Classifying Books

Beau Carlborg, Clara Richter and Sunny Chiu

Redefining our project

Original Project:

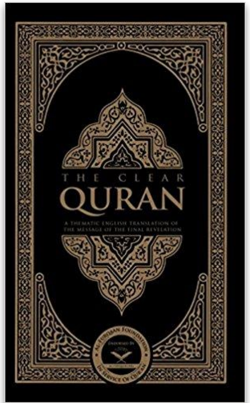
Train a model to identify the probable publication date of a text using date obtained from Project Gutenberg and publication dates retrieved from GoodReads.

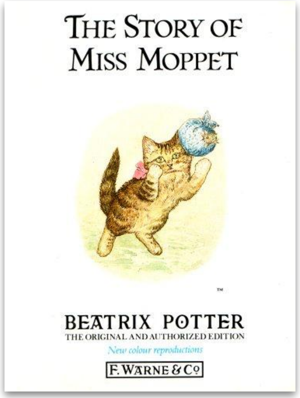
New Project:

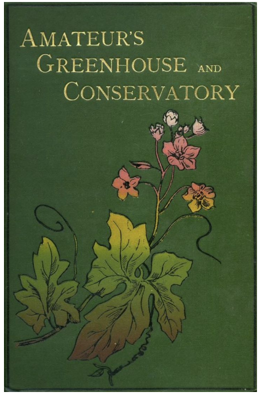
Train a model to identify the author of a text using works obtained from Project Gutenberg.

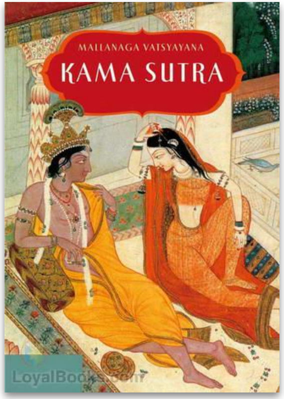
Why we re-worked the project.

- Many works in our data set were written by a small group of authors of a very large period of time.
- The works we retrieved from Project Gutenberg came from a much wider range of genres than we initially expected.


THE CLEAR
QURAN
A THEMATIC ENGLISH TRANSLATION OF THE MESSAGE OF THE FINAL REVELATION





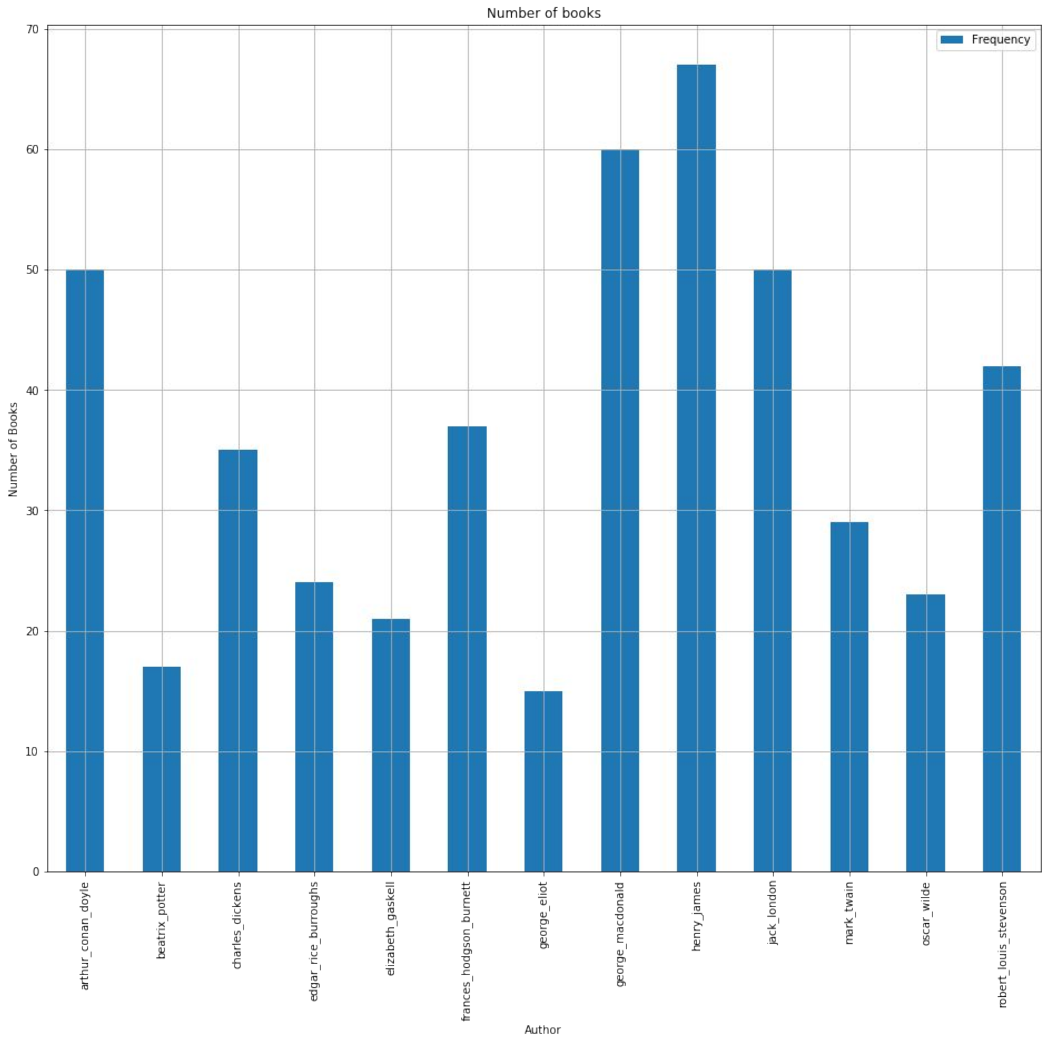


Number of Texts Per Author

A bar graph showing the texts we were able to retrieve for a small set of authors.

By focusing on authorship as opposed to publication date, we were able to drastically increase the amount of data in each of our data classes.

We were also able to use fewer buckets in total which we know to increase accuracy of the model.

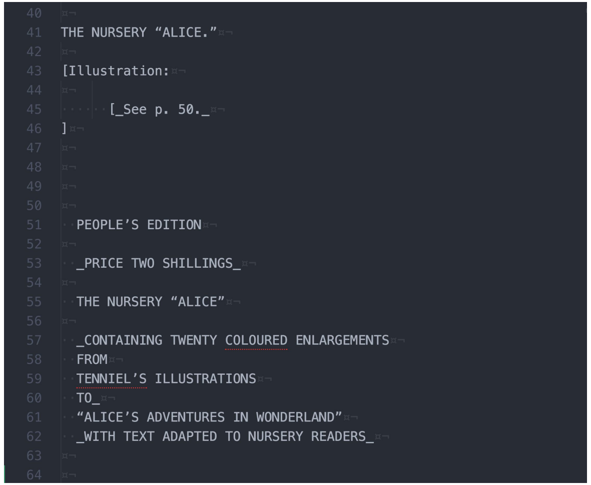


Our Process of Cleaning the Data

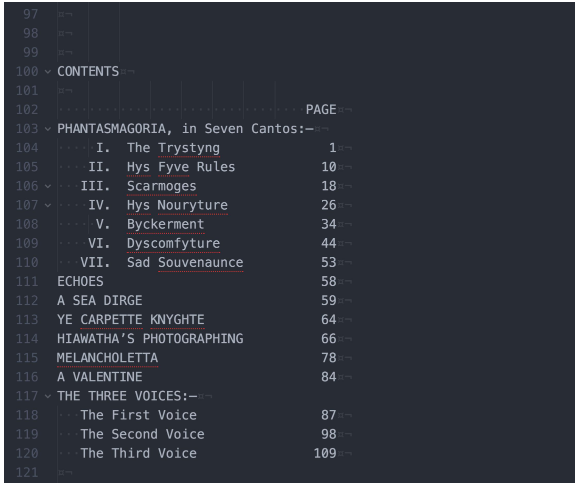
- In order to process our data effectively, we needed to **strip each document** to its bare text (more or less).
- To do this, we needed to remove a set of lines from the top and bottom of each book.

Challenges in data cleaning

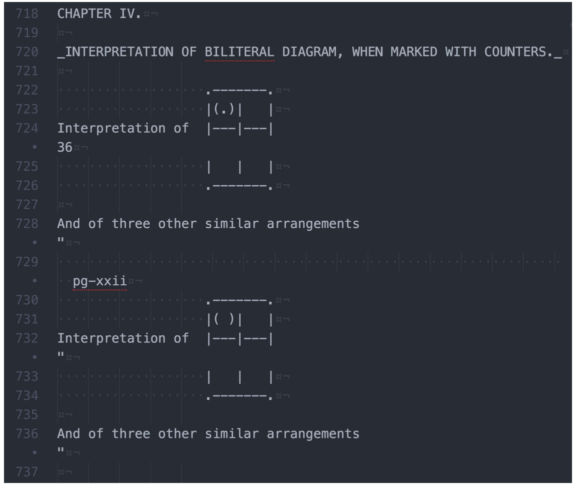
Every document had many different features, so we cleaned the data by hand



```

THE NURSERY “ALICE.”

[Illustration:

      [_See p. 50._
]




  PEOPLE’S EDITION

  _PRICE TWO SHILLINGS_

  THE NURSERY “ALICE”

  _CONTAINING TWENTY COLOURED ENLARGEMENTS
  FROM
  TENNIEL’S ILLUSTRATIONS
  TO_
  “ALICE’S ADVENTURES IN WONDERLAND”
  _WITH TEXT ADAPTED TO NURSERY READERS_


```



```



CONTENTS

                                PAGE
PHANTASMAGORIA, in Seven Cantos:—
     I.  The Trystyng              1
    II.  Hys Fyve Rules           10
   III.  Scarmoges                18
    IV.  Hys Nouryture            26
     V.  Byckerment               34
    VI.  Dyscomfyture             44
   VII.  Sad Souvenaunce          53
ECHOES                            58
A SEA DIRGE                       59
YE CARPETTE KNYGHTE               64
HIAWATHA’S PHOTOGRAPHING          66
MELANCHOLETTA                     78
A VALENTINE                       84
THE THREE VOICES:—
   The First Voice                87
   The Second Voice               98
   The Third Voice               109

```



```
CHAPTER IV.

_INTERPRETATION OF BILITERAL DIAGRAM, WHEN MARKED WITH COUNTERS._

                   .-------.
                   |(.)|   |
Interpretation of  |---|---|
36
                   |   |   |
                   .-------.

And of three other similar arrangements
"

  pg-xxii
                   .-------.
                   |( )|   |
Interpretation of  |---|---|
"
                   |   |   |
                   .-------.

And of three other similar arrangements
"

```

Preprocessing

Extracting information from the text

Stylometric features:

Qualities and characteristics of a written work which capture some small element of an author’s style.

Includes character and word specific features, syntactic features, and structural features.

Our Current Features:

- N-grams
- Parts of speech N-grams
- Sentence Length
- Word Length
- Vocabulary richness
- Function Words
- Punctuation

Challenges in preprocessing

How do we choose our features intelligently?

- Using existing research on document classification, we were able to identify commonly used features.
- We chose features which were common among other similar text classification problems.
- **But we still are not sure of the best way to identify strong features for our model.**

WILEY InterScience® DISCOVER SOMETHING GREAT

A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques

Rong Zheng
Information, Operations, and Management Science, Stern School of Business, New York University, New York, NY 10012. E-mail: rzheng@stern.nyu.edu

Jiexun Li, Hsinchun Chen, and Zan Huang
Artificial Intelligence Lab, Department of Management Information Systems, University of Arizona, Tucson, AZ 85721.

With the rapid proliferation of Internet technologies and applications, misuse of online messages for inappropriate or illegal purposes has become a major concern for society. The anonymous nature of online-message distribution makes identity tracing a critical problem. We developed a framework for authorship identification of online messages to address the identity-tracing problem. In this framework, four types of writing-style features (lexical, syntactic, structural, and content-specific features) are extracted and inductive learning algorithms are used to build feature-based classification models to identify authorship of online messages. To examine this framework, we conducted experiments on English and Chinese online-newsgroup messages. We compared the discriminating power of the four types of features and of three classification techniques: decision trees, backpropagation neural networks, and support vector machines. The experimental results showed that the proposed approach was able to identify authors of online messages with satisfactory accuracy of 70 to 95%. All four types of message features contributed to discriminating authors of online messages. Support vector machines outperformed the other two classification techniques in our experiments. The high performance we achieved for both the English and Chinese datasets showed the potential of this approach in a multiple-language context.

Introduction

The rapid development and proliferation of Internet technologies and applications have created a new way to share information across time and space. A wide range of activities have evolved over the Internet, ranging from simple information exchange and resource sharing to virtual communications and e-commerce activities. In particular, online messages are being extensively used to distribute information over Web-based channels such as e-mail, Web sites, Internet newsgroups, and Internet chat rooms.

Unfortunately, online messages also can be misused for the distribution of unsolicited or inappropriate information such as junk mail (commonly referred to "spamming") and offensive/threatening messages. Moreover, criminals have been using online messages to distribute illegal materials, including pirated software, child pornography materials, stolen properties, and so on. In addition, criminal or terrorist organizations also use online messages as one of their major communication channels. These activities have spawned the concept of "cybercrime." Cybercrime was defined by Thomas and Loader (2000) as illegal computer-mediated activities which can be conducted through global electronic networks.

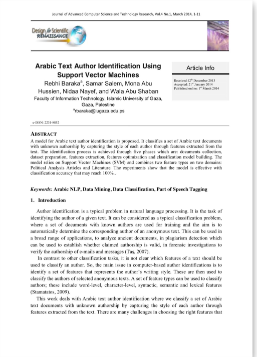
A common characteristic of online messages is anonymity. People usually do not need to provide their real identity information such as name, age, gender, and address. In many misuse or crime cases of online messages, the sender will attempt to hide his/her true identity to avoid detection. For example, the sender's address can be forged or routed through an anonymous server, or the sender can use multiple usernames to distribute online messages via different anonymous channels. Therefore, the anonymity of online messages imposes unique challenges to identity tracing in cyberspace. As a result of the sheer growth of cyber users and activities, efficient automated methods for identity tracing are becoming imperative. Authorship identification based on analyzing stylistic features of online messages is suggested to be a possible solution.

In this article, we propose a framework of authorship identification on online messages. In this framework, four types of features that are identified in authorship-analysis

Received October 1, 2003; revised August 1, 2004; accepted February 10, 2005

© 2005 Wiley Periodicals, Inc. • Published online 21 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20316

JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 57(3):378–393, 2006



Journal of Advanced Computer Science and Technology Research, Vol.4 No.1, March 2014, 1-11

Design for Scientific RENAISSANCE

Arabic Text Author Identification Using Support Vector Machines

Rebhi Baraka[a], Samar Salem, Mona Abu Hussien, Nidaa Nayef, and Wala Abu Shaban
Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine
[a]rbaraka@iugaza.edu.ps

Article Info

Received:12th December 2013
Accepted: 21st January 2014
Published online: 1st March 2014

e-ISSN: 2231-8852

Abstract

A model for Arabic text author identification is proposed. It classifies a set of Arabic text documents with unknown authorship by capturing the style of each author through features extracted from the text. The identification process is achieved through five phases which are: documents collection, dataset preparation, features extraction, features optimization and classification model building. The model relies on Support Vector Machines (SVM) and combines two feature types on two domains: Political Analysis Articles and Literature. The experiments show that the model is effective with classification accuracy that may reach 100%..

Keywords: **Arabic NLP, Data Mining, Data Classification, Part of Speech Tagging**

1. Introduction

Author identification is a typical problem in natural language processing. It is the task of identifying the author of a given text. It can be considered as a typical classification problem, where a set of documents with known authors are used for training and the aim is to automatically determine the corresponding author of an anonymous text. This can be used in a broad range of applications, to analyze ancient documents, in plagiarism detection which can be used to establish whether claimed authorship is valid, in forensic investigations to verify the authorship of e-mails and messages (Taş, 2007).

In contrast to other classification tasks, it is not clear which features of a text should be used to classify an author. So, the main issue in computer-based author identifications is to identify a set of features that represents the author's writing style. These are then used to classify the authors of selected anonymous texts. A set of feature types can be used to classify authors; these include word-level, character-level, syntactic, semantic and lexical features (Stamatatos, 2009).

This work deals with Arabic text author identification where we classify a set of Arabic text documents with unknown authorship by capturing the style of each author through features extracted from the text. There are many challenges in choosing the right features that

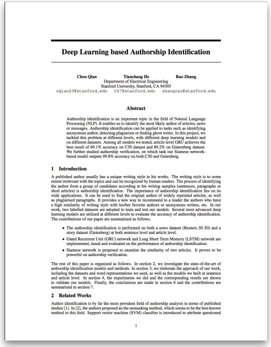
NTNU – Trondheim
Norwegian University of
Science and Technology

Authorship Identification of Research Papers

Simen Skoglund

Master of Science in Informatics
Submission date: August 2015
Supervisor: Kjetil Nørvåg, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science



Deep Learning based Authorship Identification

Chen Qian **Tianchang He** **Rao Zhang**
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
cqian23@stanford.edu th7@stanford.edu zhangrao@stanford.edu

Abstract

Authorship identification is an important topic in the field of Natural Language Processing (NLP). It enables us to identify the most likely author of articles, news or messages. Authorship identification can be applied to tasks such as identifying anonymous author, detecting plagiarism or finding ghost writer. In this project, we tackled this problem at different levels, with different deep learning models and on different datasets. Among all models we tested, article-level GRU achieves the best result of 69.1% accuracy on C50 dataset and 89.2% on Guternberg dataset. We further studied authorship verification, on which task our Siamese network-based model outputs 99.8% accuracy on both C50 and Gutenberg.

1 Introduction

A published author usually has a unique writing style in his works. The writing style is to some extent irrelevant with the topics and can be recognized by human readers. The process of identifying the author from a group of candidates according to his writing samples (sentences, paragraphs or short articles) is authorship identification. The importance of authorship identification lies on its wide applications. It can be used to find the original author of widely reprinted articles, as well as plagiarised paragraphs. It provides a new way to recommend to a reader the authors who have a high similarity of writing style with his/her favorite authors or anonymous writers, etc. In our work, two labelled datasets are adopted to train and test our models. Several most advanced deep learning models are utilized at different levels to evaluate the accuracy of authorship identification. The contributions of our paper are summarized as follows.

- The authorship identification is performed on both a news dataset (Reuters_50_50) and a story dataset (Gutenberg) at both sentence level and article level.
- Gated Recurrent Unit (GRU) network and Long Short Term Memory (LSTM) network are implemented, tuned and evaluated on the performance of authorship identification.
- Siamese network is proposed to examine the similarity of two articles. It proves to be powerful on authorship verification.

The rest of this paper is organized as follows. In section 2, we investigate the state-of-the-art of authorship identification models and methods. In section 3, we elaborate the approach of our work, including the datasets and word representations we used, as well as the models we built at sentence and article level. In section 4, the experiments we did and the corresponding results are shown to validate our models. Finally, the conclusions are made in section 6 and the contributions are summarized in section 7.

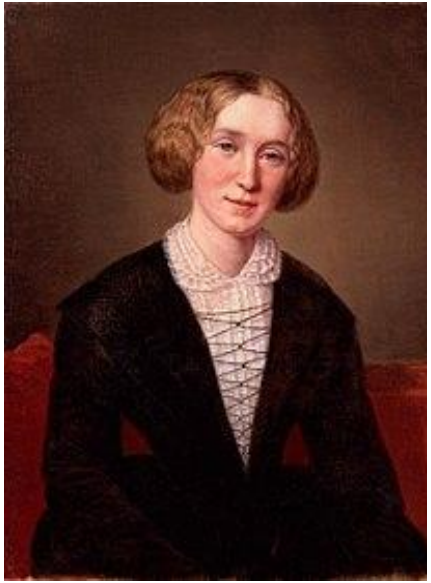
2 Related Works

Author identification is by far the most prevalent field of authorship analysis in terms of published studies [1]. In [2], the authors proposed an the unmasking method, which seems to be the best known method in this field. Support vector machine (SVM) classifier is introduced to attribute questioned

1

Attempting to assess our features.

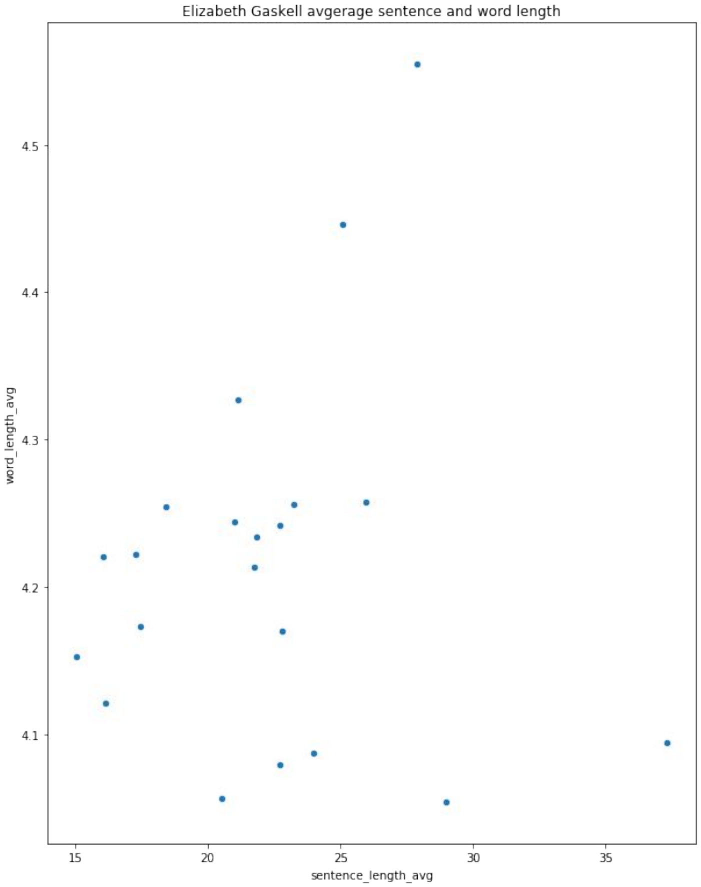
Elizabeth Gaskell
29 September 1810 –
12 November 1865



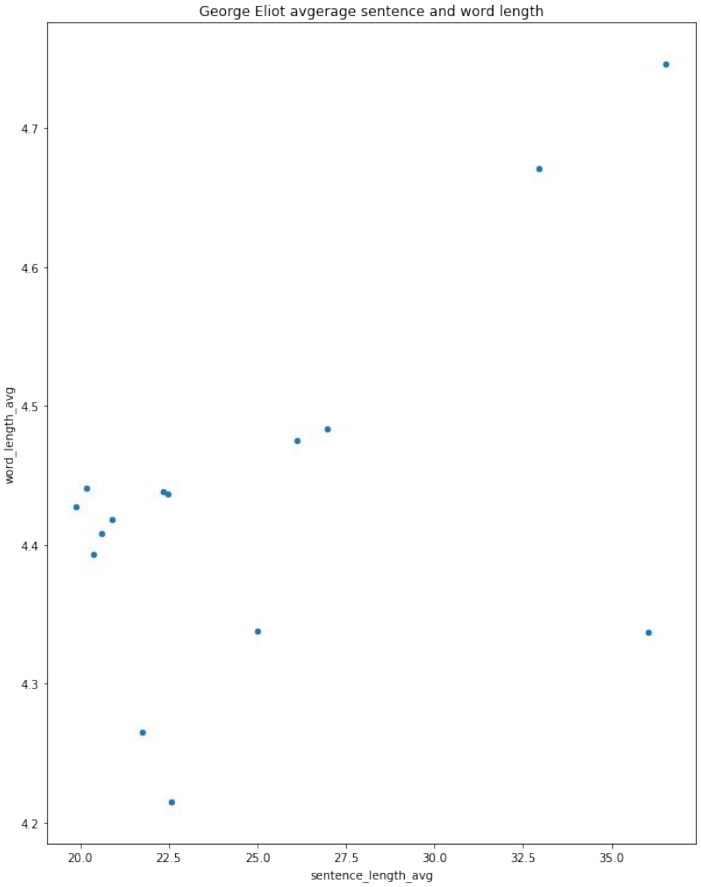
Mary Anne Evans

George Elliot

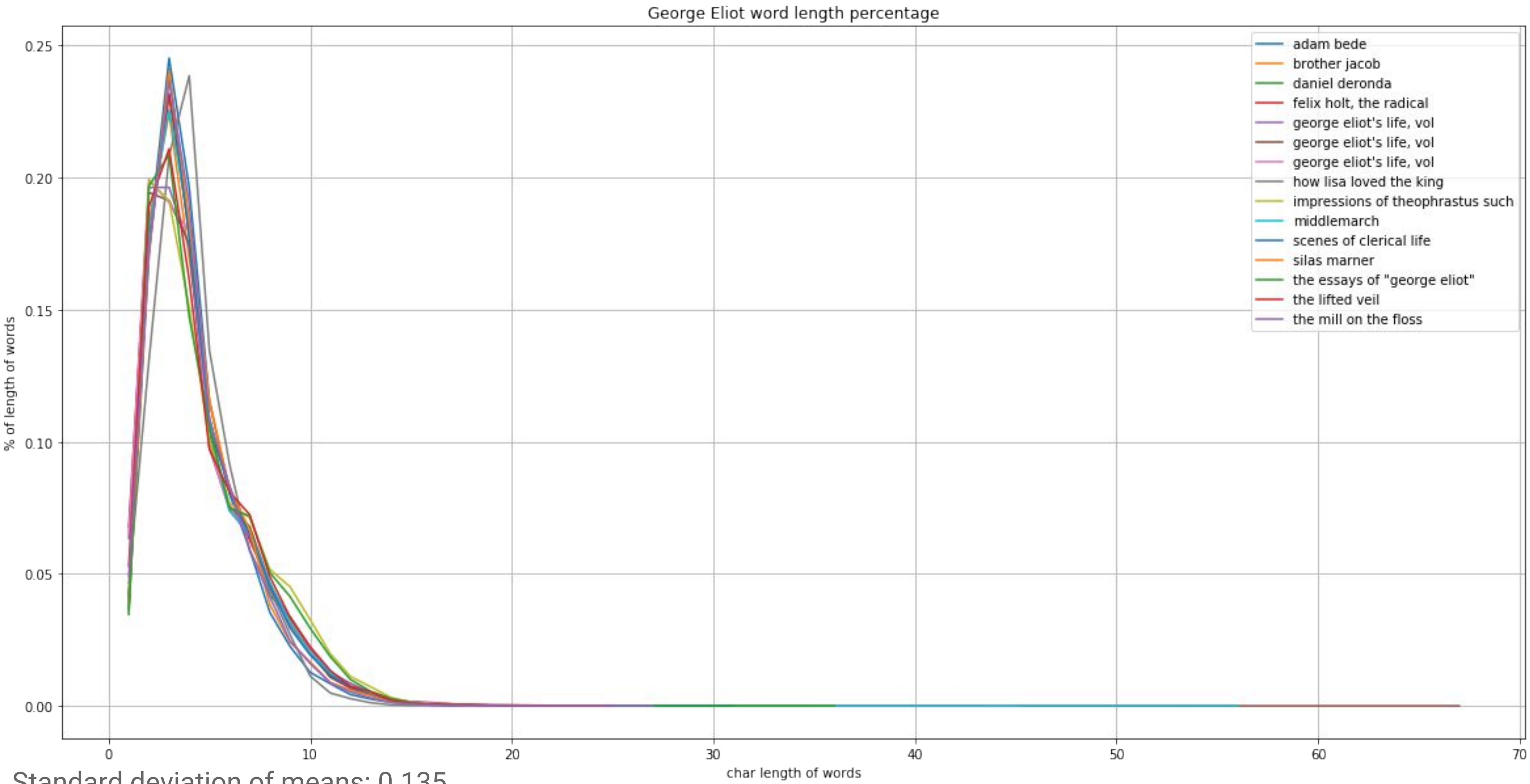
22 November 1819 – 22 December 1880



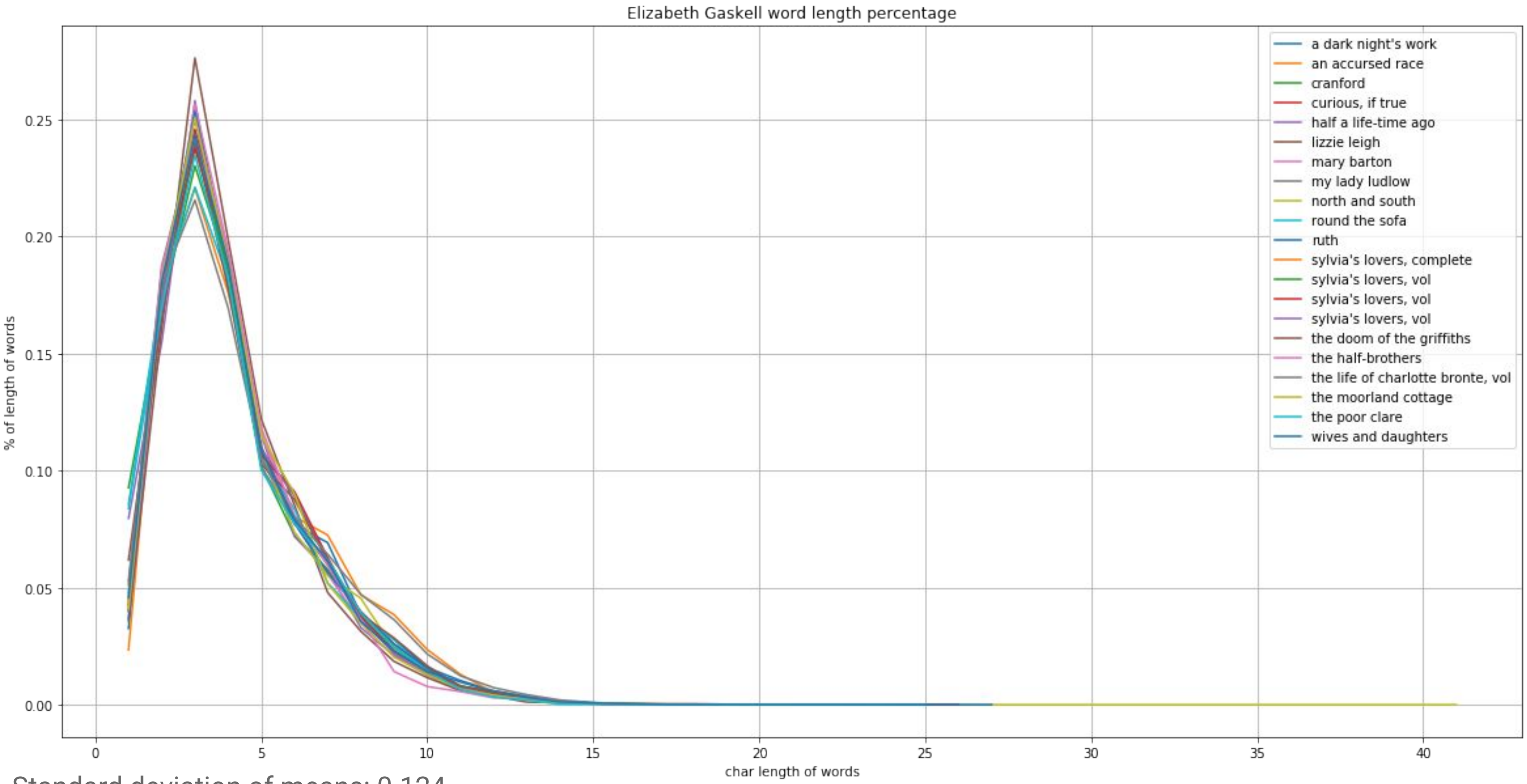
Correlation: 0.05



Correlation: 0.55



Standard deviation of means: 0.135
Standard deviation of Standard deviations: 0.202



Standard deviation of means: 0.124
Standard deviation of Standard deviations: 0.101

Our Plan Moving forward

- Choose a model to run our data on
  - Support Vector Machine
  - Artificial Neural Network
  - Decision Trees
  - Random Forest
  - K-Nearest Neighbor
- Choose our strategy for splitting testing and training data.
  - Percent split or K-Fold Cross Validation
- Measuring Success of our model
  - Start with True and False positives and Negatives
  - Possibly add metrics for precision and recall
- How we will visualize results
  - We hope to plot the percent of correct author identifications our model produces as the number of authors increases