Authorship Attribution with GatorCAAT: Data Collection and Feature Extraction

Rahul Alapati  
Department of Computer Science and Software Engineering  
Auburn UniversityAuburn, USA  
rza0037@auburn.edu

Sutanu Bhattacharya  
Department of Computer Science and Software EngineeringAuburn UniversityAuburn, USA  
szb0134@auburn.edu

Suraj Raymond D’Souza  
Department of Computer Science and Software EngineeringAuburn UniversityAuburn, USA  
szd0075@auburn.edu

*Abstract*— Authorship Attribution is the science of identifying the author from the characteristics of articles written by that author. In this paper, we discuss about data collection and feature extraction required for the development of GatorCAAT. With GatorCAAT, a user can determine authors based on their writing style.

A unigram is an n-gram of size 1. An n-gram is a continuous sequence of n items from a given sample of text or speech. In this paper, we discuss about a simple method to generate character unigrams.

Keywords—Authorship Attribution, GatorCAAT, Data Collection, Character Unigram, Feature Extraction

# Introduction

Authorship attribution (AA) is the process of determining the writer of an article, given a collection of articles whose authorship is known [1]. The fast growing internet usage across the world, has led to a tremendous increase in the amount of anonymous information, thereby increasing the need for efficient Authorship Attribution systems. Authorship Attribution can be used in plagiarism detection, identifying the author of inappropriate communications and resolving historical questions of unclear or disputed authorship [1].

An n-gram is a n-character slice of a longer string. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of (n-1)-order Markov Model. n-gram models are widely used in probability, communication theory, computational linguistics, computational biology and data compression [2].

The organization of the paper is as follows: In section 2, we briefly mention about the different newspapers and authors used in data collection. In section 3, we explain the methods used in character unigram feature extraction. Section 4 ends the paper with discussion on the breakdown of work among the three authors.

# Newspapers and Authors

As a part of the data collection process, we copy pasted the Football Game Recaps of Florida Gators for the first 3 weeks, from 7 daily newspapers within the state of Florida written by 7 different sports writers.

The 7 daily newspapers were identified from a set of 100 daily and weekly newspapers published within the state of Florida [3]. The selection was made on the basis of the continuity of the Game Recaps over a period of 3 weeks, the uniqueness and quality of the article and its author.

The 7 daily newspapers used for data collection are:

1. St. Augustine Record published in St. Augustine, Florida. [4]
2. Ocala Star Banner published in Ocala, Florida. [5]
3. Independent Florida Alligator published in the University of Florida. [6]
4. Orlando Sentinel published in Orlando, Florida and the Central Florida region. [7]
5. Tampa Bay Times published in St. Petersburg, Florida. [8]
6. Palm Beach Post published in Palm Beach County in South Florida and the parts of Treasure Coast. [9]
7. WPLG Local 10, an online news website based out of Miami, Florida. [10]

The 7 unique sports writers identified as a part of this process are:

1. Mark Long (AP Sports Writer)
2. Edgar Thompson (AP Sports Writer)
3. Robbie Andreu
4. Mark Stine
5. Jake Dreilinger
6. Matt Baker
7. Pat Dooley

Some of the challenges faced during this process are:

1. Most of the daily and weekly newspapers in Florida are a part of The USA Today Network and hence all of them publish the same Game Recap written by a single Associated Press author.
2. Due to the availability of good college football teams like FSU, UCF, University of Miami and University of North Florida in the state of Florida, coverage of the football games gets divided over all the newspapers. Also, about 15% of the newspapers within Florida exclusively cover the football games of a particular school, thereby ignoring the football games of Florida Gators.

The above mentioned challenges have led to the shortage of unique game recaps written by unique sports writers.

# Feature Extraction

After acquiring game recaps, another fundamental step is to find distinctive features from the articles. In Authorship Attribution, selection of features that define the characteristics of an author is equally important to the data collection process. There are dozens of possible features for authorship attribution that have been proposed in the past, for training state of the art algorithms. Good features for authorship attribution are the ones that are able to capture the distinctive aspects of author’s writing style and are consistent even when the author is writing on different subjects. For GatorCAAT, we extracted character unigram feature vectors from the articles.

Large ranges of character n-grams contain characteristic information about the writing style of an author [11]. In this paper, we discuss about the character unigram feature extractor we have developed for GatorCAAT. Our character unigram feature extractor takes as input a directory consisting of the writing samples and provides the user with 2 files, one consisting of the raw feature vectors and the other consisting of the normalized feature vectors.

The raw feature vector consists of the term frequencies for the 95 printable ASCII characters (ranging from 32 to 126) [12] in the writing samples. The normalized feature vectors are obtained by dividing the term frequencies in the raw feature vectors by the magnitude of the raw feature vector. The magnitude of the raw feature vector is the Euclidean distance between itself and the origin, as stated in equation 1.

where, rv is a raw feature vector.

## CASIS-25 Dataset

The provided CASIS-25 Dataset consists of 100 writing samples written by 25 authors, i.e. 4 writing samples per author. We used our feature extractor to extract raw and normalized feature vectors from the 100 writing samples. The following is the raw feature vector extracted from the 1st writing sample of author 1001:

1001,113.0,0.0,0.0,0.0,0.0,0.0,0.0,2.0,1.0,1.0,0.0,0.0,3.0,4.0,17.0,4.0,7.0,4.0,1.0,1.0,1.0,2.0,0.0,0.0,3.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,5.0,1.0,0.0,0.0,0.0,0.0,0.0,3.0,4.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,4.0,1.0,0.0,1.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,66.0,6.0,19.0,20.0,63.0,8.0,15.0,32.0,42.0,0.0,2.0,18.0,21.0,47.0,39.0,13.0,0.0,41.0,36.0,59.0,20.0,7.0,16.0,2.0,14.0,0.0,0.0,0.0,0.0,0.0

The first element is the label, which denotes the author name. Then follow the term frequencies of 95 printable ASCII characters. For example, there are 113 spaces in the 1st writing sample of author 1001.

The following is the normalized feature vector for the 1st writing sample of author 1001:

1001,0.583607310325,0.0,0.0,0.0,0.0,0.0,0.0,0.0103293329261,0.00516466646305,0.00516466646305,0.0,0.0,0.0154939993892,0.0206586658522, …

The first element is the label, which denotes the author name. Then follow the normalized features obtained by dividing the 95 term frequencies with the magnitude of the raw feature vector. Also, we have compared our feature vectors with those provided as a part of the CASIS-25 dataset and our feature vectors exactly match with the provided feature vectors.

## SEC Sports Writers Dataset

The SEC Sports Writers Dataset consists of 21 writing samples from 7 different newspapers, collected over a period of 3 weeks, written by 7 different authors. When passed as an input to our feature extractor, a file consisting of raw feature vectors and a file consisting of normalized feature vectors were generated. The following is the raw feature vector extracted from the Week 2 Game Recap (Florida v/s Kentucky), published in Tampa Bay Times by Matt Baker:

Tampa-Bay-Times\_Baker,722.0,0.0,16.0,0.0,0.0,0.0,0.0,18.0,5.0,5.0,0.0,0.0,34.0,19.0,55.0,0.0,10.0,29.0,13.0,10.0,6.0,7.0,6.0,6.0,3.0,3.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,6.0,3.0,2.0,3.0,3.0,17.0,7.0,3.0,9.0,0.0,13.0,2.0,13.0,5.0,1.0,1.0,0.0,0.0,13.0,21.0,11.0,0.0,10.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,274.0,47.0,91.0,140.0,395.0,77.0,49.0,173.0,174.0,2.0,57.0,127.0,57.0,215.0,220.0,53.0,8.0,184.0,203.0,301.0,91.0,32.0,65.0,5.0,61.0,0.0,0.0,0.0,0.0,0.0

The first element is the label which consists of the newspaper and author’s last name. Then follow the term frequencies of 95 printable ASCII characters.

The following is the normalized feature vector for the Week 2 Game Recap (Florida v/s Kentucky), published in Tampa Bay Times by Matt Baker:

Tampa-Bay-Times\_Baker,0.669686493302,0.0,0.0148406979125,0.0,0.0,0.0,0.0,0.0166957851516,0.00463771809766,0.00463771809766,0.0,0.0,0.0315364830641,0.0176233287711,0.0510148990743,0.0,0.00927543619532,0.0268987649664,0.0120580670539,0.00927543619532,0.00556526171719,0.00649280533673,0.00556526171719,0.00556526171719, …

We manually cross verified some raw feature vectors with the character counts in their writing samples.

## Dataset Statistics

The CASIS-25 Dataset has an average character count of 1368 characters, an average word count of 278 words and an average line count of 8 lines.

Table 1 consists of the average number of characters, words and lines in SEC Sports Writers Dataset per week.

Table 1: SEC Sports Writers Dataset Statistics per week

|  |  |  |  |
| --- | --- | --- | --- |
| Week | Avg Chars | Avg Words | Avg Lines |
| Week\_1 | 3492 | 731 | 25 |
| Week\_2 | 3638 | 776 | 29 |
| Week\_3 | 3213 | 682 | 23 |

A screenshot of a cell phone

Description generated with very high confidence

Figure 1: SEC Sports Writers Dataset Statistics over 3 weeks

Figure 1 shows the change in the average number of characters, words and lines as the dataset grows over 3 weeks. A slight increase in the averages can be observed when week 2’s game recaps are added to the exiting dataset and a notable decrease can be observed after the week 3’s game recaps are added. We would like to investigate this trend and its effect on the accuracy of training algorithms over the 12 weeks.

# Breakdown of Work

Rahul was responsible for the collection of game recaps from newspapers. He was also responsible for implementing the character unigram feature extractor.

Sutanu was responsible for the analysis of datasets. He implemented a program to calculate the average number of characters, words and lines in the CASIS-25 and SEC Sports Writers Dataset.

Suraj was responsible for validating the raw and normalized feature vectors generated by our feature extractor.

##### References

1. I. N. Bozkurt, O. Baglioglu and E. Uyar, Authorship attribution, 22nd international symposium on computer and information sciences, Ankara, 2007, pp. 1-5.
2. https://en.wikipedia.org/wiki/N-gram#.
3. https://en.wikipedia.org/wiki/List\_of\_newspapers\_in\_Florida#Daily\_and\_weekly\_newspapers\_(currently\_published).
4. http://www.staugustine.com/.
5. http://www.ocala.com/.
6. https://www.alligator.org/.
7. https://www.orlandosentinel.com/.
8. http://www.tampabay.com/.
9. https://www.palmbeachpost.com/.
10. https://www.local10.com/.
11. Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling, The Power of Character Ngrams in Native Language Identification, 2017, In Proceedings of the 12th Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics, Copenhagen, Denmark.
12. https://en.wikipedia.org/wiki/ASCII#Printable\_characters.