Classifying Fiction and Non-Fiction Works Using Machine Learning

Rachna Gupta[1]

[1]Illinois Mathematics and Science Academy (Aurora, Illinois)

Abstract

The objective of this project was to create a program that can determine whether an unknown text is a work of fiction or non-fiction using machine learning. Various datasets of speeches, ebooks, poems, scientific papers, and texts from Project Gutenberg and the Wolfram Example Data were utilized to train and test a Markov Chain machine learning model. A microsite was deployed with the final product that returns a probability of fictionality based on input from the user with 95% accuracy.

Introduction

- Fiction is defined as the class of literature comprising works of imaginative narration, while non-fiction is defined as all content discussing real events and facts
- Although most of the time the class of a work is obvious due to fantastical elements and writing style, the class can be unclear because of the various types of fiction and non-fiction and the complexities of the boundaries between them.
- Because of the impossibility of verifying the content of the works, machine learning must be used to identify patterns and classify based on those patterns
- Computational linguistics can be used to identify patterns in literature
- **GOAL: To create a classifier that can accurately understand those boundaries in order to classify an unknown work**

Methods/Results

Obtaining and Processing Data

- Imported 300 full .txt files of various speeches, ebooks, poems, essays, scientific papers, and more texts from the public dataset for Project Gutenberg
- Manually assigned classes to the data and sorted into Fiction and Non-Fiction based on research on the works
- Made a dataset containing the name, class, and full text of each file
- Partitioned each text into sections of 5000 characters each and associate the sections with their corresponding class

FileName	Nonfiction	FullText
Aldous Huxley___Crome Yellow.txt	False	CROME YELLOW ⋱
Aldous Huxley___Mortal Coils.txt	False	MORTAL COILS ⋱
Aldous Huxley___The Defeat of Youth and Other Poems.txt	False	THE DEFEAT OF YOUTH AND ⋱
Alexander Pope___An Essay on Criticism.txt	True	AN ESSAY ON CRITICISM. ⋱
Alexander Pope___Essay on Man.txt	True	AN ESSAY ON MAN. ⋱
Alexander Pope___The Poetical Works of Alexander Pope, Volume 1.txt	False	THE POETICAL WORKS OF ALE
Alexander Pope___The Poetical Works of Alexander Pope, Volume 2.txt	False	THE ⋱
Alexander Pope___The Rape of the Lock and Other Poems.txt	False	THE RAPE OF THE LOCK ⋱
Alexander Pope___The Works of Alexander Pope, Volume 1.txt	False	THE WORKS ⋱
Alfred Russel Wallace___Contributions to the Theory of Natural Selection.txt	True	_CONTRIBUTIONS TO_ THE THE
Alfred Russel Wallace___Darwinism.txt	True	DARWINISM ⋱
Alfred Russel Wallace___Island Life.txt	True	FRONTISPIECE ⋱
Alfred Russel Wallace___The Malay Archipelago, Volume 1.txt	True	THE MALAY ARCHIPELAGO, VOL
Alfred Russel Wallace___The Malay Archipelago, Volume 2.txt	True	THE MALAY ARCHIPELAGO ⋱
Ambrose Bierce___A Cynic Looks at Life.txt	True	LITTLE BLUE BOOK NO. 1099 ⋱
Ambrose Bierce___An Occurrence at Owl Creek Bridge.txt	False	AN OCCURRENCE AT OWL CREE
Ambrose Bierce___A Son of the Gods, and A Horseman in the Sky.txt	False	A SON OF THE GODS ⋱

Figure 1: Depiction of Dataset and Associated Data

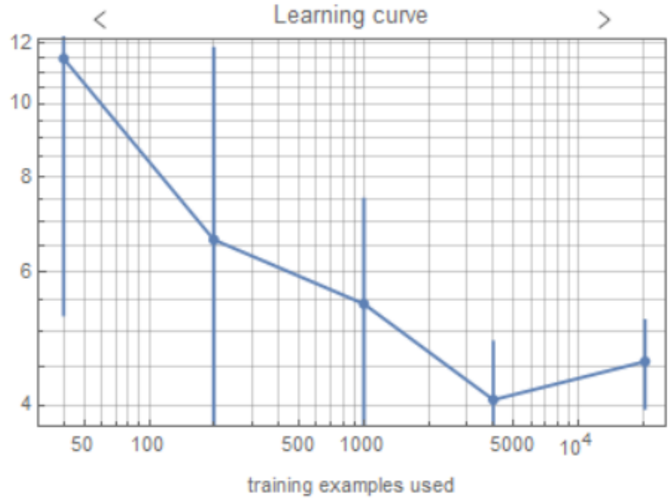

Figure 2: Learning Curve of Markov Model

Training and Deploying the Machine Learning Model

- Experimented with various types of machine learning such as Neural Networks, Random Forest Models, and Linear Regression Models
- Markov Method of machine learning yielded the highest accuracy of approximately 95%
- Created a microsite that takes a text from the user and returns the probability that it is fiction

Discussion/Conclusion

- A model was successfully created that can accurately classify fiction and non-fiction texts.
- The classifier takes an input of any English text greater than 500 characters and gives an output of the probability that the given work is fiction or non-fiction
- In the future, more data can be used from the Wolfram Data Repository to increase the accuracy.
- The classifier could be more specific, incorporating genre identification and mapping the "fictionality" of the work on a scale

I would like to thank my mentor, Sylvia Haas, as well as all the other mentors from the Wolfram Summer Camp, for guiding me and helping me with this project. I would also like to thank Rory Foulger..