Exploring the Author Identification Task

Thomas Pace & Sofia Nystrom

Group: Author Identification, Spring 2020

Abstract

The task of Author Identification is to identify authors behind written work from a set of plausible candidates. Here we apply cluster analysis on a new text corpus using standard features while both including and excluding *stop words* and TF-IDF. While we find that TF-IDF increases seperation between different clusters, overall, we find little separation across the techniques. We further evaluate the results by calculating the Shannon entropy of the author distribution within clusters which highlighted TF-IDF using LDA as the best performing method. The complexity behind Author Identification warrants further work on this dynamic area of research.

Keywords: author identification, n-grams, text analysis, cluster analysis

Contents

Introduction

While attributed to early work by Mosteller & Wallace in the mid 1960's, the task of *Author Identification* is to identify the author of a written work from a set of plausible candidates. The technique stems from *Stylometry*, in where it is hypothesized that a writer has a *stylistic fingerprint* that can be studied and *learned* from (Stamatatos, 2009). To implement the task, the style of the written work (e.g., average word length/sentence complexity) can be analyzed using both supervised and unsupervised learning. Author identification is an emerging field with applications in a wide range of areas, including various online platforms, writings, and books.

Figure 1 illustrates the breath of research done on the author identification task. Indeed, existing work ranges from supervised and unsupervised learning in areas of deep learning, neural networks and clustering. In this paper we explore author identification using a large set of standard features and TF-IDF through dimensionality reduction techniques and cluster analysis using a large set of features on a novel English text corpus[1] in the *unsupervised* setting.

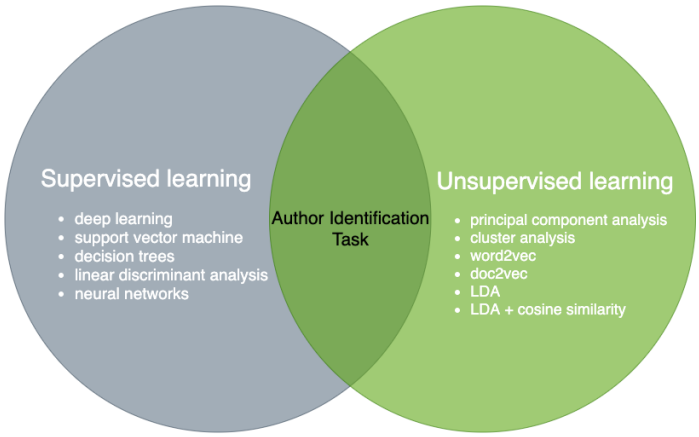

Figure 1: Venn diagram showing how the author identification task intersects between supervised and unsupervised learning.

[1]It may be noted that in the past 20 years, author identification has been centered around English written documents (Waheed et al, 2019).

Overview

Figure 2 depicts a Flow diagram of the author identification task we apply here. The process can be described in **5 steps: 1) identifying the raw input** (e-books), **2) developing a web scraping** tool **3) pre-processing the data** (creating features and standardizing the data *or* applying id-tdf) **4) applying dimensionality reduction techniques** in form of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and **5) using unsupervised methods**, in form of K-means clustering to identify the author.

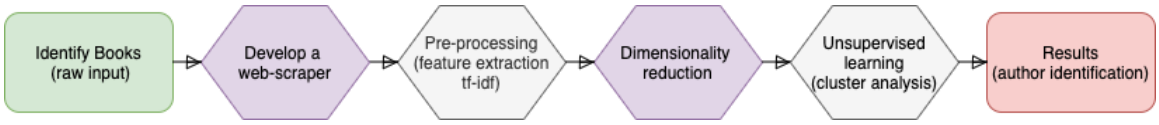

Figure 2: Flowdiagram of the author identification task.

Data

Data Source & Developing a Web-scraping Tool

We use data from https://www.gutenberg.org/. The site contains tens of thousands of e-books available to freely download. To efficiently extract the books, we built a novel web-scraping tool in the Python environment. This was a significant time investment. In broad strokes, this application downloads "specified" content from a given webpage and then returns the e-books as individual text files. The text files are used to create Python *string* objects for each book and placed into a data frame for feature extraction. The Python package *Beautifulsoup* was primarily used in building this tool. A total of 50,000 books were downloaded. After reviewing the corpus, we decided to focus our work on a subset of the books reflecting the most prolific authors. To the author's knowledge, the authors identification task has not been done on this text corpus before.

Methods

Our objective is to use K-means clustering to group the e-books by their respective author. Specifically, our goal is to compare the performance of the cluster analyses on: 1) standard features with *stop words*, 2) standard features excluding *stop words*, and 3) a TF-IDF

high dimensional vector. We then evaluate the performance of the K-means clustering by calculating the Shannon entropy of the author distribution within clusters. Specifically, the Shannon entropy is defined as $-\sum_k p_k \log p_k$ where there are k distinct classes and p_k is the proportion of the k^{th} class. For example, the entropy will be minimized if no author has books present in multiple clusters. In contrast, the entropy will be maximized if all authors have books evenly distributed between clusters.

Pre-processing

To capture the style of the author, we make use of n-grams separating each remaining e-book by words. An important consideration concerns the inclusion of *stop words*. Stop words include words such *a*, *as the*, *that*, and *this* and typically make up a large share of n-grams in a given document. While some argue that these high-frequency words diminishes the ability to distinguish work between different authors due their low lexical content, others argue that they are part of the author's style (Anwar et al., 2019). To explore this on our text corpus, our pre-processing both includes and excludes these high-frequency words to then analyze if their inclusion makes a difference.

Since our goal is to group work by author, we focus on prolific authors who have many books in the corpus. We further exclude books with unknown attribution, leaving a total of 43 authors left for analysis. The mean number of books per author was 47 (min=19, max=146).

Feature Extraction

A plethora of features have been used in the Author Identification task. Here we use a combination of 12 distinct (standard) features, including the *lexicon diversity*, *mean words/syllabus per sentence*, *smog index*, and a *bag of words*. A complete list of all features is available in the Appendix. The features were calculated using a combination of custom implementations and the Python packages *TextStat* and *NLTK*. As mentioned, we calculate the features on two set of pre-processed text documents: one that includes the *stop* words and one that excludes them.

TF-IDF

TF-IDF, or term frequency–inverse document frequency is commonly used in the author identification task. TF-IDF is a statistic that measures the *importance* of a word within a

document. The method discounts commonly used words while placing more weight on less commonly used words. In technical terms, the approach maps the high-frequency words and then re-weighs them to generate an idf vector across the text corpus. Our TF-IDF vector had a shape of (2,753, 100,000).

Dimensionality Reduction

We normalized the standard features (mean=0 and variance=1) and then use *Principal Component Analysis* (PCA) and Linear *Discriminant Analysis* (LDA) to create a linear 2-dimensional dataframe.[2] In broad strokes, PCA takes a set of dimensions d and makes a linear combination, while maximizing the retained variance. Dimensionality reduction is a commonly used technique as it often retains the majority of the total variability in the data. We further apply the same dimensionality reduction procedure to the TF-IDF high-dimensional vector.

Discussion

Cluster Analysis

Figure 3 depicts selected scatter plots where the e-books are colored by their respective authors versus the results of the K-means clustering aimed to cluster the features into distinct groups for either PCA or LDA.[3] Prior to running K-means, which we initialized with K-means++ we explored the *Elbow method* to assess the number of clusters for each method (seen in Figure 4). In terms of the standard features, while natural grouping are seen in the scatter plots, little separation is seen between the different clusters. This holds true for when the stop words are included/excluded. In contrast, TF-IDF appear to slightly increase seperation between the clusters, in particular for points in the right area of the plot.[4] Figure 5 shows the Shannon entropy output. As seen, TF-IDF using LDA performed the best in comparison to the other methods as seen by its relatively distant position from a random assignment.[5]

[2]We assessed the correlation between the features via a correlation matrix. As expected, many of the features are highly correlated (in both directions).

[3]Note that the cluster plots show only one particular iteration after running the K-means algorithm.

[4]It should be noted that using PCA on our standard features resulted in a less than optimal clustering of authors as demonstrated by a simple scatter plot

[5]Note that "ns" stands for "no stop words included".

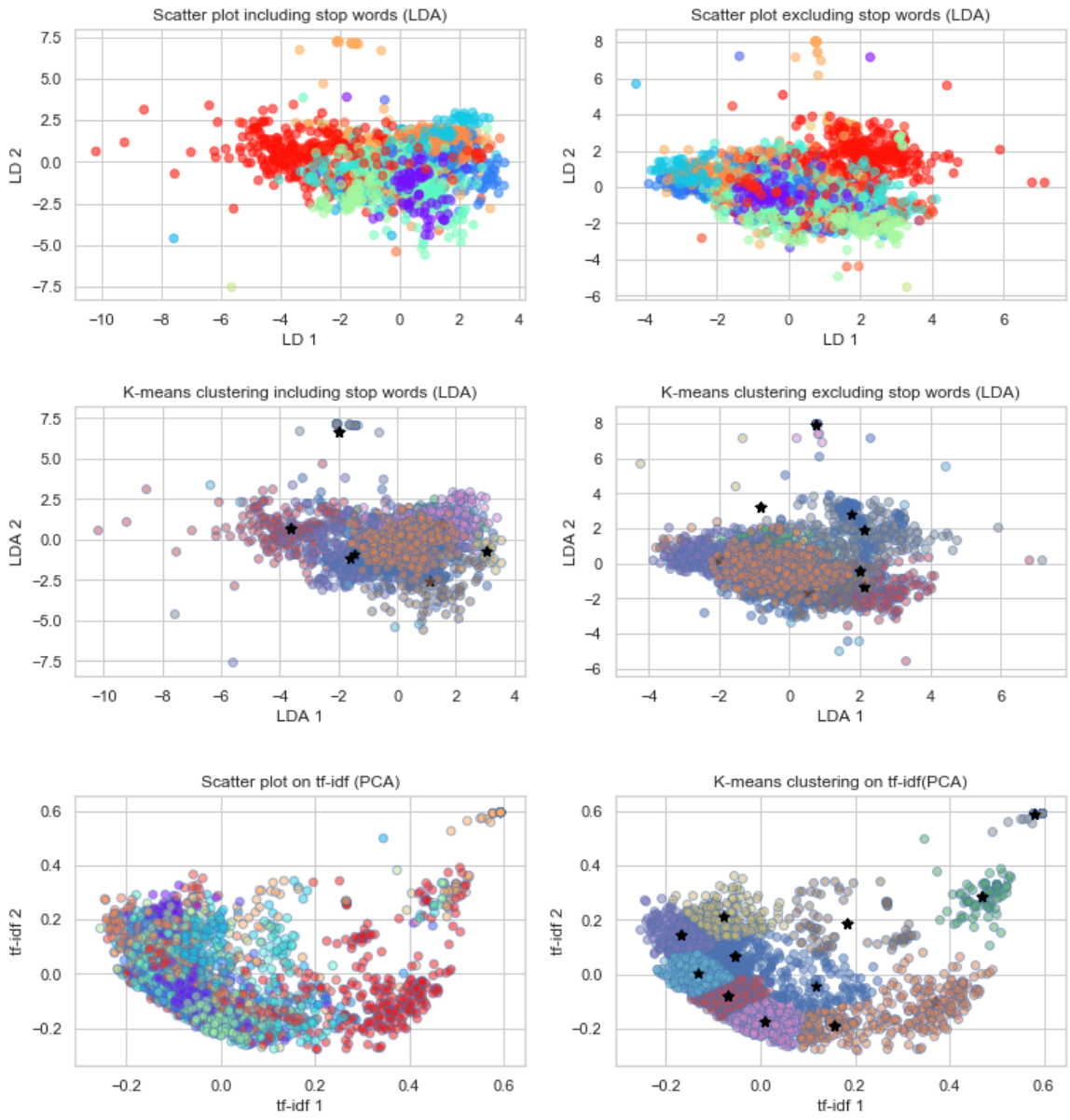

Figure 3: Scatter plots vs. K-means clustering for our standard features and TF-IDF.

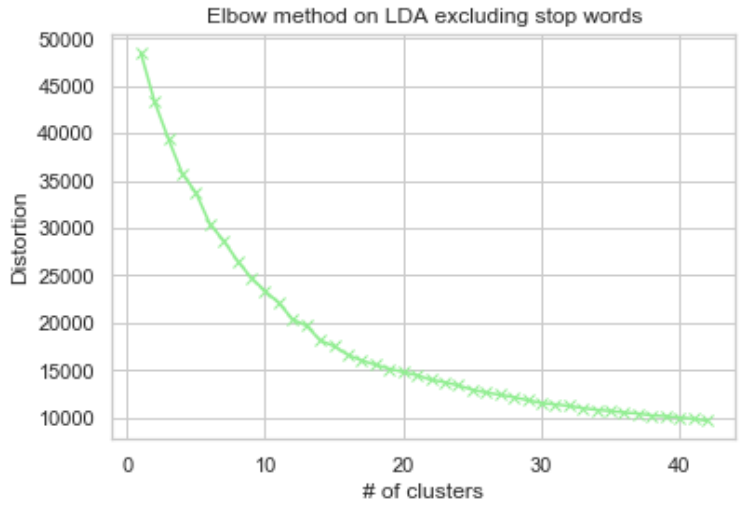

Figure 4: Example of using the Elbow method in assessing the number of clusters.

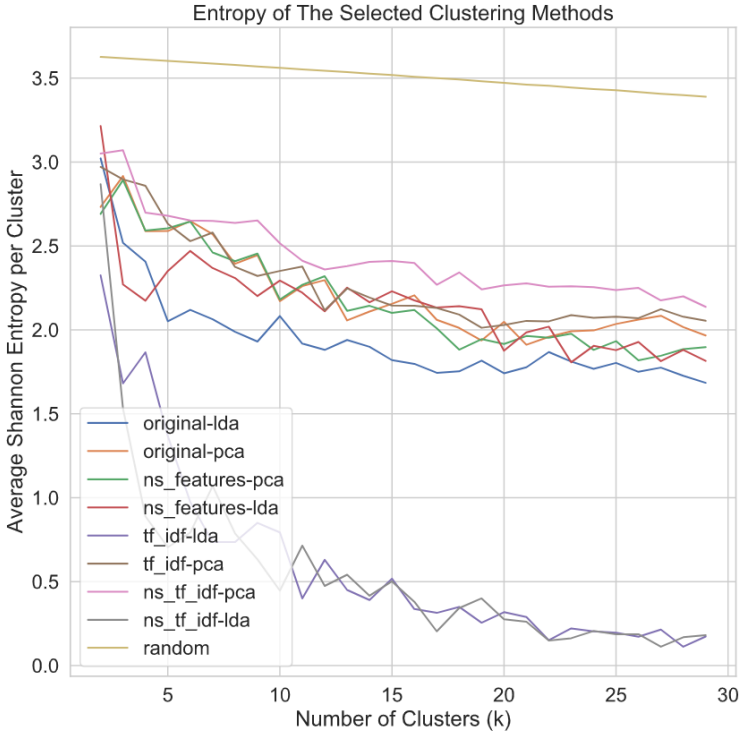

Figure 5: Figure showing the Shannon entropy for the selected clustering methods and values of k, the number of clusters.

Conclusion & Lessons Learned

In this paper we explored the emerging field of *Author Identification*. We found that clustering analysis that used TF-IDF in conjunction with LDA performed the best among the various methods if the Shannon entropy of the author distribution was used to analyze the results. A discussion around the results relieved that our while some groupings were detected, our approach need further refinements to create further separation amongst the clusters.

Many *lessons were learned*, including the technical aspects of developing a web-scraping tool, implementing the algorithms and the pre-processing necessary to do the analyses. We also learned a great deal about the research domain at hand and its many complexities. In addition, the project greatly reinforced many of the essential course-work algorithms related to text analysis on a real-world application. Extensions of our work in this domain area include techniques that increases separation between different authors as well as accounting for the use of *editors* and it's impact on detecting the author's work as well as written work with multiple authors.

References

A. M. Mohsen, N. M. El-Makky and N. Ghanem, *Author Identification Using Deep Learning*, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 898-903. doi: 10.1109/ICMLA.2016.0161

M. Kestemont et. al (2018). *Overview of the Author Identification Task at PAN-2018 Cross-domain Authorship Attribution and Style Change Detection.* Available here http://ceur-ws.org/Vol-2125/invited_paper_2.pdf

N. E. Benzebouchi, N. Azizi, N. E. Hammami, D. Schwab, M. C. E. Khelaifia and M. Aldwairi, *Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector*, 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), Istanbul, Turkey, 2019, pp. 371-376.

Stamatatos, E. (2009) *A Survey of Modern Authorship Attribution Methods.* J. Am. Soc. Inf. Sci., 60: 538-556. doi:10.1002/asi.21001

Waheed Anwar, Imran Sarwar Bajwa and Shabana Ramzan (2019). *Design and Implementation of a Machine Learning-Based Authorship Identification Model.* Available here https://www.hindawi.com/journals/sp/2019/9431073/

O. Halvani, M. Steinebach, & R. Zimmermann (2013). *Authorship Verification via k-Nearest Neighbor Estimation.* Available here http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-HalvaniEt2013.pdf

Appendix

List of Features

- flesch_reading_ease
- smog_index
- flesch_kincaid_grade
- coleman_liau_index
- automated_readability_index
- dale_chall_readability_score
- difficult_words
- linsear_write_formula
- gunning_fog
- text_standard
- mean_syllable_per_word
- ratio_unique_words
- ratio_difficult_words

Author Contributions

We emphasized relying on each team member's individual strengths while also making sure each person was exposed to learning new technical tasks. We saw the work as being split fairly (a 50/50 split) and communicated regularly about ideas and solutions to apply to our project. Overall, it was a great experience and we both learned a great deal.