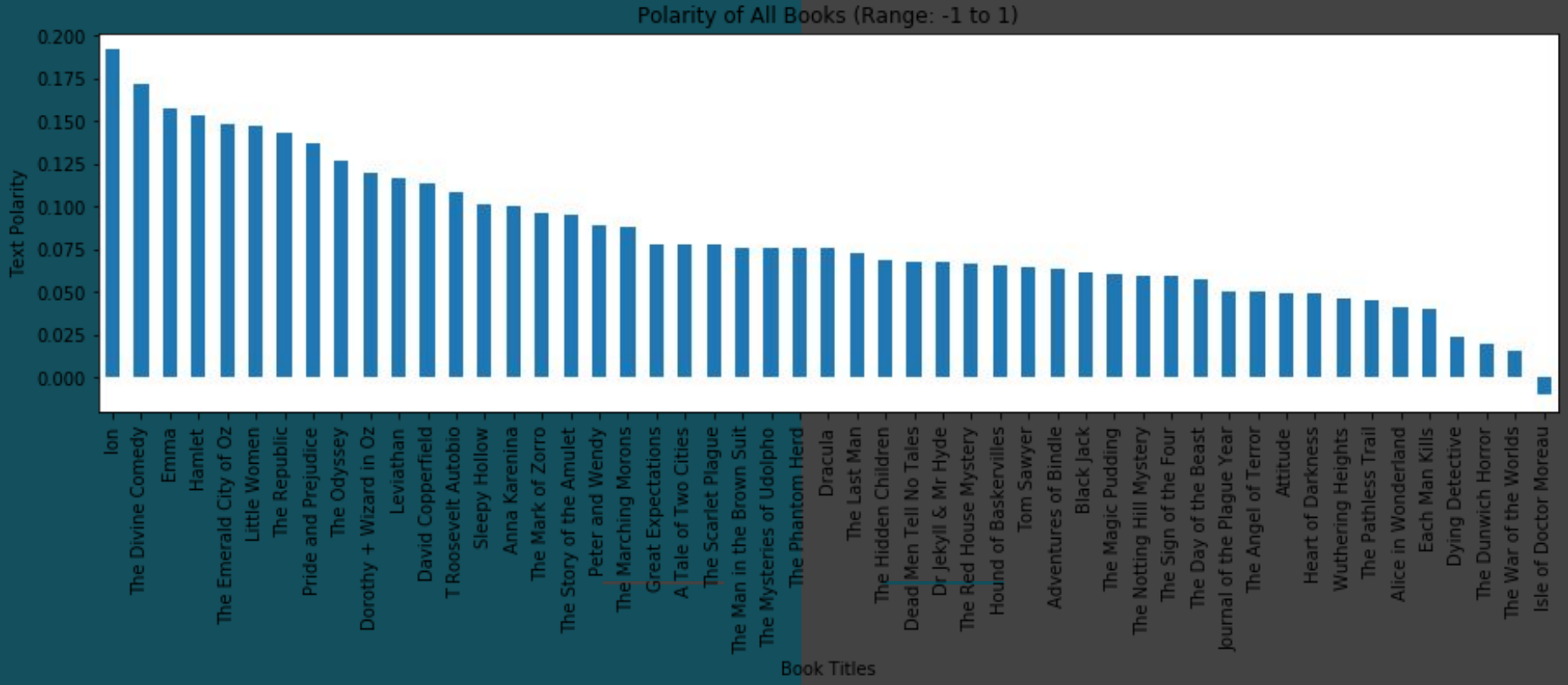
http://www.gutenberg.org/



Sentiment Analysis

Polarity

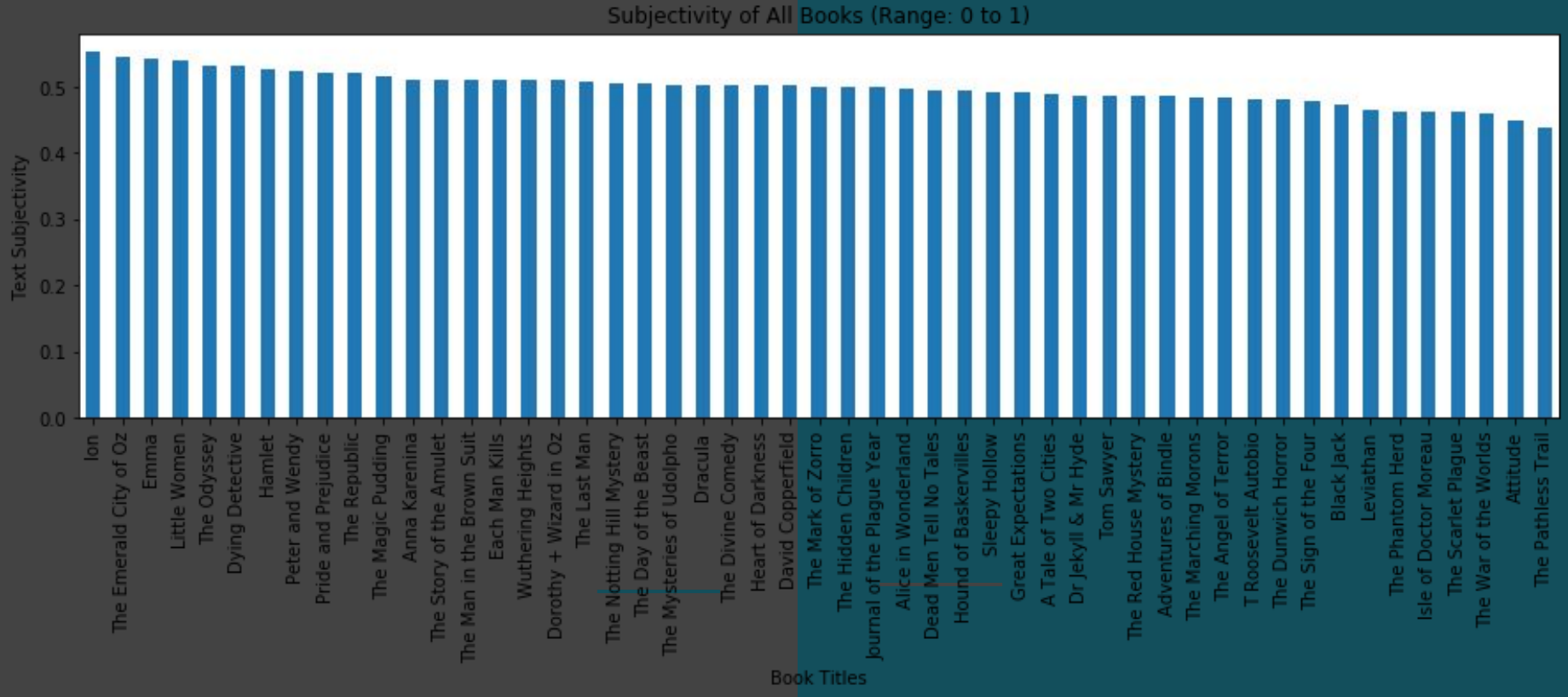
Text polarity attempts classify the emotions expressed within the text as generally positive, neutral, or negative.



Sentiment Analysis
Subjectivity

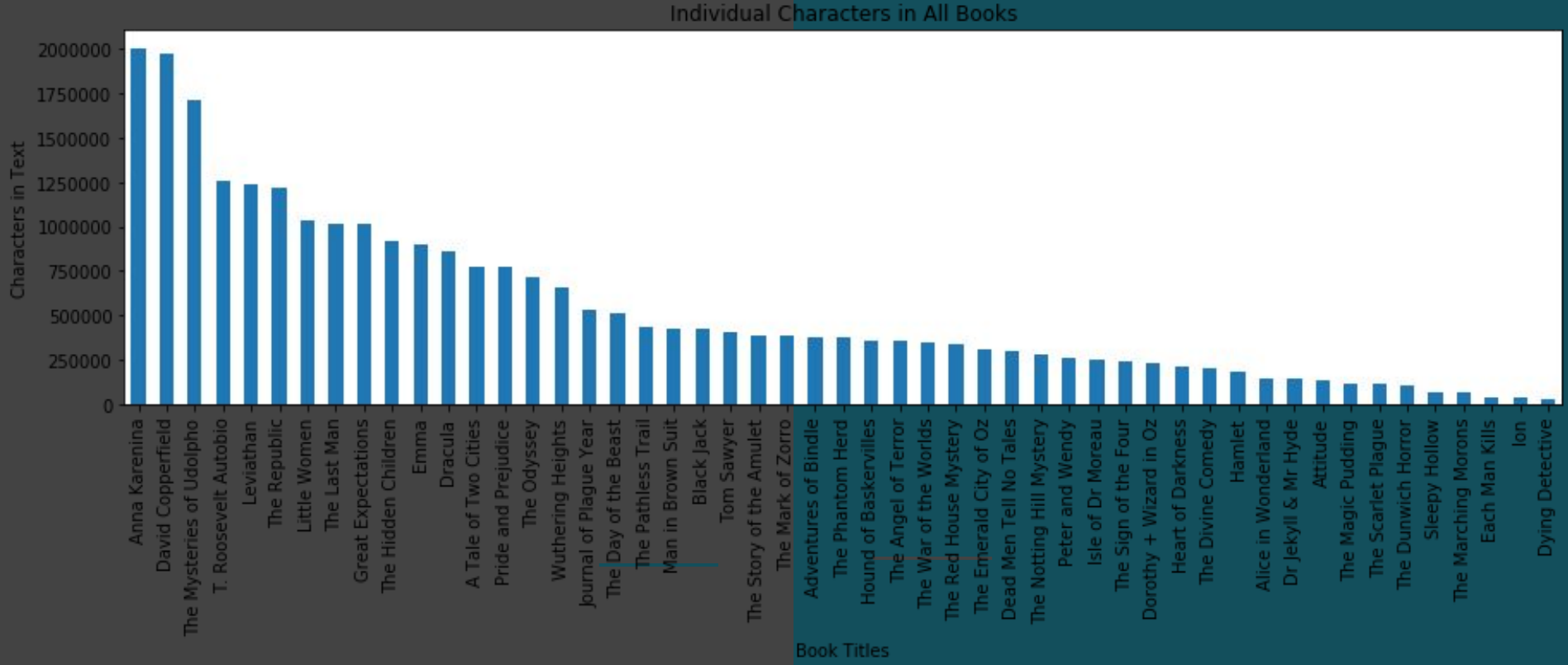
Text subjectivity describes a text based on its use of factual statements (objective) or emotional/opinionated (subjective).

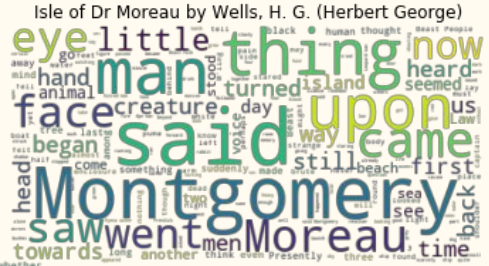
Scale: 0 - 1; 1 being the most subjective.

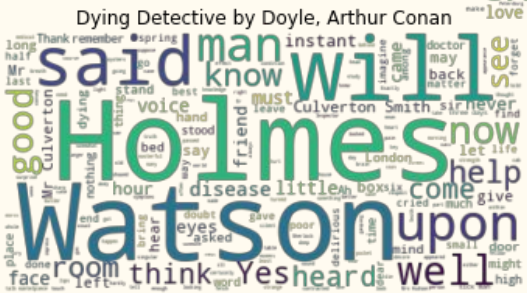


Book Length

This plot shows the amount of characters in each book. As you can see there is quite the range of book size.



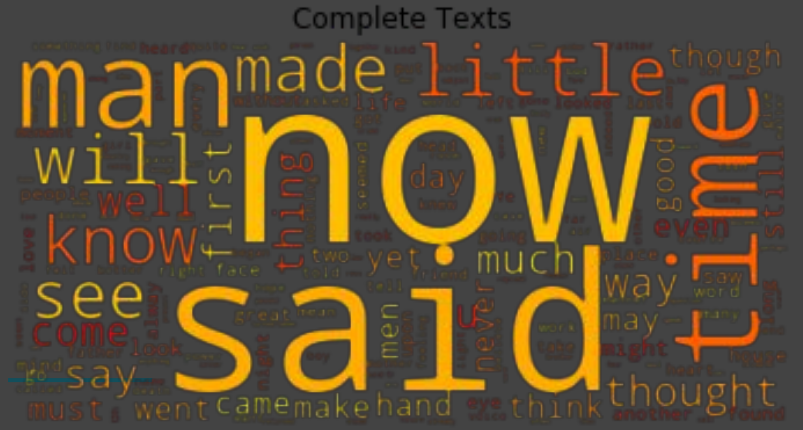




Word Clouds

An easy visual way to see what words pop up the most in books.

Notice the generic words in the complete texts and the specific names in the individual books.



Text Processing

This is where the majority of time was spent; both programming and computing.

All books underwent the following:

- Punctuation removal
- Tokenization; each individual word is separated into its own string
- Stop-word removal; common words that won't aid in classifying the book
- Part of speech tagging; each word's PoS is added; verb, noun, etc.
- Lemmatization; uses PoS tags to reduce words to their root
- Vectorization; transforms texts into numerical values that can be analyzed

Machine Learning - Kmeans

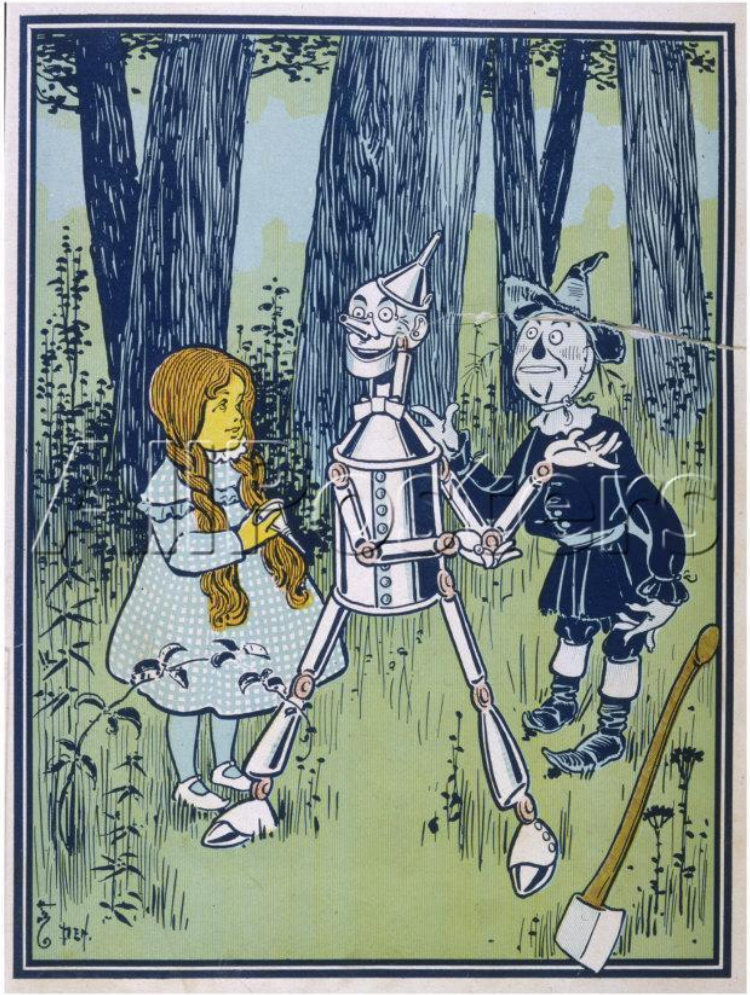
Using the elbow and silhouette methods of selecting clusters, the optimal range is between 3 and 6.

```
Cluster 0:
 peter
 wendy
 mcgregor
 mr
 say
Cluster 1:
 one
 say
 make
 would
 come
Cluster 2:
 dorothy
 scarecrow
 oz
 say
 wizard
Cluster 3:
 señor
 zorro
 diego
 señorita
 gonzales
Cluster 4:
 say
 go
 one
 mr
 come
```

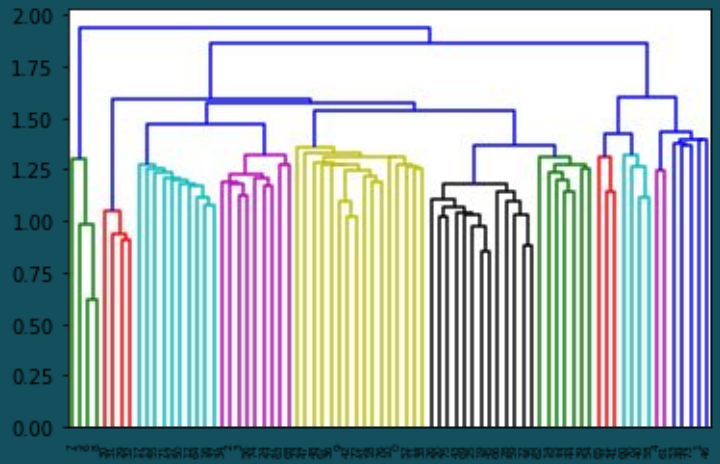
Let's look at the top words in each cluster (n=5) using the out of box model and compare with the prediction placement for the test books.

```
Prediction
4 Sense and Sensibility
2 Ozma of Oz
4 Cabin Fever
4 Jane Eyre: An Autobiography
4 The Mysterious Affair at Styles
4 The Nigger of The "Narcissus"
1 A Journal of the Plague Year
4 David Copperfield
4 The Adventures of Sherlock Holmes
```

Clearly #2 is Cluster of Oz, #3 is very Spain oriented, and the remaining clusters are fairly generic with #4 taking over most of the predictions.



Machine Learning - Hierarchical



Looking at a dendrogram it looks like there are 7 main clusters that are defined.

```
Training Labels:
6 Little Women
2 The Piebald Hippogriff
0 Pride and Prejudice
0 Emma
2 Peter Pan
3 Dorothy and the Wizard in Oz
3 The Wonderful Wizard of Oz
3 The Marvelous Land of Oz
3 The Emerald City of Oz
6 The Phantom Herd
```

```
Predicted Labels:
1 Sense and Sensibility
5 Ozma of Oz
0 Cabin Fever
1 Jane Eyre: An Autobiography
1 The Mysterious Affair at Styles
4 The Nigger of The "Narcissus"
4 A Journal of the Plague Year
1 David Copperfield
1 The Adventures of Sherlock Holmes
4 Each Man Kills
0 The Mysterious Rider
3 The Odyssey
4 The Legend of Sleepy Hollow
4 The Jungle Book
0 The Call of the Wild
```

Comparing slices of training and predicted labels it doesn't seem like anything is lining up! From a glance it doesn't look like this model will do very well.

Non-negative Matrix Factorization

This model won't be used to predict labels but as a topic clusterer - it expresses each book as a combination of created topics/themes.

Looking at a few models with various parameters we can see the top words for each topic. Some are more recognizable than others; specifically Oz and Shakespeare themes.

```
Out of Box Model:
Topic #0: one go come say could would get make time see
Topic #1: one hath thou make may say socrates thy would man
Topic #2: dorothy oz scarecrow say wizard ozma woodman zeb nome emerald
Topic #3: mr say would go could come look one know emma
Topic #4: holmes watson say upon man mcmurdo one would sir well
Topic #5: say go dont get see one come well know look
```

```
Frobenius Norm Model:
Topic #0: say one go come would make could see time get
Topic #1: hamlet horatio polonius laertes rosencrantz ophelia guildenstern lord king marcellus
Topic #2: dorothy oz scarecrow wizard ozma woodman zeb nome toto emerald
Topic #3: bindle mr wot earty hearty stiffson millie ai gupperduck macfie
Topic #4: holmes watson upon baskerville culverton man mcmurdo sir stapleton barrymore
Topic #5: puddin bunyip bill possum bluegum say thieves sam wombat owners
```

```
Kullback-Leibler Divergence Model:
Topic #0: young youth way window wife wink whole whose whether ask
Topic #1: dissolve austerity fourths supplant prudence wood frequent capability besmear favorable
Topic #2: cloth bravely chasm cheer branch butcher chariot blew bread prisoner
Topic #3: bill harm chose uncertainty hitherto farthing agree unfolded within assume
Topic #4: beards deep brow dearest civilize bishareen come wire borderer axe
Topic #5: babylon carrot olympus world broaden fillip comin aster woodland worm
```

Evaluation

I decided to create my own accuracy function to put real values on each model's ability.

Evaluating unsupervised models is tough - there are no specific metrics to look for. It's easiest to work with data you're familiar with and compare cluster labels to predicted labels.

The estimated accuracy is based on the subjects of the books. Here's an example from Project Gutenberg; Alice's Adventures in Wonderland.

```
({'Alice (Fictitious character from Carroll) -- Juvenile fiction',
  "Children's stories",
  'Fantasy fiction',
  'Imaginary places -- Juvenile fiction',
  'PR',
  'PZ'})
```

If this book is predicted as cluster 3 the function gathers all words from training books in cluster 3 and calculates how many are in the test book's subject.

The output is a percentage which is the amount of words in the test book's subject also found in the trained cluster's subjects divided by total words in the test book's subject.

Evaluation

```
Hierarchical Model

       3 clusters
Average Accuracy: 15.39%

       4 clusters
Average Accuracy: 16.21%

       5 clusters
Average Accuracy: 26.13%

       6 clusters
Average Accuracy: 27.17%

       7 clusters
Average Accuracy: 24.13%

       8 clusters
Average Accuracy: 26.30%
```

```
KMeans Model

       3 clusters
Average Accuracy: 82.07%

       4 clusters
Average Accuracy: 81.86%

       5 clusters
Average Accuracy: 81.65%

       6 clusters
Average Accuracy: 81.45%

       7 clusters
Average Accuracy: 81.45%

       8 clusters
Average Accuracy: 70.66%
```

Looking at some average estimated accuracies from the two models with a range of clusters it's clear the KMeans model wins out every time.

While it looks like 3 clusters is the best, it's important to keep in mind that with this method of accuracy estimation, when there are fewer clusters it's easier to have a higher score for each book.

"The Monkeys caught Dorothy in their arms and flew away with her."

Conclusion

KMeans does a relatively good job of clustering and predicting. Here are some specific estimated accuracies within the test books using the optimal number of clusters. "The Call of the Wild" is surprisingly low, even lower than one of the few philosophy books included of Ion by Plato.

-	Ozma of Oz - L. Frank Baum	90.00%
-	The Call of the Wild - Jack London	37.50%
-	Sherlock Holmes - Arthur Doyle	92.31%
-	Sense & Sensibility - Jane Austen	85.71%
-	The Odyssey - Homer	75.00%
-	The Mysterious Rider - Zane Grey	100%
-	Ion - Plato	44.44%

Conclusion

While this model clearly struggles on some genres like philosophy and some adventure - it does very well with most children's/juvenile and mystery books. This can most likely be contributed to the small and narrow selection of books.

The methods used here show great promise in the ability of unsupervised learning that may soon be successfully applied to a book recommendation system.

Future Work

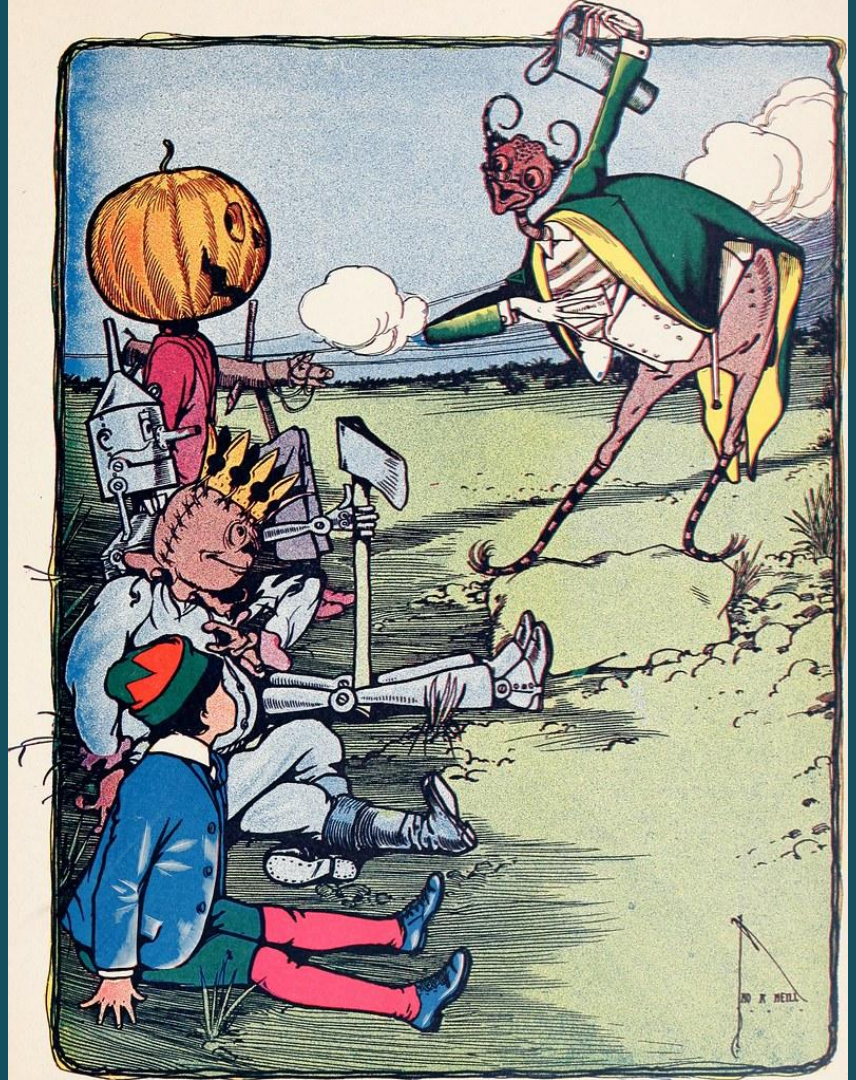
Of course the first thing to be improved on is the range and size of the catalog of books used. In all cases, more training data leads to a better model. Other areas that may provide improvements and should be explored include:

- N-gram modeling (word phrases)
- Latent Dirichlet Allocation
- Dimensionality Reduction
- Improved subject descriptions for a more precise accuracy estimator

Thank You!

I appreciate you taking the time to look through my project. I hope to improve and add to it soon.

If you have any questions, comments, or concerns, please feel free to contact me at evan.j.hintz@gmail.com



Illustrations taken from original Wizard of Oz and Alice's Adventures in Wonderland books.

"She caught Toto by the ear."