Book Classification using Text of Book

Aaron Kohn

Business Problem

Using the text of books can they be accurately categorized? As a start in this endeavor this project focuses on two predictions. Using samples of text selected from books 1. can a model be created to identify the book title? 2. Can the model predict the author when there are multiple books by the same author?

Background/History

Book recommendations are often made based on external data. Such features may include data collected regarding others that have read and liked the book. These recommendations are made using the assumption that those with similar likes will like the same books. While the assumption has merit, there may be some flaws as well.

The method proposed here is to categorize books based on the text. Using the text of the book to extract features to categorize books should be a more direct approach to book recommendations. To show the effectiveness of categorizing books by text, books will be categorized based on available labels. In this study the labels available are book title and author. Based on samples of text selected from books predictions will be made to identify the book title of the sample. Further attempts will be made to identify the author of a book, based on other works of the same author.

Data Explanation (Data Prep/Data Dictionary/etc)

The texts of the books have been collected from the Project Gutenberg website. Random file numbers were collected and downloaded from the website. Steps were taken to prevent duplicate files. Similarly books with anonymous authors or unnamed authors were not included for this project. A dataset of two thousand books was created. The columns of the data set are the book name, author, URL, length, and ten samples selected from the text. The book name and author are extracted from the title of the book on the website and are used as the target values for this project. The URL and length columns are used for data preparation purposes and not included in the model.

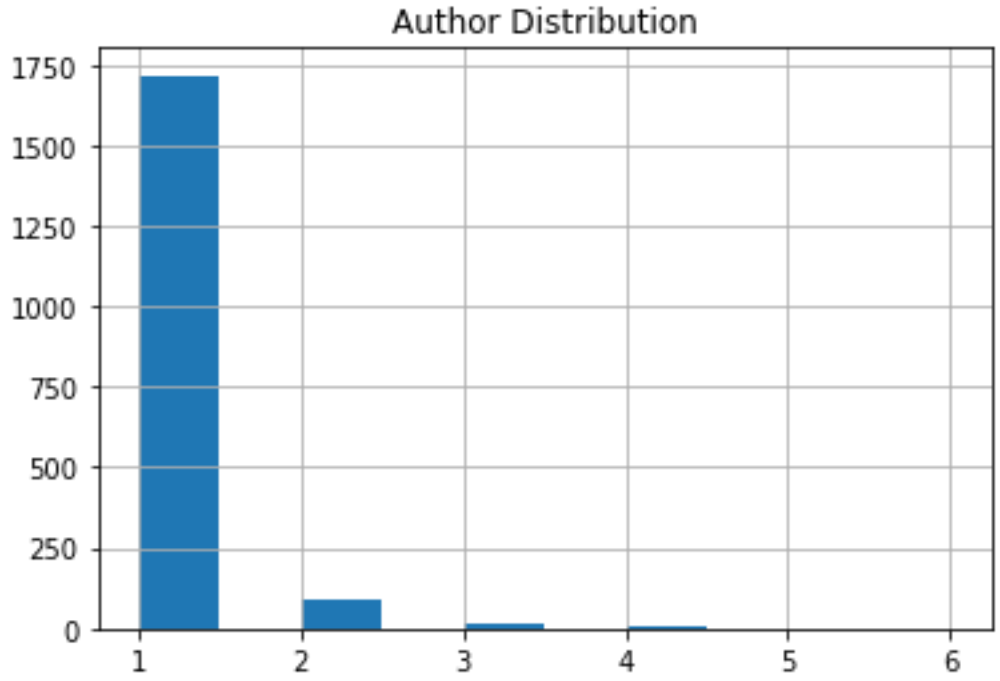
The sample texts which are used as the basis for the features in the model were created selecting random five-hundred-word samples of the text. An attempt was made to exclude text added by the website to the actual books. The results were not fully effective and requires more work.

The dataset was transformed into long format with each row containing one sample. The sample texts were tagged, tokenized, and stemmed. Stop words were removed. As a final step in data preparation tfidf vectorization was applied to the stemmed samples. Vectorization was done for unigrams and for n-grams with in the one to five range. Models were attempted for both.

Methods

The data selected while random certain steps were taken to ensure that all variables will be present. Therefore, there was no issues regarding missing data. There may be some concern regarding books sharing names. There are 11 names that appear more than once. The name “Poems” appears five times and the name “the islands and their peoples” three. Nine other names appear twice. More than one hundred authors appear more than once this should provide some information on the effectiveness

of predicting the author across multiple works.



Graph of author occurrences in data set. While most authors appear just once there are many that have multiple appearances.

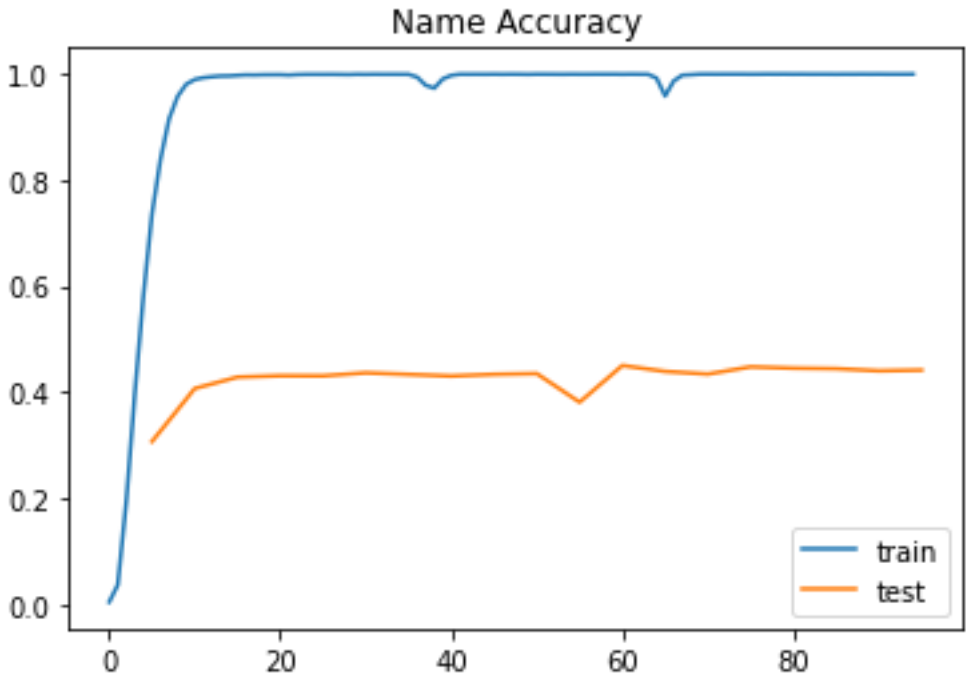
In the initial phase of the project a random train test split was performed on the data. The split randomly selected sample rows to set aside for validation. In the final steps of the model testing, testing will be done to determine the model's effectiveness on predicting the author of books not present in the training set at all.

The Keras classifier wrapper was used for a basic deep neural network with two hidden layers. The model is tested on unigram features and n-grams within a range of 1-5. Multiple max feature sizes are tried to determine the most effective size. The model is trained using a series of epochs to guard against overfitting.

Analysis

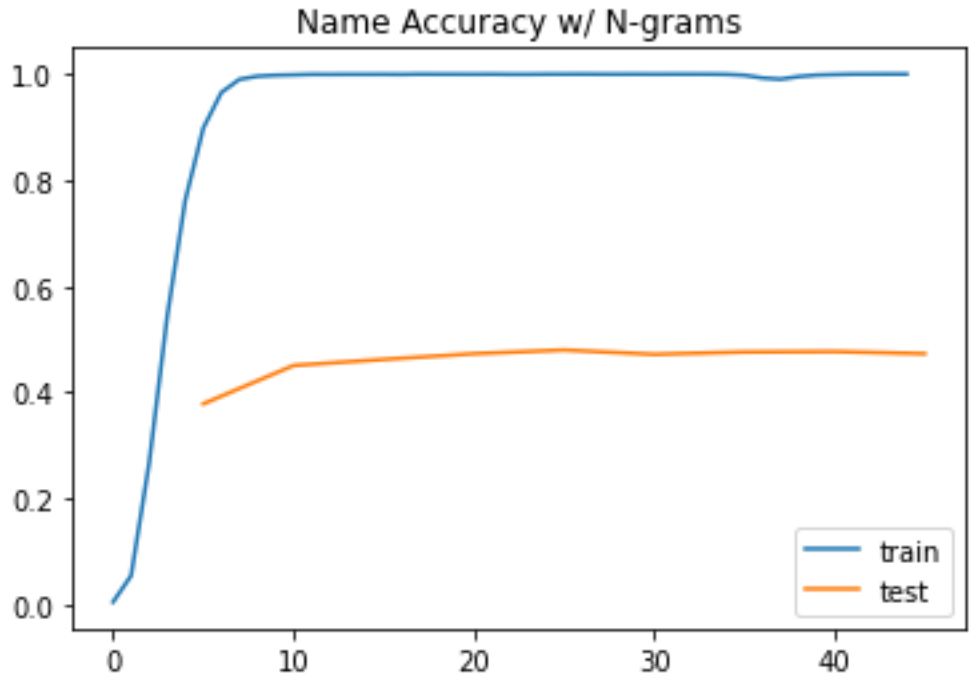
All models appear to converge after only a few epochs. The models reach ninety-nine percent accuracy in less than twelve epochs. However, on the validation set accuracies top out at about fifty percent. While these numbers are not as high as might be hoped for the number of categories being predicted, 1836 authors and 1985 book names, it is not that low.

The base model for book names was allowed to run up to 95 epochs. The model seems to top



Accuracy of model for the train and test data. Model created to predict book name using top 5000 unigram features.

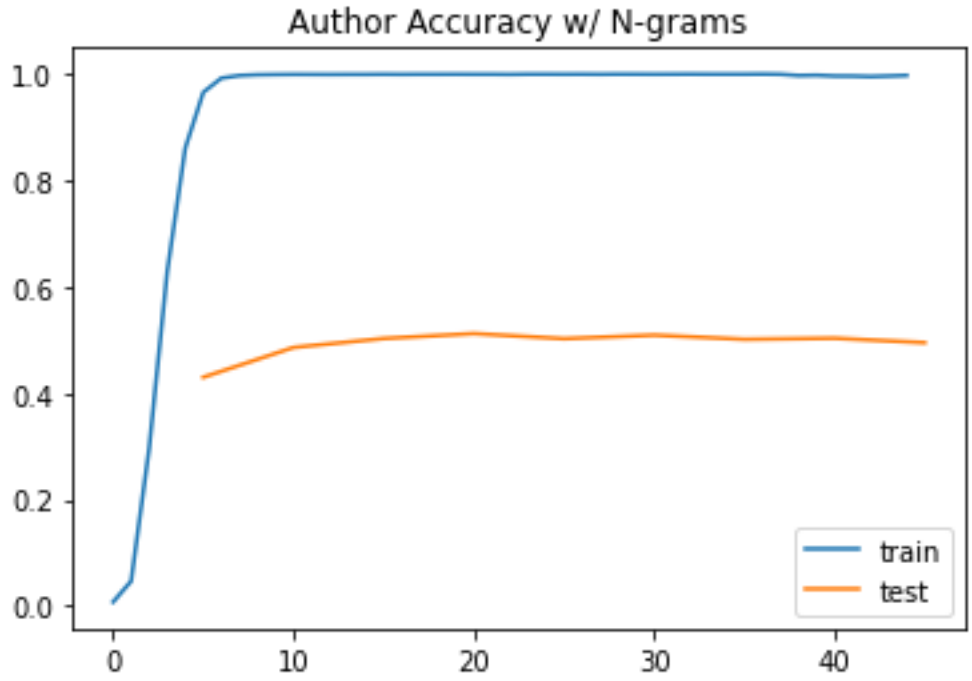
out for the test data at approximately thirty epochs, with an accuracy of 0.4363. After that the accuracy of the becomes unstable although there is a slight improvement in later epochs. Introducing multiple n-gram levels improves the performance of the model. Using the top 15000 features with up 5-gram sets achieves forty-eight percent accuracy at twenty-five epochs.



Accuracy of model for the train and test data. Model created to predict book name using top 15000 multigram features.

When training the model to predict authors using 15000 multi-gram features as was done for book name. The model achieves an accuracy of fifty-one percent at about twenty epochs. This was the

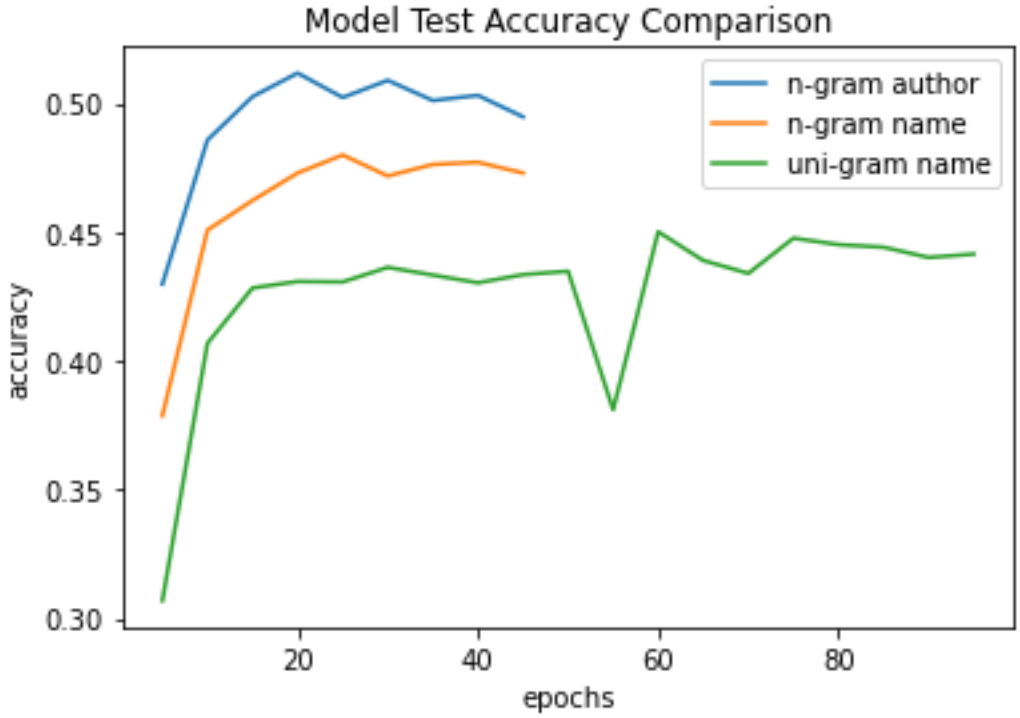
best performance of the three models.



Accuracy of model by epoch for the train and test data. Model created to predict book authors using top 15000 multigram features

Conclusion

There is some predictive power to these models. While the results are not at optimal levels the models still do have the ability to predict at some level. There is a benefit to using the n-gram features.



Accuracy results for all three models. The n-gram models were run for up to 45 epochs. The unigram model was run for 95.

It would be necessary to test if adding features over the 15000 would further improve the model results.

Assumptions

As described above the data was collected from the Project Gutenberg website. There were some challenges when collecting the data as not all book formats are uniform. Some books may include added text that is not part of the original work. There may be some instances where the same text may appear in multiple books as a result. Similarly, the method used for selecting samples of text may have led to some overlap.

Limitations

The model was trained using the books made available by the Gutenberg Project. These works are limited to those that have had their copyright rights expired. In general, this limits the books of a certain age. For a more relevant model a wider range of books should be included.

Future Uses/Additional Applications

As pointed out earlier the large number of categories that are attempted in the prediction makes achieving a high level of accuracy challenging. Categories such as genre may provide a more realistic target variable. Further models may be created to predict similarity for books liked by a specific reader. Features deemed predictive would be given higher weights when scoring similarity.

Ethical Assessment

There does not appear to be much ethical concern in the study itself. Using such a model to attribute the work to an author may lead to withholding credit from the correct author. One area where there may some concern is if a model would be created to predict age appropriateness. Relying on such a model may result in giving literature unsuited to those of a younger age.

References

Data retrieved from https://www.gutenberg.org/

Badreesh Shetty, (November 24, 2008). Natural Language Processing (NLP) for Machine Learning. Retrieved from Natural Language Processing (NLP) for Machine Learning | by Badreesh Shetty | Towards Data Science

Questions

1. Are there specific authors harder to predict than others?
2. Are author prediction and book prediction really different or different names for the same target?
3. Other than tfidf vectorization were other methods attempted?
4. Why was tfidf vectorization used?
5. Can specific features be identified?
6. What are the benefits of using the model if accuracy rates are low?
7. Why were names and authors selected as target variables?
8. What are some categories that would be ideal for prediction?
9. Why does the model perform better for authors than names?
10. What steps would you suggest next?