**Spooky Authors:  
Predictive Analysis of Natural Language**

**Team Members**

Kali Blevins (Team Lead)

Anais Dotis-Georgiou (Tester)

Danny Lu (Data Engineer)

Bharathy Natarajan (Scrum Master)

**Overview**

We used many of the data analysis tools we have learned so far, plus some additional help from the Natural Language Toolkit library, to identify a writer’s writing style and used that information to assign authorship to a unlabeled dataset and then quality check those assignments.

**Data Sources & Cleaning**

Data Sources

* Uses Kaggle dataset for the Spooky Author Competition (<https://www.kaggle.com/c/spooky-author-identification>)

Data Cleaning

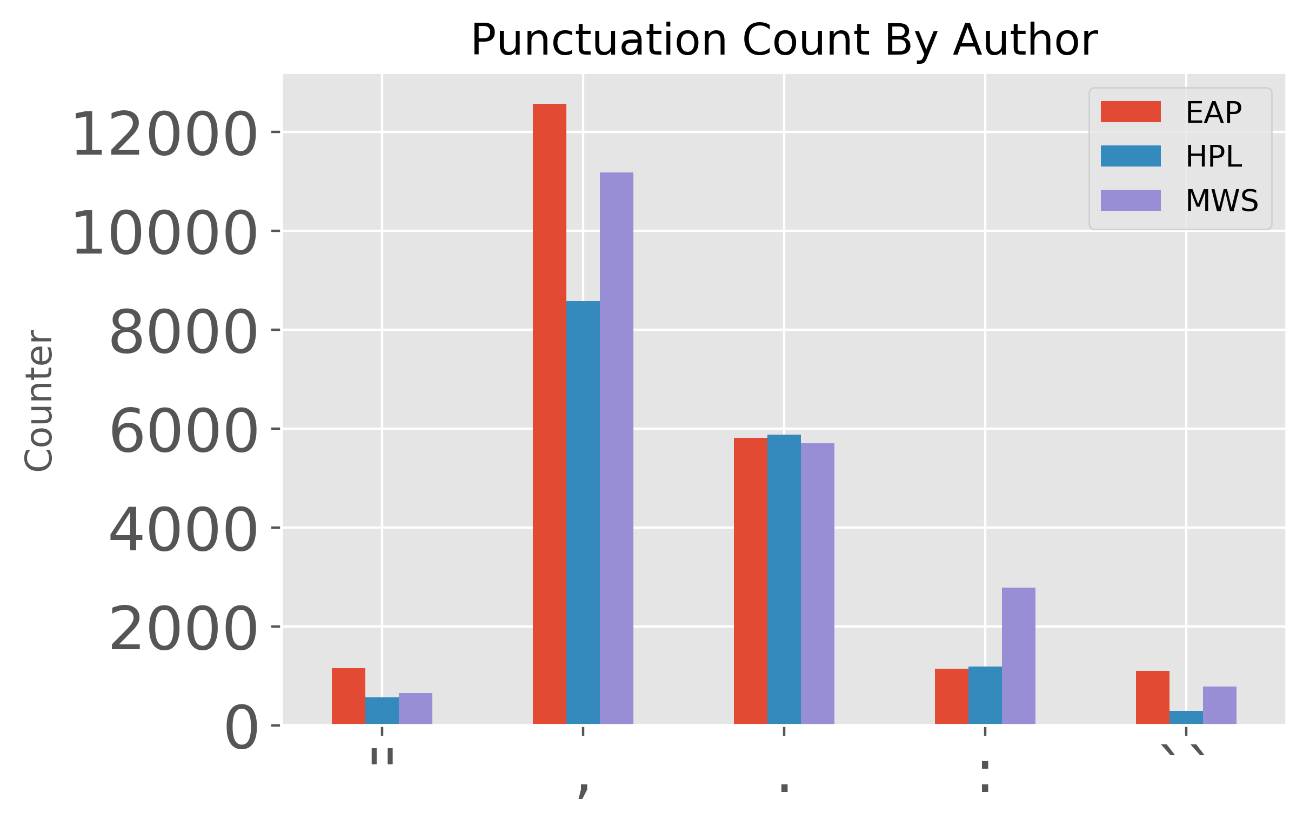
* Group dataset by author
* Create equal length datasets for each author in the training data
  + Count length of each author dataset
  + Randomly sample the datasets for the authors with longer datasets to subset for equal lengths
* Remove outliers for really long passages

**How do the authors vary with respect to sentence structure & vocabulary?**

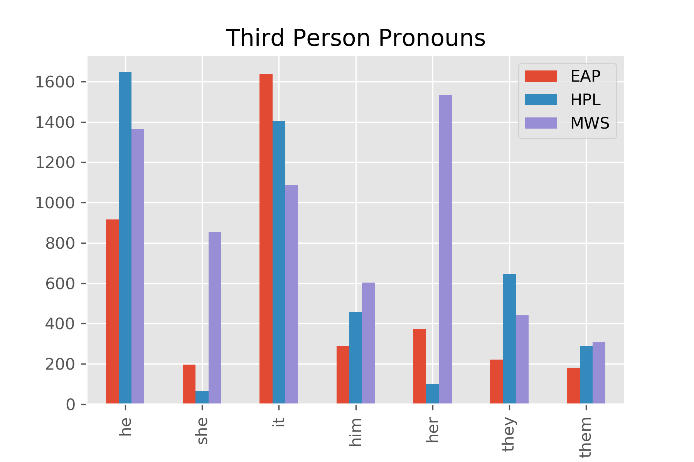
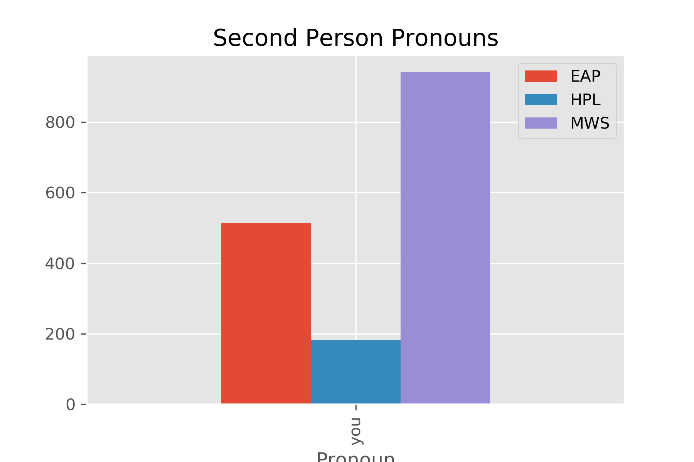
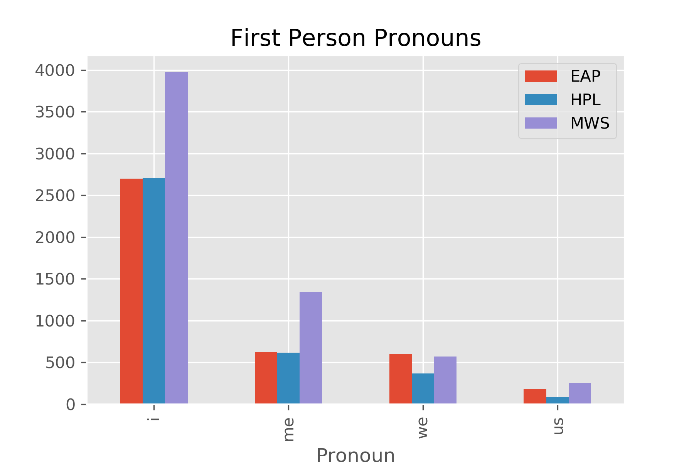
Based on our analysis, the best identifiers of authorship are word length and sentiment scores, specifically positive and neutral (see table with ANOVA results).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **sent\_len** | **word\_var** | **word\_len** | **compound** | **pos** | **neu** | **neg** |
| F | 30.11198 | 58.58519 | 235.5922 | 111.2549 | 384.527 | 445.7831 | 91.9655 |
| p | 8.83E-14 | 4.41E-26 | 5.38E-103 | 9.95E-49 | 4.91E-164 | 2.16E-189 | 1.89E-40 |

Special character usage in the stories can tell you about the authorship of the text. Looking at the graph, we clearly see that Edgar Allan Poe is fond of quotations and commas, whereas Mary Shelley enjoys using colons more than the other two.



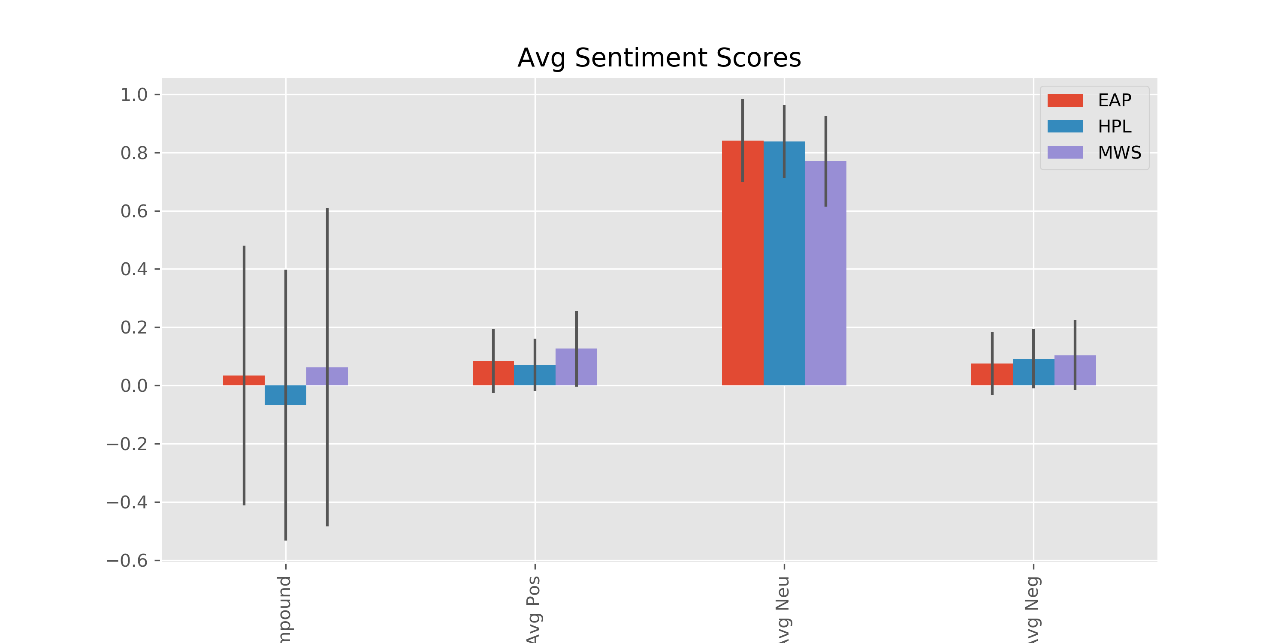
Pronoun usage is also a good indicator to identify the authorship of the text. Some author prefers to use first person and second person pronouns. Here we see MWS is more likely to use the first person and second person pronouns in her work compared to other two authors.



Third person pronouns usage could be used to identify author’s favor on male/female main characters in their work.

**What are the sentiment scores for each author?**

Sentiment scores allow us to determine the mood of the story the author wants to deliver to their reader. From this graph, we see that MWS has the highest compound score when compared to other two authors, even though she likes to write many positive and negative sentences. HPL, on another hand, has the lowest compound score. This shows us unique writing style of each author.



**How do we assign authorship to sentences?**

1. Create dictionary of word frequencies for each author for training dataset
2. For each sentence in the test dataset:
   1. create a table with the frequency with which each author uses each word
   2. For each author, multiply together all the frequencies for the words in the sentence
   3. Identify which author has the highest number – that’s the winner!

We tested this method using 90% of the training dataset as the training dataset and 10% of the training dataset as the test dataset. The overall accuracy was 85% and the predictions for individual authors is summarized in the table below, the correct author is on the left and the predicted author is across the top.

|  |  |  |  |
| --- | --- | --- | --- |
|  | EAP | HPL | MWS |
| EAP | 80% | 8% | 12% |
| HPL | 8% | 88% | 5% |
| MWS | 7% | 3% | 90% |

**How does our assigned authorship compare to data with author already assigned?**

By comparing plots of the same analyses for both the training & test datasets after the author has been assigned, we can see that our author assignment seems to have done fairly well! In the plots below, the training data can be seen on the left and the test data is found on the right.

