**Abstract**

*Through Feature Extraction, various statistics found throughout data sets have proven to be useful. Finding feature vectors that produce patterns help create an understanding of authorship attribution. Finding the feature vectors is simply the first step in recreating authors’ articles to create a sense of anonymity. Along with this possibility, an unknown piece of work can be identified as a certain author’s work by using the patterns created by the feature vectors. This paper outlines the first step in completing the longer process of Authorship Attribution.*

**Introduction**

Author Attribution is an important topic in the twenty-first century due to the rise of media outlets and the overarching shadow a possible cyber attack poses on the security of our information. It has become imperative we develop systems to strengthen our ability to authenticate authors to prevent from attacks on our information, so to start our progression to a system which can handle this data, we have developed a feature extractor which will store the amount of unigrams (single characters) from a writing sample as a vector. After it stores this vector, we will then normalize the vectors. The unigram feature extractor will serve as a base to a better system in the future which we hope will be able to pick up on unique stylistic qualities to differentiate between authors of a growing set.

This unigram feature extractor will give us a statistic for how often each author will use a specific character. Then, we can try to compare our feature vectors of each author as the set of samples grow to see if we can find a correlation between the original feature vectors and the new feature vectors each week.

**Newspaper and Authors**

For our work we created a data set that consists of six separate newspapers and six different authors. Therefore we have twelve articles thus far. As expected, all the articles are very similar due to the fact that they all cover the same events. It has even gone so far that some newspapers contain the same sports writers since they are within the same state. This actually led to struggles while scraping for articles to add to the data set. Nonetheless, our data set is comprised of various papers from different authors.

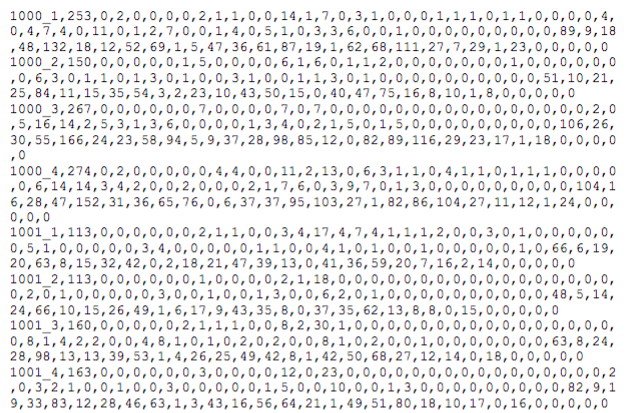
To start off, we have two articles written by Tyler Horka from the Clarion-Ledger newspaper. He covered Mississippi State football for the first two weeks of the season. Followed by the Clarion-Ledger paper, gathering articles from the Starkville Daily News (Joel Coleman and Dave Skretta), The Dispatch (Brett Hudson), and The Reflector. Following these state newspapers, ESPN and SEC is where we went next. Both news outlets relied on the Associated Press to give the recap of games so it will be very interesting to see how this data set evolves over the course of this football season.

As stated above, it was difficult to find more articles due to the watered-down nature of sports reporting. With this said, our data-set of articles is wide-ranging and extremely useful.

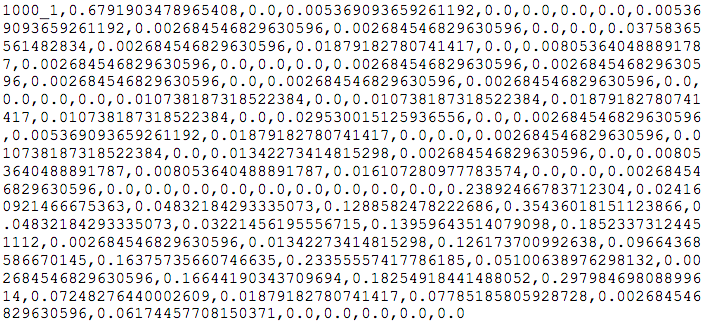
**Feature Extraction**

We wrote our unigram feature extractor in Python, and we gave it two datasets to traverse, our CASIS-25 dataset and our SEC Sports Writer Dataset (Mississippi State). It counts the number of 95 different characters from each article in the dataset and stores the count for each one in a vector. It appends each vector from each sample in each dataset to a file, so there is a different file for each dataset with all the vectors of every sample. We also create another vector from the first we made, which is the starting vector normalized, and we append the normalized vectors to a different file.

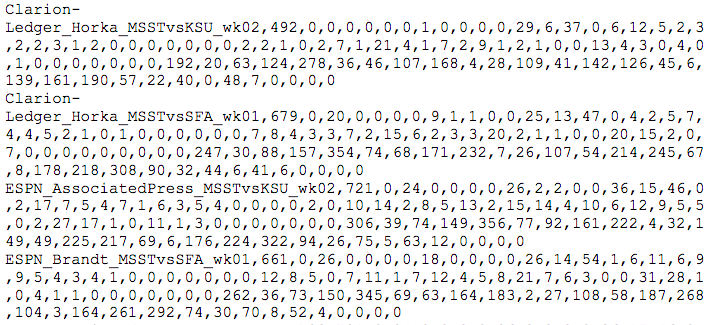
CASIS-25 raw vector sample:



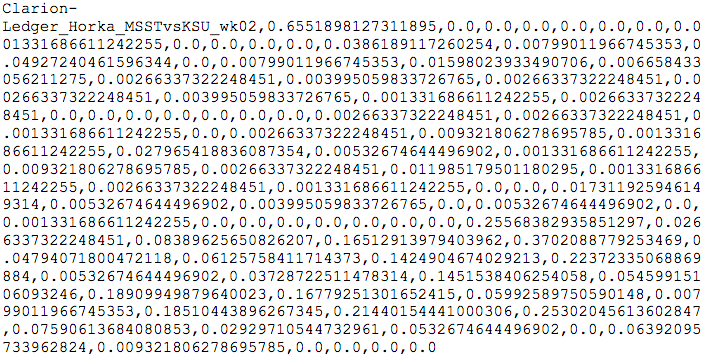
CASIS-25 normalized vector sample:



SEC raw vector sample:



SEC normalized vector sample:



As you can see, the normalized vectors give us a probability from [0 to 1] of each character. We then calculated the average amount of characters, words, and sentences of each dataset.

SEC Dataset:

Avg. # of Characters: 3752

Avg. # of Words: 654

Avg. # of Sentences: 43

CASIS-25 Dataset:

Avg. # of Characters: 1640

Avg # of Words: 278

Avg # of Sentences: 14

**Breakdown of Work**

**Sarp Aykent:** Coded Unigram Feature Extractor and dataset averages calculator

**Blake Schilleci:** Gathered sports articles and wrote abstract, newspapers/articles, and references sections of paper.

**Jordan Cox:** Wrote introduction, feature extraction, and breakdown of work sections of paper and made final revisions of paper.

**References**

[1] Associated Press. “No. 18 Mississippi State tops SFA 63-6 in Moorhead's debut.” *ESPN,* 1 Sept. 2018. [Online]. [www.espn.com](http://www.espn.com/).

[2] Associated Press. “Fitzgerald leads No. 18 Miss St to 31-10 rout of K-State.” *ESPN,* 8 Sept. 2018. [Online]. [www.espn.com](http://www.espn.com/).

[3] Associated Press. “No. 18 Mississippi State breezes by SF Austin.” *SEC Sports,* 8 Sept. 2018. [Online]. www.secsports.com.

[4] Associated Press. “No. 18 Mississippi State stifles Kansas State 31-10.” *SEC Sports,* 8 Sept. 2018. [Online]. www.secsports.com.

[5] Cloud, Hunter. “Defensive Dog Collars, Slam Dunks, are all part of a culture change.” *The Reflector,*12 Sept. 2018. [Online]. www.reflector-online.com.

[6] Cloud, Hunter. “Bulldog offense lifts off, defense smothers in first win for Moorhead.” *The Reflector,*2 Sept. 2018. [Online]. www.reflector-online.com.

[7] Coleman, Joel. “Bulldogs Becoming Disruptive on Defense.” *Bulldogs Becoming Disruptive on Defense | Starkville Daily News*, 2 Sept. 2018. [Online]. Starkvilledailynews.com.

[8] Horka, Tyler. “Running Back Kylin Hill Leads Mississippi State to 31-10 Victory over Kansas State.” *The Clarion Ledger*, The Clarion-Ledger, 8 Sept. 2018. [Online]. www.clarionledger.com.

[9] Horka, Tyler. “Thompson, Hill Lead Mississippi State to 63-6 Stomping of Stephen F. Austin.” *The Clarion Ledger*, The Clarion-Ledger, 2 Sept. 2018. [Online]. www.clarionledger.com.

[10] Hudson, Brett. “MSU's Defense Shows Versatility in Dominating Kansas State.” *The Commercial Dispatch*, 10 Sept. 2018. [Online]. www.cdispatch.com.

[11] Hudson, Brett. “Columbus' Hill has rousing start to season.” *The Commercial Dispatch*, 2 Sept. 2018. [Online]. www.cdispatch.com.

[12] Skretta, Dave. “Thrill from Hill: Sophomore Running Back Produces in MSU Victory.” *Bulldogs Becoming Disruptive on Defense | Starkville Daily News*, 8 Sept. 2018. [Online]. starkvilledailynews.com.