Aaron Kohn

Topic

Book classification using the actual text of the book.

Business Problem

The goal of this project is to create a model that can learn to identify writing patterns and correctly classify books. Can a segment of text be identified correctly as to which book it belongs to? Can we recognize authors across different books? Are there other patterns that can be identified that can help categorize the books?

Datasets

The book texts will be randomly selected from Project Gutenberg. Multiple segments of text will be selected from each book. Target labels such as book name, author and year written will be extracted from the information in the book. An effort will be made to find multiple works from the same authors in order to predict authors.

Methods

After selecting the data and identifying target goals creating predictive features will be necessary. Tokenizing and cleaning the text will be a large part of the data wrangling portion of this project. The plan is to use n-grams of multiple lengths to try to find writing style patterns.

Once usable features have been created, machine learning models, such as neural networks will be used. Accuracy of predictions will be the primary measure of model success.

Further analysis may include unsupervised learning. K-means clustering, and other grouping methods may be applied to see if anything more can be learned from the data.

Ethical Considerations

There are no evident ethical concerns regarding the study itself. A note of caution is due if there is any attempt to implement the models. Whenever identification of any sort is automated there is some concern regarding error. Using the model to give credit to an author may result in wrongfully withholding credit from the rightful owner.

Challenges/Issues

One issue already encountered is the lack of easily available target labels. One use that such models could be applied to is predicting age appropriateness of books. Unfortunately, lists of titles with suitable ages and full book contents are not readily available. The targets chosen so far are limiting.

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