Authorship Attribution & Stylometry

...

(CS4120) - Natural Language Processing

Yuval Timen

Introduction

The Washington Post

TUESDAY, SEPTEMBER 19, 1995

A SUPPLEMENT TO THE WASHINGTON POST

Weather

118TH YEAR · No. 268

Unabomber Manuscript Is Published

'Public Safety Reason Cited in Joint Decisic By Post, N.Y. Times

By Howard Kurtz
Washington Post Staff Writer

After weighing the questi nearly three months, The Was Post and New York Times have to publish in today's Post a 35 word manuscript submitted Unabomber, the serial mail who has possessed to halt his tacks if either newspaper identity critique of industrial Donald E. Graham, The I Editor, and Arthur O. Sulz

This text was sent last June to The New York Times and The Washington Post by the person who calls himself "FC," identified by the FBI as the Unabomber, whom authorities have implicated in three murders and 16 bombings. The author threatened to send a bomb to an unspecified destination "with intent to kill" unless one of the newspapers published this manuscript. The Attorney General and the Director of the FBI recommended publication. An article about the decision to publish the document appears on the front page of today's paper.

INDUSTRIAL SOCIETY AND ITS FUTURE

INTRODUCTION

1. The Industrial Revolution and its consequences have been a disaster for the human race. They have greatly increased the life expectancy of those of us who live in "advanced" countries, but they have destabilized society, have made life unfulfilling, have subjected human beings to indignities, have led to widespread psychological suffering (in the Third World to physical suffering as well) and have inflicted severe

9. The two psychological tendencies that underlie modern leftism we call "feelings of inferiority" and "oversocialization." Feelings of inferiority are characteristic of modern leftism as a whole, while oversocialization is characteristic only of a certain segment of modern leftism; but this segment is highly influential.

FEELINGS OF INFERIORITY

his own needs. The leftist is antagonistic to the concept of competition because, deep inside, he feels like a loser.

17. Art forms that appeal to modern leftish intellectuals tend to focus on scrofulous, defeat and despair, or else they take an inane tone, throwing off rational control as if there were no hope of accomplishing anything through rational calculation and all that was left was to immerse oneself in the sensations of the moment.

18. Modern leftish philosophers tend to dismiss reason, science, objective reality and to assert that everything is culturally relative. It is true that one can ask serious questions about the foundations of scientific knowledge and about how, if at all, the concept of objective reality can be defined. But it is obvious that modern leftish philosophers are not simply cool-headed logicians systematically analyzing the foundations of knowledge. They are deeply involved emotionally in their attack on truth and reality. They attack these concepts because of their own psychological needs. For one thing, their attack is an outlet for hostility, and, to the extent

Historical Uses of AA

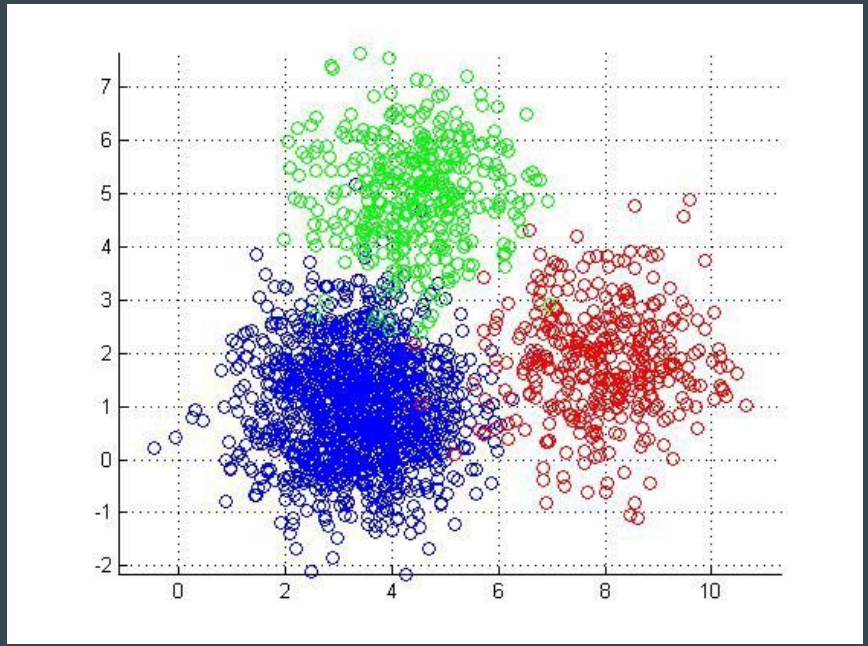
- Determining the authors of the Federalist Papers
- Revealing identity of anonymous writers
- Resolve conflicts about contested writings
- Plagiarism detection
- Can be applied to programs/code
- Forensic linguistics - catching the Unabomber

Outline

1. Methods of Distinguishing Stylometry
2. Examination of Used Datasets
3. Methodology
4. Results/Conclusion
5. Future Works
6. Sources Cited

Methods of Distinguishing Between Writing Style

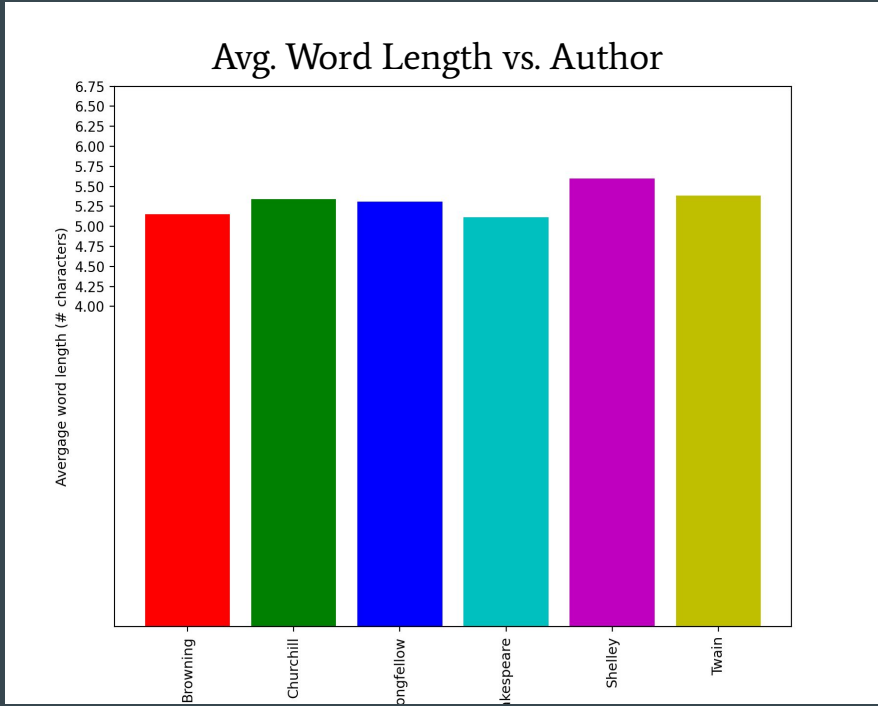
- Authorship Attribution is hard:
  - Open vs. Closed set of authors
  - Inconsistent style within author corpus
- Statistical techniques
  - Grammatical Statistics
  - High dimensional N-grams/embeddings and clustering
- Neural Networks
- Topic Modeling

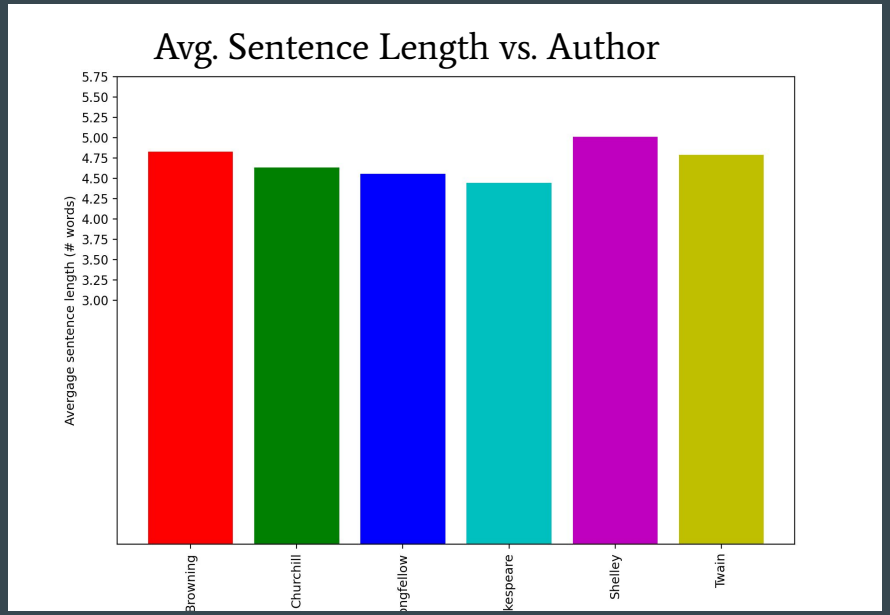


Examination of Used Datasets

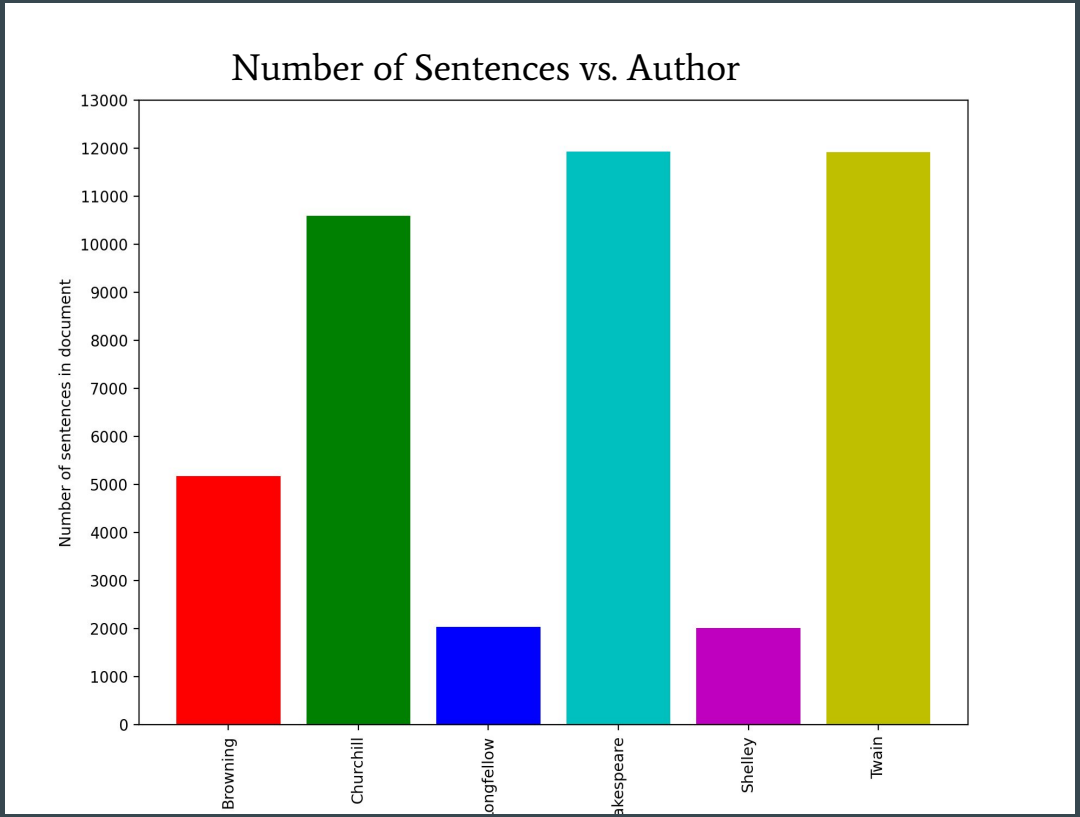
- Robert Browning
- Winston Churchill
- Henry Wadsworth Longfellow
- William Shakespeare
- Percy Bysshe Shelley
- Mark Twain

Examination of Used Datasets - Can we use basic methods?

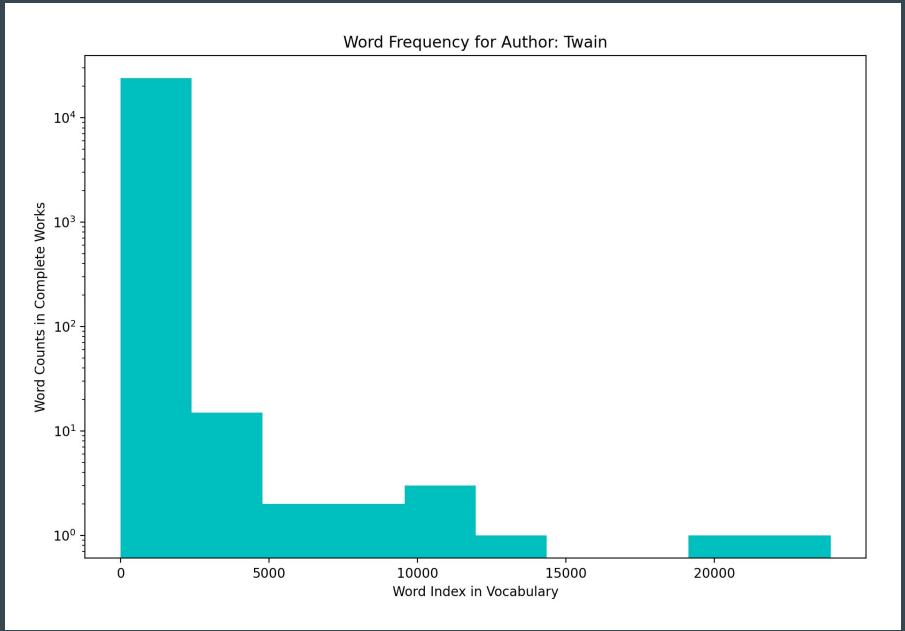


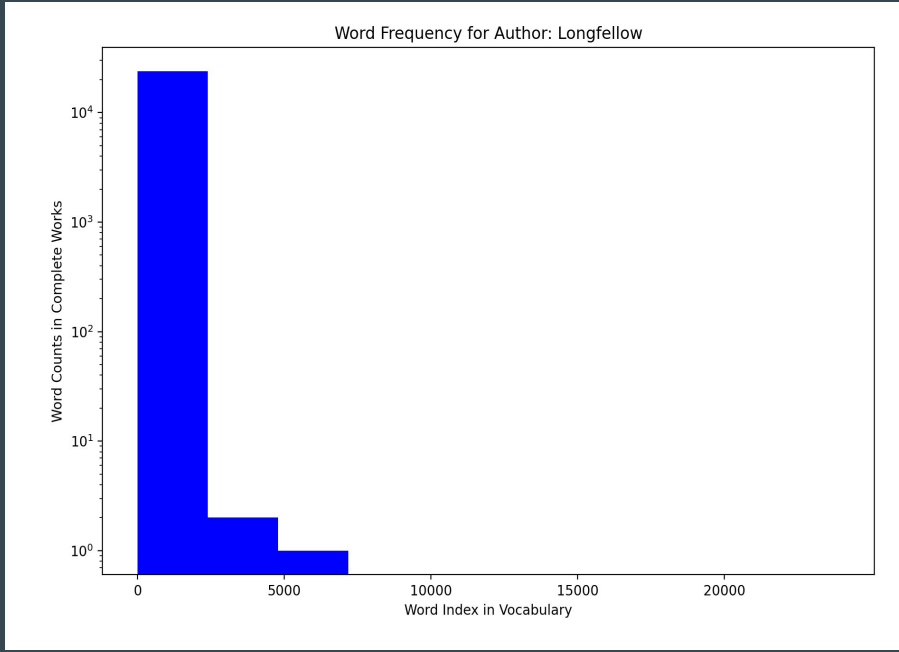


Examination of Used Datasets - How will we evaluate?

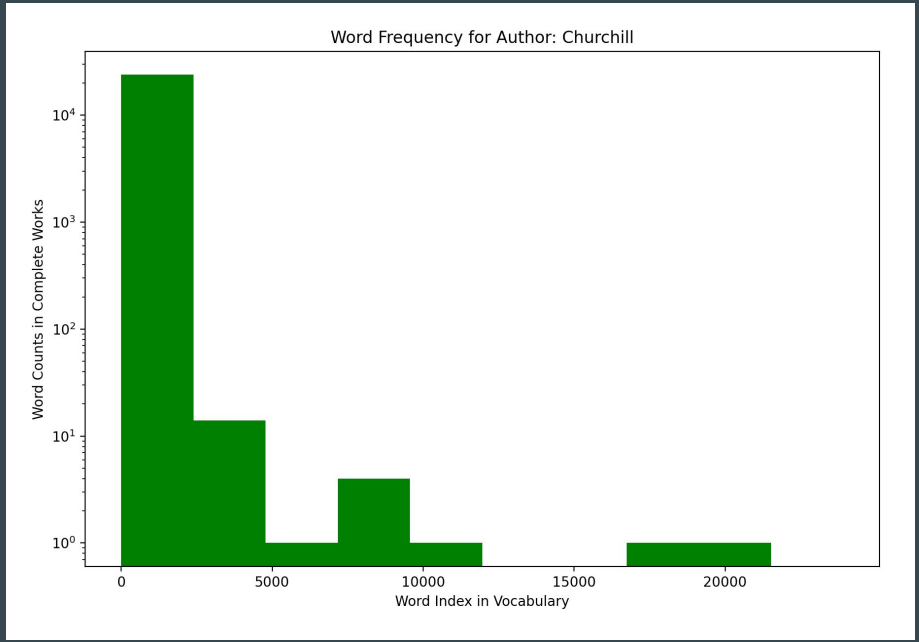


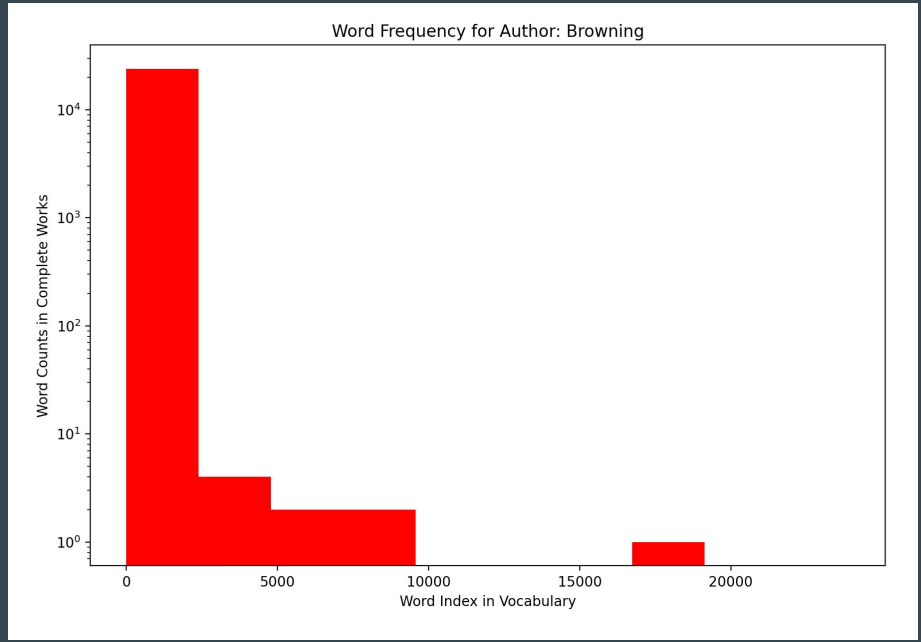
Examination of Used Datasets



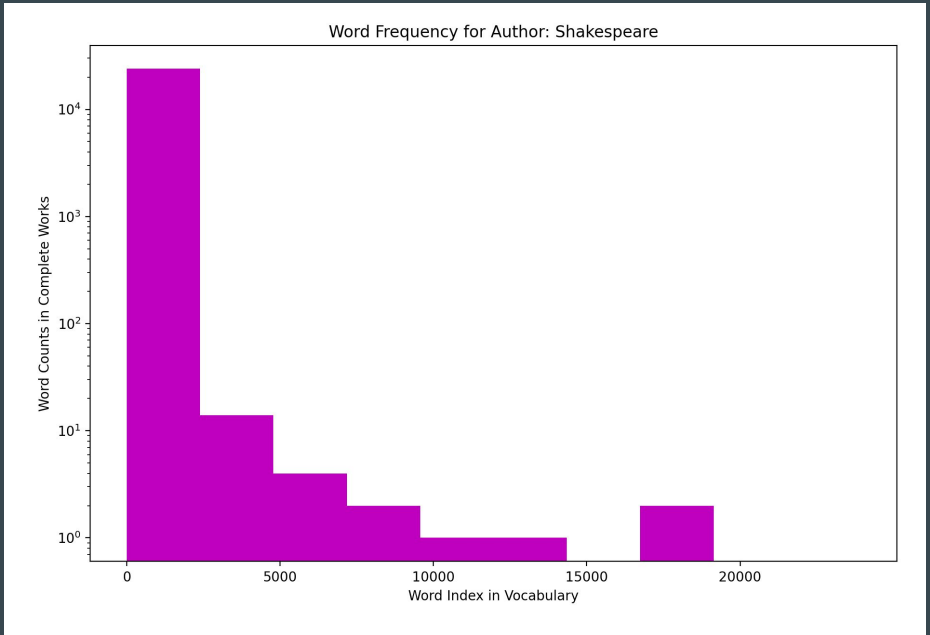


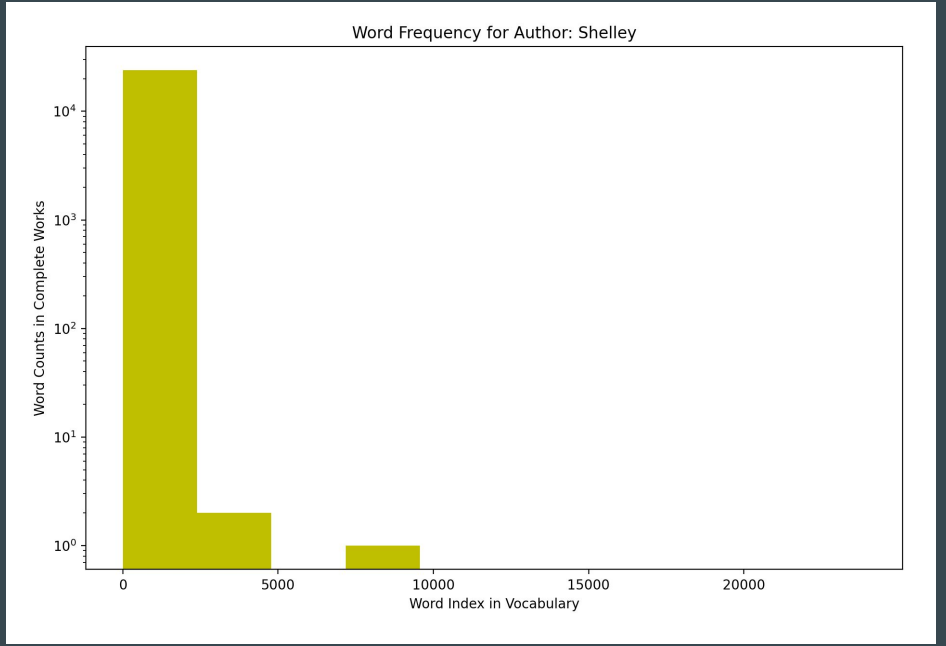
Examination of Used Datasets

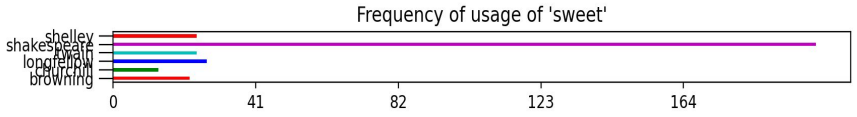


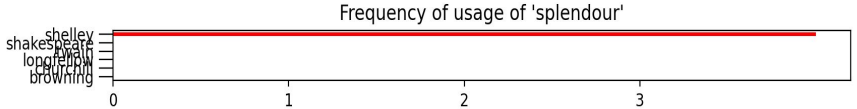


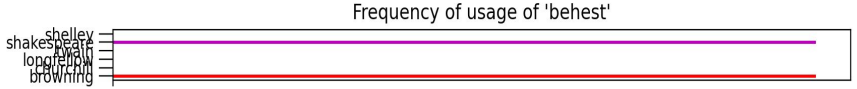
Examination of Used Datasets

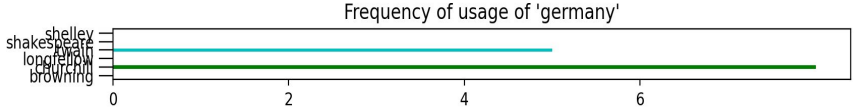


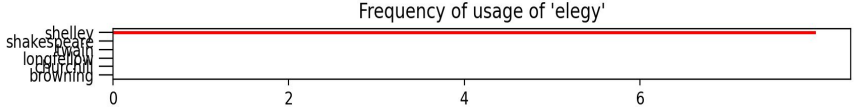


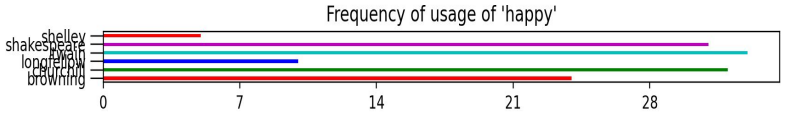

Frequency of usage of 'sweet'
shelley
shakespeare
longfellow
0
41
82
123
164

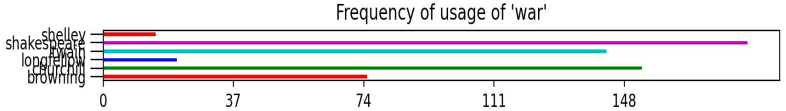

Frequency of usage of 'splendour'
shelley
shakespeare
longfellow
browning
0
1
2
3

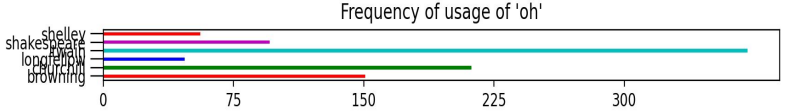

Frequency of usage of 'behest'
shelley
shakespeare
longfellow
browning
0

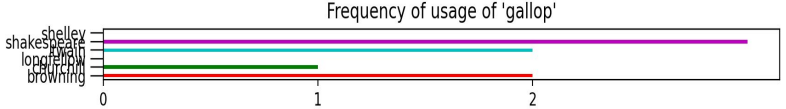

Frequency of usage of 'germany'
shelley
shakespeare
longfellow
browning
0
2
4
6

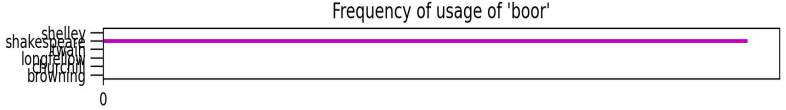

Frequency of usage of 'elegy'
shelley
shakespeare
longfellow
browning
0
2
4
6


Frequency of usage of 'happy'
shelley
shakespeare
longfellow
browning
0
7
14
21
28


Frequency of usage of 'war'
shelley
shakespeare
longfellow
browning
0
37
74
111
148


Frequency of usage of 'oh'
shelley
shakespeare
longfellow
browning
0
75
150
225
300

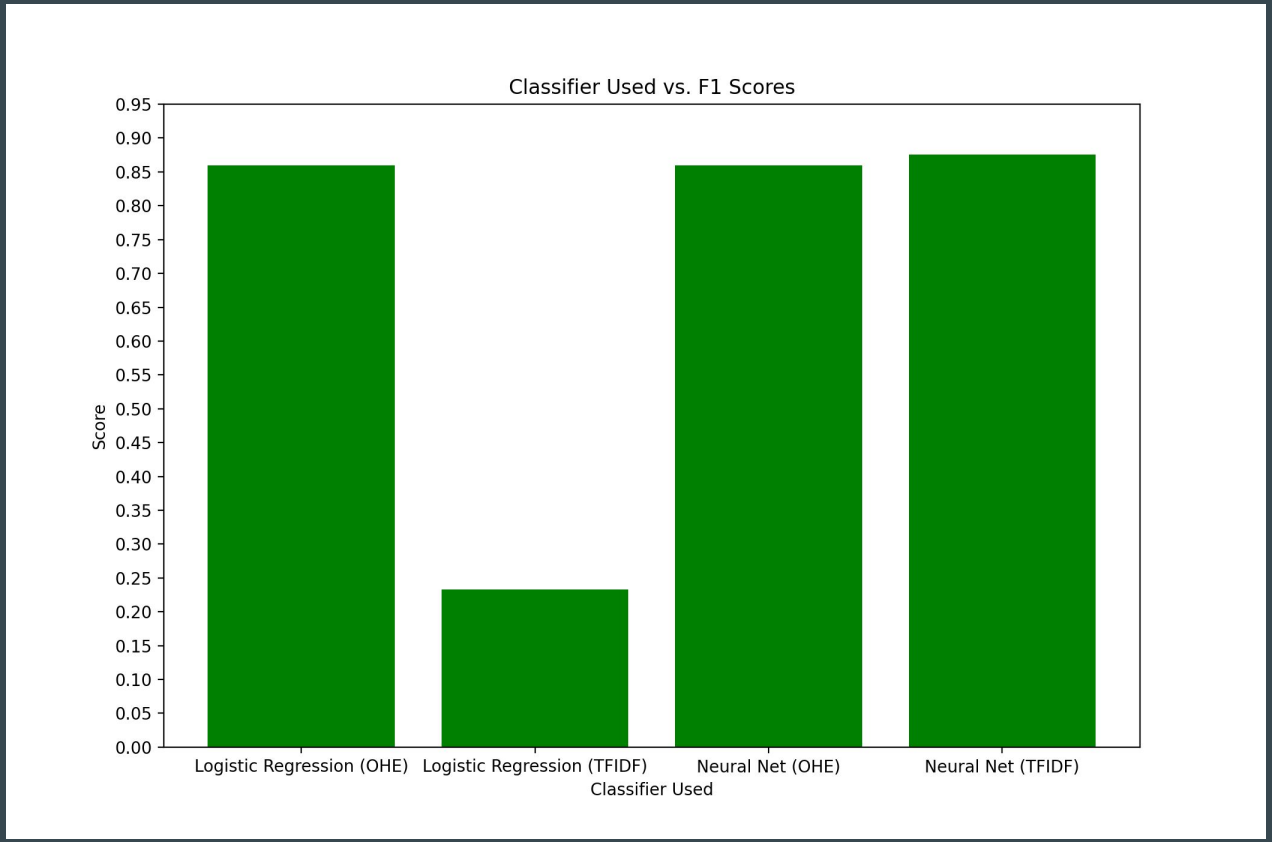

Frequency of usage of 'gallop'
shelley
shakespeare
longfellow
browning
0
1
2


Frequency of usage of 'boor'
shelley
shakespeare
longfellow
browning
0

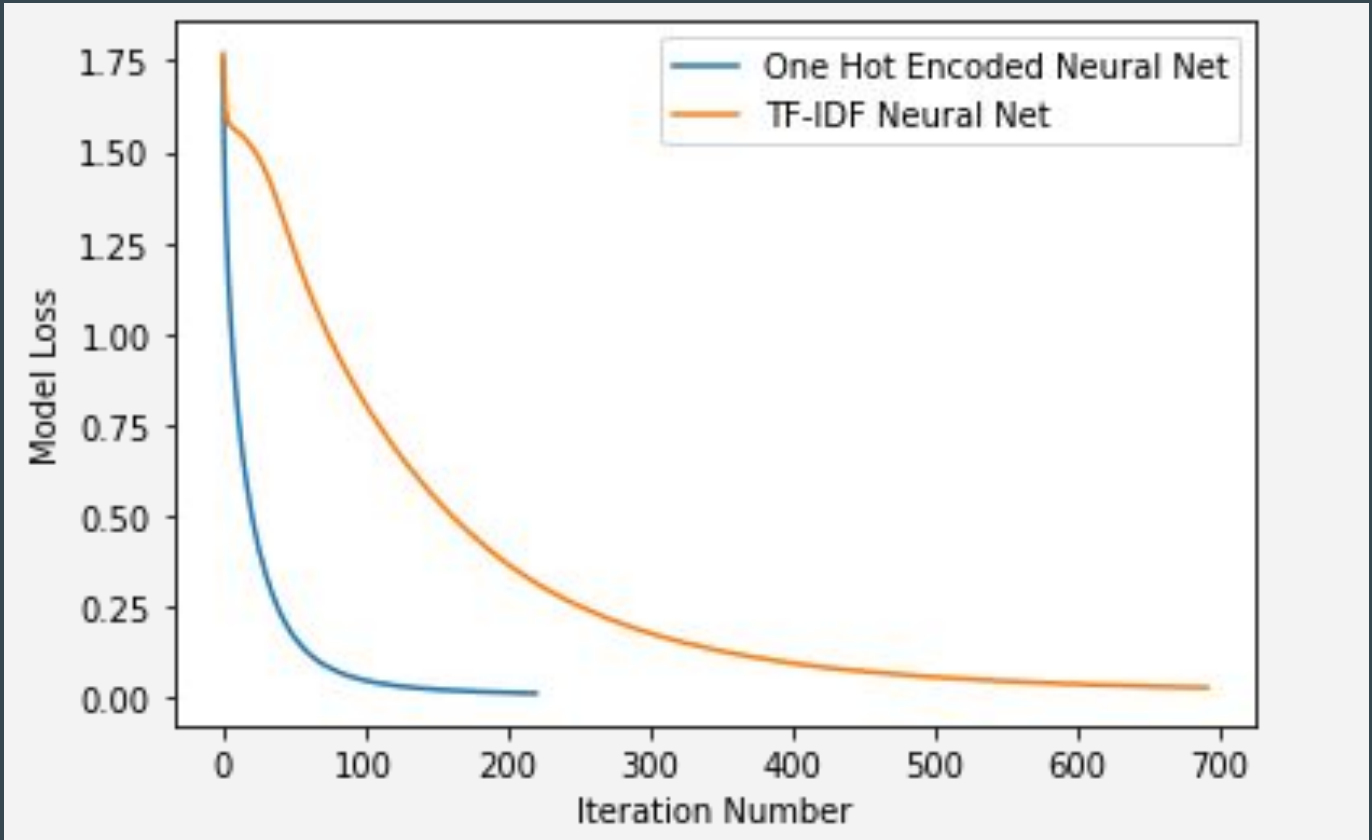
Methodology

1. Data cleaning
2. Deciding which datasets to use
3. Featurizing
    a. One Hot Encodings
    b. TF-IDF
4. Models
    a. Logistic Regression
    b. Feedforward Neural Network
5. Evaluations
    a. P/R/F1 with weighted averaging

Results/Conclusion



Results/Conclusion



Summary

- Authorship Attribution is a hard problem
- Useful in many fields and applications
- Basic statistical methods don’t work very well
- Unsupervised methods > Supervised methods
- How does one characterize stylometry?

Future Works

- Use unsupervised techniques
  - Latent Dirichlet Allocation
  - K-means clustering
- Explore Open Set AA
- Add information about grammatical structure as a feature

Sources

1. https://www.quora.com/LDA-Topic-Modelling-output-what-do-the-output-values-represent
2. https://www.aclweb.org/anthology/C18-1029
3. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.679.2951&rep=rep1&type=pdf
4. https://towardsdatascience.com/hyperparameter-tuning-c5619e7 e6624
5. https://www.aclweb.org/anthology/C18-1029.pdf
6. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.679.2951&rep=rep1&type=pdf
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256385/