Duquesne University
Duquesne Scholarship Collection

Electronic Theses and Dissertations

Fall 2011

Best Practices in Authorship Attribution of English Essays

Darren Vescovi

Follow this and additional works at: https://dsc.duq.edu/etd

Recommended Citation
Vescovi, D. (2011). Best Practices in Authorship Attribution of English Essays (Master's thesis, Duquesne University). Retrieved from https://dsc.duq.edu/etd/1310

This Immediate Access is brought to you for free and open access by Duquesne Scholarship Collection. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Duquesne Scholarship Collection. For more information, please contact phillipsg@duq.edu.

BEST PRACTICES IN AUTHORSHIP ATTRIBUTION OF ENGLISH ESSAYS

A Thesis

Submitted to the McAnulty College and Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Masters of Science in Computational Mathematics

By

Darren M. Vescovi

December 2011

Darren M. Vescovi

Best Practices in Authorship Attribution of English Essays

Master of Science in Computational Mathematics

Department of Mathematics & Computer Science
Duquesne University, Pittsburgh, PA, USA

November 30, 2011

APPROVED_____
Patrick Juola Ph.D., Associate Professor
Department of Mathematics & Computer Science

APPROVED_____
John Kern, Ph.D., Associate Professor
Department of Mathematics & Computer Science

APPROVED_____
Donald Simon, Ph.D., Director of Graduate Study
Department of Mathematics & Computer Science

APPROVED_____
James Swindal, Ph.D., Dean
McAnulty College and Graduate School of Liberal Arts

ABSTRACT

BEST PRACTICES IN AUTHORSHIP ATTRIBUTION OF ENGLISH ESSAYS

By

Darren M. Vescovi

December 2011

Thesis supervised by Patrick Juola

Logistic regression analysis is used to determine the best practices in authorship attribution for English essays, specifically examining the methods available in JGAAP version 4.3 and their performance on problem A of the AAAC corpus. Best practices were determined by ranking the logistic regression coefficient estimates and odds ratios for the set of predictor variables.

Contents

1 Background **1**
- 1.1 Authorship Attribution . . . 1
- 1.2 JGAAP . . . 3
- 1.3 Corpora . . . 4

2 Search for Best Practices **6**
- 2.1 Logistic Regression . . . 6
- 2.2 Odds Ratios Using Logistic Regression . . . 9

3 Methods and Materials **11**
- 3.1 Data Collection . . . 12
- 3.2 SAS Logistic Regression . . . 15

4 Results **17**
- 4.1 Canonicizers . . . 17
- 4.2 Event Sets . . . 18
- 4.3 Event Cullers . . . 21
- 4.4 Analysis Drivers . . . 24

5 Discussion **26**
- 5.1 Canonicizers . . . 26
- 5.2 Event Sets . . . 27
- 5.3 Event Cullers . . . 28
- 5.4 Analysis Drivers . . . 29

6 Future Work **31**

7 Conclusion **32**

A Distance Formulas **35**

List of Tables

3.1 List of Canonicizers 13
3.2 List of Event Sets 14
3.3 List of Event Cullers 15
3.4 List of Analysis Drivers 15

4.1 Maximum Likelihood Estimates: Canonicizers 18
4.2 Odds Ratio Estimates: Canonicizers 18
4.3 Maximum Likelihood Estimates: Event Sets 19
4.4 Odds Ratio Estimates: Event Sets 20
4.5 Maximum Likelihood Estimates: Event Cullers 22
4.6 Odds Ratio Estimates: Event Cullers 23
4.7 Maximum Likelihood Estimates: Analysis Drivers 24
4.8 Odds Ratio Estimates: Analysis Drivers 25

Chapter 1

Background

1.1 Authorship Attribution

The main assumption of authorship attribution studies is that an author has a distinct set of characteristics in his or her writing style that produces an authorial fingerprint detectable in their writings. A term synonymous with authorship attribution is that of stylometry. Juola [5] goes on to describe that an authorial fingerprint makes theoretical sense because each author must learn language on their own, producing a specific style for each author. By using the main assumption of authorship attribution studies, [4] lays out two specific authorship attribution problems. The first is a closed class problem: given a sample document believed to be written by one of a set of authors, determine the author who wrote the document. The second is an open class problem: given a sample document believed to be written by one (or more) authors determine which, if any, author wrote the document. One can notice the increase in difficulty for both the open class and closed class authorship attribution problems as the set of authors to choose from grows. This is based on the assumption that an author's style can be defined by a set of measurable patterns that are unique to an author [2]. Holmes [2] also notes that no methodology has yet been discovered

which projects style better than methods based on lexical measures.

Many authorship attribution techniques use simple univariate statistics such as average word length, vocabulary richness, distributions of syllables, sentence length, or parts of speech [2]. According to [2] numerous statistical models and tests have been examined to assign authorship to disputed texts with varying degrees of success. But with the above assumption and description of authorship attribution problems one can see that this authorial fingerprint could be quite complex and not well represented by the univariate techniques described above.

With the emergence of multivariate techniques tuned to distributional features such as cluster analysis, factor analysis, and discriminant analysis, researchers hope to be able to reliably describe the authorial fingerprint left by an author. With the lack of a clear underlying mathematical model to describe ones authorial fingerprint, [2] describes the use of principal component analysis to create new ranked components, linear combinations of the set of original features used in multivariate analysis, and assign authorship based on these new features with the hope that most of the variation in the original data is from the first few components. Although this makes sense that an authorial fingerprint should be able to be described by a set of summary statistics, there is still an unsettling amount of discontinuity between scholars in the search for best practices in authorship attribution. Rudman [7] describes numerous problems with contemporary authorship attribution techniques leading to the main theme of a lack of consensus on accepted or correct techniques. Rudman cites numerous occasions in which a previously published work comes under fire from another researcher showing that their results are flawed or biased. One can see how this can be damaging to the field of authorship attribution by the variation of so-called proven results. One such example is that of Nuemann's reliance on discriminant analysis,

described as problematic by Mealand [7]. Rudman [7] also mentions the lack of long-term devoted researchers to the field of authorship attribution, with the majority of researchers conducting experiments on a problem specific level. The lack of such researchers hinders the advancement of the field of authorship attribution and this lack of commitment in combination with the inconsistency of research results makes authorship attribution easily discredited in situations demanding a result be irrefutably true beyond a reasonable doubt. Hence the need for determining the best practices in authorship attribution, as well as suggestion for future research in specific authorship attribution methods described below.

1.2 JGAAP

There is a vast array of methods for authorship attribution presented by numerous researchers with marginal evidence to support their claim of validity across a wide range of document types. Rudman [7] suggests thousands. Juola's [3] Java Graphical Authorship Attribution Project, JGAAP, provides a number of methods to do authorship attribution. JGAAP consists of a three phase modular design summarized below and described in detail at [5] [4]:

- Canonicization: Standardizing realization of the same event that the computer would treat as separate events. For example this may include changing consecutive whitespace characters to single whitespace characters, changing all characters to lower case, removing page numbers, etc.

- Event set determination: Partitioning the document into disjoint events such as words, word lengths, characters, character bi-grams, etc. At this time uninformative events can be discarded.

- Statistical Inference: The remaining events can be subjected to a variety of inferential statistics, ranging from simple analysis of event distributions through complex pattern based analysis. Authorship is then assigned as a result of these inferential statistics.

JGAAP also provides the option of limiting the event space (Event Culler) to the fifty most common events, fifty least common events, and the extreme events (events appearing in all documents), thus effectively giving four phases since version 4.3.

The authorship attribution process in JGAAP can be described as follows. First the set of unknown and known(training) documents are loaded into JGAAP. From here the documents are then subject to canonicization to put them into a standard form for analysis. After canonicization is complete the set of events is determined and the document is partitioned into events, such as character bi-grams, two consecutive characters. Once the event set is created the user has the option to limit the number of events to be analyzed by applying an event culler, such as having the most common events or events only appearing in all documents(extreme culling). The event set then moves to an analysis driver to determine the authorship of the document. A typical distance-based analysis method determines the distance between the unknown document and each training document, and then assigns authorship to the unknown document with the notion that the author of the closest document, distance wise, authored the unknown document.

1.3 Corpora

With the emergence of a software package that pulls together a variety of authorship attribution techniques, so does the need for a suitable corpus to test how well each

method performs on typical authorship attribution problems. Juola's 2004 Ad-hoc Authorship Attribution Competition (AAAC) [5] [4] established a set of problems to serve as an empirical test bed for comparative analysis of authorship attribution methods. From this standardized corpus a researcher can now test how well his or her authorship attribution method performs on a set of typical closed and open class authorship attribution problems. The AAAC includes 13 problems across a variety of styles, languages, lengths, and genres mostly gathered from the Web. Unfortunately, the AAAC is too small to adequately distinguish between good and bad methods of authorship attribution across the entire scope of authorship attribution problems. Consider the Brennan-Greenstadt Obfuscation corpus [1], purposely constructed as an example of authors deliberately trying to mask his or her authorial style though imitation or obfuscation, not represented in the AAAC. The AAAC will serve as a suitable corpus for the purpose of this research study, because it offers a suitable representation of an authorship attribution problem for English essays.

Chapter 2

Search for Best Practices

With the ease of conducting such large scale experiments on a standardized corpus of authorship attribution problems by using JGAAP, or similar software, one can see how a large set of data on the attribution of a specific document can be produced. However, simple statistical analysis of this data can produce results on the combination of a different set of contributing variables for the attribution of a document; but it provides little information on the interaction of those specific contributing variables. By combining the abundant data on the performance of different combinations of variables for authorship attribution with logistic regression analysis, a researcher can weed through the less valuable variables and move in the direction of a more reliable definition of the best practices for the authorship attribution.

2.1 Logistic Regression

Logistic regression is a commonly used method to classify dichotomous outcome variables such as, whether a document was correctly attributed to a specific author. The logistic model is a mathematical model that can relate several predictor variables x_1,x_2 $\ldots$ x_n to a dichotomous dependent variable y typically coded as $\{0, 1\}$ [6]. The

logistic model describes the expected value of y as,

$$E(y) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}, \tag{2.1}$$

It then follows from basic statistical principles that,

$$p(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}, \tag{2.2}$$

for (0, 1) random variables [6]. The logistic function, that is

$$f(z) = \frac{1}{1 + e^{-z}}, \quad \text{where } z = (\beta_0 + \sum_{i=1}^{n} \beta_i x_i), \tag{2.3}$$

is well suited to model a probability because the function is an increasing function with range $(0, 1)$ and domain of $(-\infty, \infty)$.

Like most regression models logistic regression hinges on the maximum likelihood estimates of the regression coefficients. The term maximum likelihood refers to the estimation of population parameters by using a very general algorithm. If we define the likelihood function $L(\theta) = p(y|\theta)$ as the probability of the observed data given a set of parameters θ, then $L(\theta)$ gives the probability distribution of the observed data, y as a function of the unknown parameters θ. ML addresses the problem of finding $\hat{\theta}$, or the value of θ that maximizes $L(\theta)$. [6] states that $\hat{\theta}$ is the numerical value that agrees the most with the observed data in the sense of providing the largest possible value for the probability $L(\theta)$. From this a researcher can then use the estimated parameters $\hat{\theta}$ to make inferences on the true parameters θ. One of the main goals of regression analysis is to test the null hypothesis that the regression coefficient $\beta_i = 0$. To do this the researcher uses the Wald statistic. When a large data set and an appropriate likelihood function is used the ML estimator is essentially unbiased, has

a small variance, and is approximately normally distributed, thus allowing us to use the standard normal statistic called the Wald statistic,

$$Z = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{var}(\hat{\beta}_i)}} = \frac{\hat{\beta}_i}{\sqrt{\widehat{var}(\hat{\beta}_i)}}. \tag{2.4}$$

Alternatively a chi-square statistic can be used since Z^2 is χ^2 with one degree of freedom [6].

Logistic regression modeling relies on maximizing one of two likelihood functions, the conditional and the unconditional likelihood function. Before we discuss the likelihood functions we must first talk about the statistical distribution of the outcome variable. The outcome variable is a Bernoulli random variable taking value 1 with probability θ and value 0 with probability $(1-\theta)$ with the simple discrete probability distribution $p(y|\theta) = \theta^y(1-\theta)^{(1-y)} \quad y = 0, 1$. For a study of n samples the Bernoulli random variable for the ith sample is $p(y_i|\theta_i) = (\theta_i)^{y_i}(1-\theta_i)^{(1-y_i)} \quad y_i = 0, 1$. Consider a sample of $y_1, y_2, \ldots, y_n$ mutually independent observations. The likelihood function is obtained from the product of the marginal distributions for the y_i's and so,

$$L(Y|\theta) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{(1-y_i)}. \tag{2.5}$$

By using the fact that

$$\theta_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{n} \beta_j x_j)}}, i = 1, 2, \ldots, n, \tag{2.6}$$

where $\beta_1, \beta_2, \ldots, \beta_k$ are the unknown regression coefficients to be estimated, and some

algebraic manipulation we can write the likelihood function as the following,

$$L(Y|\beta) = \frac{\prod_{i \ni y_i=1}[e^{-(\beta_0+\sum_{j=1}^{n}\beta_j x_{ij})}]}{\prod_{i=1}^{n}[1+e^{-(\beta_0+\sum_{j=1}^{n}\beta_j x_{ij})}]} \tag{2.7}$$

The above is also known as the unconditional likelihood function, and refers to the unconditional probability of obtaining the particular set of data under consideration [6]. Alternatively one can use the conditional likelihood function which refers to the conditional probability of obtaining the data configuration actually observed, given all possible configurations. The conditional likelihood function is used when the data set is "small" and also when dealing with matched data, but this will not be the case for this research study.

2.2 Odds Ratios Using Logistic Regression

Logistic regression also provides a useful statistic that allows a researcher to compare two or more groups with respect to the outcome variable. The odds of an event is the ratio of the probability of an event occurring divided by the probability that same event will not occur. Hence the odds of event D happening is,

$$odds(D) = \frac{pr(D)}{1-pr(D)}. \tag{2.8}$$

For example, an odds of one-third can be interpreted as the probability of event D occurring, is one-third the probability of D not happening. An odds ratio (**OR**) is the ratio of two odds, that is

$$\mathbf{OR}_{A\ vs.\ B} = \frac{\frac{pr(D_A)}{1-pr(D_A)}}{\frac{pr(D_B)}{1-pr(D_B)}} \tag{2.9}$$

where A and B two groups being studied. Consider the correct authorship of a document as the event occurring, and A denotes using the unify case canonicizer and B denotes using the normalize white space canonicizer. If $\mathbf{OR}_{A\ vs.\ B} = 2$ then the odds of unify case canonicizer correctly attributing the authorship are twice that of the normalize whitespace canonicizer. An $\mathbf{OR}_{A\ vs.\ B} = 1$ means that the odds of the event happening for A is the same as for B.

By using reference cell coding we can estimate the odds ratio of different levels for the variables in the logistic model created. By using $\{0, 1\}$ coding the odds ratio will be, $\mathbf{OR} = e^{\beta_i}$. By comparing multiple variable that can change we obtain an adjusted odds ratio that controls for the other variables in the model. Adjusted odds ratios can be obtained by exponentiating the parameter estimate, β_i, for a $\{0, 1\}$ variable provided there are no interaction terms in the model.

Chapter 3

Methods and Materials

JGAAP 4.3 provides a suitable number of authorship attribution methods to allow for a large scale test on how well these methods predict the authorship of a document correctly. Since JGAAP was designed with a three phase modular structure, variables for this model are easily obtained and are: event culler, canonicizer, event set, and analysis driver. For this research study only problem A of the AAAC will be used to test the methods selected from JGAAP, with the outcome variable coded as 1 for a correct attribution and 0 for an incorrect attribution. Analysis will be performed on an unknown document, comparing it to a set of three training documents for which each of 13 authors are believed to have written the unknown document. Problem A is representative of a real life authorship attribution of student essays, not exceeding 1200 words, gathered in a first year writing class [4]. The documents will be subject to a number of authorship attribution methods provided by JGAAP, and the result of the attribution will be recorded as well as what event cullers, canonicizers, event set, and analysis method used.

3.1 Data Collection

Logistic regression requires a dichotomous response variable to be described by a number of factors. The collection of this data needs to be done in such a manor as to not create an underlying bias in the data. The data will be collected from a relational database containing approximately 187 million records of authorship attribution methods consisting of various combinations of event sets, event cullers, canonicizers, and analysis drivers used on individual unknown document from problem A of the AAAC. The data includes whether or not the method attributed the correct author to the unknown document. Correct attribution for each document will be the dependent variable in this study.

Collection of the data consisted of randomly selecting 10,000 different attribution methods for each of the 13 unknown documents. The 10,000 methods were selected (of the 187 million) without replacement with respect to each unknown document. This was done to control for the effect in which the unknown document has on the outcome variable. JGAAP allows one to use a combination of multiple canonicizers, however canonicization was limited to only one canonicizer and not a combination of different canonicizers because the number of combinations of canonicizers would be too large to study effectively. Similarly, JGAAP allows you to use more than one event culler on the event sets, with the order of application determining the resulting event set. Hence, applying extreme culling, then least common culling will produce a different event set than the result of applying least common culling, and then extreme culling would produces. Since the order of application is of particular interest and the number of event cullers is small, we will study each permutation of event cullers, with each permutation being treated as a separate level in the event culler variable. A description of the variables and levels is provided below.

Table 3.1: List of Canonicizers

Canonicizer	**Description**
Normalize Whitespace	Changes all consecutive white space characters to a single space
Punctuation separator	Put a single space before and after each punctuation mark, to keep them separate from adjacent words.
Strip Alpha-Numeric	Removes all letters and numbers leaving only punctuation and symbols
Strip Numbers	Removes all numbers from the documents
Strip Punctuation	Removes all punctuation from the documents
Unify Case	Changes all letters to lowercase

I chose to only look at authorship attribution methods that used a single canonicizer, so each canonicizer is its own level, for a total of six. This is shown in Table 3.1.

Only one event set can be used at a time in JGAAP, so there will be twenty-six levels of event drivers. This is shown in Table 3.2.

A permutation of the event cullers, in Table 3.3, can be used with the order of application determining the resulting event set. Therefore, I chose to treat each permutation as a separate level, making a total of 15 levels in the event culler variable.

As with event sets, only one analysis driver can be used at a time in JGAAP, so there are fifteen levels of analysis drivers. This is shown in Table 3.4.

Reference cell coding was used to determine the individual effects of each level. The reference cell for the canonicizer, event set, event culler, and analysis driver variable are normalize white space, 2 to 3 letter words, least common events, and Canberra

Table 3.2: List of Event Sets

Event Set	Description
2-3 Letter Words	Two to three letter words
2-4 Letter Words	Two to four letter words
3-4 Letter Words	Three to four letter words
Character Bi-grams	Two consecutive characters
Character Tri-grams	Three consecutive characters
Characters	Individual characters
Coarse POS Tagger	A simplification of the normal part of speech tagger
Dis Legomena	Words appearing only twice per document
First Word In Sentence	The first word in each sentence
Hapax Legomena	Words appearing only once per document
Hapax/Dis Legomena	Words appearing only once or twice per document
MW Function Words	Function Words from Mosteller-Wallace
POS	Parts of Speech
POS Bi-grams	Two consecutive parts of speech
Sentence Length	The number of words in a sentence
Syllables Per Word	The number of syllables in each word
Vowel 2-3 Letter Words	Words beginning with a vowel with two to three letters
Vowel 2-4 Letter Words	Words beginning with a vowel with two to four letters
Vowel 3-4 Letter Words	Words beginning with a vowel with three to four letters
Vowel-initial Words	Words beginning with a vowel
Word Bi-grams	Two consecutive words
Word Tri-grams	Three consecutive words
Word Tetra-grams	Four consecutive words
Word Length	The number of letters in a word
Word Stems w/ Irregular	Word stems with special handling of irregular nouns and verbs
Words	Single words

distance, respectively. Using reference cell coding will allow us to determine adjusted odds ratios, which control for the other variables, for each level compared to the reference cell. By using the fact, the adjusted odds ratio for predictor x_i is computed as $\mathbf{OR} = e^{\beta_i}$ for $\{0, 1\}$ coding. The odds ratio for reference cell coding will give the odds of a predicator correctly attributing the document versus the reference cell correctly attributing the document. This will help in determining the best canonicizer, event

Table 3.3: List of Event Cullers

Event Culler	Description
Extreme Culler	Limits events to those appearing in all documents
Most Common Culler	Limits the events to the 50 most common events in a document
Least Common Culler	Limits the events to the 50 least common events in a document

Table 3.4: List of Analysis Drivers

Analysis Driver	Description
Canberra Distance	See Appendix A
Cosine Distance	See Appendix A
Histogram Distance	Euclidean or L2 Norm
IntersectionDistance	See Appendix A
JW Cross Entropy	Juola-Wyner Cross Entropy
KS Distance	KolmogorovSmirnov test
Kendall Correlation Distance	See Appendix A
Keselj-Weighted Distance	Histogram Distance (L2 Norm) with Keselj-weighting based on overall frequency
Kullback Leibler Distance	Kullback-Leibler divergence
LDA	Linear Discriminent Analysis
LZW Divergence	Lempel-Ziv-Welch Divergence
Levenshtein Distance	See Appendix A
Markov Chain Analysis	First Order Markov Chain Analysis
Naïve Bayes Classifier	Naïve Bayes Probability Model with Maximum A Posterior Rule Analysis
RN Cross Entropy	Ryan-Noecker Cross-Entropy

culler, event set, and analysis driver to use by ranking the β_i's and the adjusted odds ratios.

3.2 SAS Logistic Regression

The LOGISTIC procedure from the SAS statistical software package was used to perform the parameter estimation using the maximum likelihood method and the unconditional likelihood function described above. Forward selection was used to add the event set, canonicizer, event culler, and analysis driver variables to the model

with each variable significantly contributing to the model. The global null hypothesis of $\beta = 0$, or that none of the independent variables in the model are related to the change in probability of the event occurring, was rejected at the standard $\alpha = 0.05$. SAS reported the tests of the local null hypotheses of $\beta_i = 0$, for each individual level within the variable and will be discussed in detail below.

Chapter 4

Results

4.1 Canonicizers

To determine which canonicizers are statistically significant in predicting the authorship of a document, we look at the Wald χ^2 statistic for each predictors coefficient. The null hypothesis of each coefficient is that $\beta_i = 0$. The effect of the predictor on the overall model is shown by the sign and magnitude of a significant coefficient. If the sign of the coefficient is positive then the predictor positively contributes to the overall model and correct authorship of an unknown document with respect to the reference cell.

As shown in the table 4.1 all but one canonicizer, Punctuation Separator, are significant at the $\alpha = 0.05$ level. The canonicizers that positively contribute to the model with respect to the reference cell were Unify Case and Strip Alpha Numeric. Strip Punctuation and Strip Numbers contribute negatively to the model with respect to the reference cell.

Table 4.1: Maximum Likelihood Estimates: Canonicizers

Predictor	Estimate β_i	SE	Wald's Test	p
Unify Case	0.0625	0.0318	3.8546	0.0496
Strip Punctuation	-0.1304	0.0330	15.5878	<.0001
Strip Numbers	-0.0845	0.0328	6.6559	0.0099
Strip AlphaNumeric	0.0848	0.0325	6.8204	0.0090
Punctuation Separator	-0.0132	0.0324	0.1662	0.6835

Table 4.2: Odds Ratio Estimates: Canonicizers

Effect vs Normalize White Space	Point Estimate	95% Wald Confidence Limits
Unify Case	1.065	(1.000, 1.133)
Strip Punctuation	0.878	(0.823, 0.936)
Strip Numbers	0.919	(0.862, 0.980)
Strip AlphaNumeric	1.088	(1.021, 1.160)
Punctuation Seperater	0.987	(0.926, 1.052)

4.2 Event Sets

Similarly to canonicizers, determination of significant event sets was done by examining the Wald χ^2. All but one event set showed significance, ($p < 0.0001$) and contributed positively to the overall model with respect to the reference cell as shown in table 4.3. Syllables Per Word was the only event set to fail to reject the null hypothesis of $\beta = 0$.

Table 4.3: Maximum Likelihood Estimates: Event Sets

Predictor	Estimate β_i	SE	Wald's Test	p
Words	0.5759	0.0770	55.9613	<.0001
Word stems w/ Irregular	0.6218	0.0764	66.3165	<.0001
Word TriGrams	0.3686	0.0831	19.6948	<.0001
Word TetraGrams	0.4216	0.0823	26.2209	<.0001
Word Lengths	0.7175	0.0751	91.2787	<.0001
Word BiGrams	0.4044	0.0818	24.4219	<.0001
Vowel-initial words	0.7758	0.0755	105.6362	<.0001
Vowel 3–4 letter Words	0.3135	0.0822	14.5406	0.0001
Vowel 2–4 letter Words	0.6207	0.0775	64.1093	<.0001
Vowel 2–3 letter Words	0.6704	0.0768	76.2015	<.0001
Syllables Per Word	-0.1357	0.0867	2.4481	0.1177
Sentence Length	0.3822	0.0832	21.1165	<.0001
POS BiGrams	0.7336	0.0754	94.7243	<.0001
POS	0.9997	0.0724	190.7196	<.0001
MW Function Words	0.3493	0.0810	18.5979	<.0001
Hapax/Dis Legomena	0.5022	0.0814	38.0552	<.0001
Hapax Legomena	0.4508	0.0822	30.0506	<.0001
First Word In Sentence	0.5478	0.0837	42.8567	<.0001
Dis Legomena	0.4875	0.0820	35.3565	<.0001
Coarse POS Tagger	0.5044	0.0771	42.8249	<.0001
Characters	0.9091	0.0735	153.1211	<.0001
Character TriGrams	0.7291	0.0756	92.8842	<.0001
Character TetraGrams	0.6528	0.0762	73.4582	<.0001
Character BiGrams	0.7497	0.0753	99.1455	<.0001
3–4 letter Words	0.8131	0.0754	116.1810	<.0001
2–4 letter Words	0.4006	0.0795	25.3787	<.0001

Table 4.4: Odds Ratio Estimates: Event Sets

Effect vs 2–3 letter Words	Point Estimate	95% Wald Confidence Limits
Words	1.779	(1.530, 2.068)
Word stems w/ Irregular	1.862	(1.603, 2.163)
Word TriGrams	1.446	(1.229, 1.701)
Word TetraGrams	1.524	(1.297, 1.791)
Word Lengths	2.049	(1.769, 2.374)
Word BiGrams	1.498	(1.276, 1.759)
Vowel-initial words	2.172	(1.874, 2.519)
Vowel 3–4 letter Words	1.368	(1.165, 1.607)
Vowel 2–4 letter Words	1.860	(1.598, 2.165)
Vowel 2–3 letter Words	1.955	(1.682, 2.273)
Syllables Per Word	0.873	(0.737, 1.035)
Sentence Length	1.465	(1.245, 1.725)
POS BiGrams	2.083	(1.797, 2.414)
POS	2.717	(2.358, 3.132)
MW Function Words	1.418	(1.210, 1.662)
Hapax/Dis Legomena	1.652	(1.409, 1.938)
Hapax Legomena	1.570	(1.336, 1.844)
First Word In Sentence	1.729	(1.468, 2.038)
Dis Legomena	1.628	(1.387, 1.912)
Coarse POS Tagger	1.656	(1.424, 1.926)
Characters	2.482	(2.149, 2.866)
Character TriGrams	2.073	(1.787, 2.404)
Character TetraGrams	1.921	(1.655, 2.230)
Character BiGrams	2.116	(1.826, 2.453)
3–4 letter Words	2.255	(1.945, 2.614)
2–4 letter Words	1.493	(1.277, 1.744)

4.3 Event Cullers

Without loss of generality, event cullers, where examined in a similar fashion to the first two classes of predictor variables discussed previously. Most Common Events and the combination of Most Common Events/Least Common Events rejected the null hypothesis of $\beta_i = 0$ with $p < 0.0001$. Other event cullers, X-treme Culler/Most Common Events, X-treme Culler/Least Common Events, Most Common Events/Extreme Culler, and Most Common Events/Least Common Event/Extreme Culler rejected the null hypothesis with $0.05 > p > 0.0001$. All of the significant event cullers contributed positively to the overall model with respect to the reference cell. Shown in table 4.5.

Table 4.5: Maximum Likelihood Estimates: Event Cullers

Predictor	Estimate β_i	SE	Wald's Test	p
X-treme Culler, Most Common Events, Least Common Events	0.0980	0.0526	3.4751	0.0623
X-treme Culler, Most Common Events	0.1097	0.0522	4.4183	0.0356
X-treme Culler, Least Common Events, Most Common Events	0.0754	0.0527	2.0511	0.1521
X-treme Culler, Least Common Events	0.1297	0.0521	6.1888	0.0129
X-treme Culler	0.0611	0.0528	1.3374	0.2475
Most Common Events, X-treme Culler, Least Common Events	0.0162	0.0533	0.0927	0.7607
Most Common Events, X-treme Culler	0.1052	0.0525	4.0102	0.0452
Most Common Events, Least Common Events, X-treme Culler	0.1036	0.0523	3.9155	0.0478
Most Common Events, Least Common Events	0.3990	0.0499	64.0674	$<.0001$
Most Common Events	0.4276	0.0496	74.1991	$<.0001$
Least Common Events, X-treme Culler, Most Common Events	-0.0393	0.0565	0.4838	0.4867
Least Common Events, X-treme Culler	-0.0415	0.0569	0.5331	0.4653
Least Common Events, Most Common Events,X-treme Culler	0.0084	0.0561	0.0222	0.8816
Least Common Events, Most Common Events	0.0327	0.0534	0.3752	0.5402

Table 4.6: Odds Ratio Estimates: Event Cullers

Effect vs Least Common Events	Point Estimate	95% Wald Confidence Limits
X-treme Culler, Most Common Events, Least Common Events	1.103	(0.995, 1.223)
X-treme Culler, Most Common Events	1.116	(1.007, 1.236)
X-treme Culler, Least Common Events, Most Common Events	1.078	(0.973, 1.196)
X-treme Culler, Least Common Events	1.138	(1.028, 1.261)
X-treme Culler	1.063	(0.958, 1.179)
Most Common Events, X-treme Culler, Least Common Events	1.016	(0.916, 1.128)
Most Common Events, X-treme Culler	1.111	(1.002, 1.231)
Most Common Events, Least Common Events, X-treme Culler	1.109	(1.001, 1.229)
Most Common Events, Least Common Events	1.490	(1.352, 1.643)
Most Common Events	1.534	(1.391, 1.690)
Least Common Events, X-treme Culler, Most Common Events	0.961	(0.861, 1.074)
Least Common Events, X-treme Culler	0.959	(0.858, 1.072)
Least Common Events, Most Common Events, X-treme Culler	1.008	(0.903, 1.126)
Least Common Events, Most Common Events	1.033	(0.931, 1.147)

4.4 Analysis Drivers

All analysis drivers except Manhattan Distance, Keselj-weighted Distance, KS Distance, and Histogram Distance rejected the null hypothesis of $\beta_i = 0$. Kendall Correlation Distance, JW Cross Entropy, Intersection Distance, and Cosine Distance contributed positively to the overall model with respect to the reference cell. While RN Cross Entropy, Nave Bayes Classifier, Markov Chain Analysis, Levenshtein Distance, LZW Divergance, LDA, and Kullback Leibler Distance contributed negatively to the overall model with respect to the reference cell. This is shown in table 4.7.

Table 4.7: Maximum Likelihood Estimates: Analysis Drivers

Predictor	estimate	SE	Wald's Test	p
RN Cross Entropy	-0.1544	0.0496	9.6980	0.0018
Naive Bayes Classifier	-0.8962	0.0600	223.1415	<.0001
Markov Chain Analysis	-1.3393	0.0698	367.9983	<.0001
Manhattan Distance	0.0437	0.0476	0.8427	0.3586
Levenshtein Distance	-0.4446	0.0530	70.2831	<.0001
LZW Divergence	-0.3084	0.0517	35.5929	<.0001
LDA	-0.3541	0.0517	46.8477	<.0001
Kullback Leibler Distance	-0.1003	0.0492	4.1548	0.0415
Keselj-weighted Distance	0.0597	0.0478	1.5574	0.2120
Kendall Correlation Distance	0.5785	0.0559	107.2551	<.0001
KS Distance	0.0565	0.0478	1.3986	0.2370
JW Cross Entropy	0.1526	0.0516	8.7543	0.0031
Intersection Distance	0.1846	0.0514	12.9163	0.0003
Histogram Distance	0.0064	0.0482	0.0178	0.8937
Cosine Distance	0.4889	0.0492	98.5927	<.0001

Table 4.8: Odds Ratio Estimates: Analysis Drivers

Effect vs Canberra Distance	Point Estimate	95% Wald Confidence Limits
RN Cross Entropy	0.857	(0.778, 0.944)
Naive Bayes Classifier	0.408	(0.363, 0.459)
Markov Chain Analysis	0.262	(0.229, 0.300)
Manhattan Distance	1.045	(0.952, 1.147)
Levenshtein Distance	0.641	(0.578, 0.711)
LZW Divergence	0.735	(0.664, 0.813)
LDA	0.702	(0.634, 0.777)
Kullback Leibler Distance	0.905	(0.821, 0.996)
Keselj-weighted Distance	1.062	(0.967, 1.166)
Kendall Correlation Distance	1.783	(1.598, 1.990)
KS Distance	1.058	(0.964, 1.162)
JW Cross Entropy	1.165	(1.053, 1.289)
Intersection Distance	1.203	(1.088, 1.330)
Histogram Distance	1.006	(0.916, 1.106)
Cosine Distance	1.631	(1.481, 1.796)

Chapter 5

Discussion

Examining the magnitude and sign of a parameter estimate, as well as the computed odds ratio, allows one to determine how well a parameter coded with reference cell coding performed compared to the reference cell. The more positive a significant parameter estimate will result in a higher odds ratio, and thus show that the predictor is more likely to predict a desired event. In this case the correct attribution of a document.

5.1 Canonicizers

Unify Case has a parameter estimate of $\beta = 0.0625(p = 0.0496)$ and is significant at $\alpha = 0.05$, allowing us to reject the null hypothesis. Unify Case has an odds ratio of 1.065(Shown in Table 4.2) and shows that changing the case of a document to all lower case is 1.065 times more likely to attribute the correct author to a document than only using Normalize White Space. The standardization of case can take out misleading differences in text that can be introduced into a document by the editor, publisher, the order of appearance in a sentence, etc. Consider the word "start". Depending on the placement of "start" in the sentence two separate realizations will be interpreted by the computer. "Start" when leading off a sentence and "start" when

not. This can produce inaccurate frequencies for events and inaccurately describe the authorial fingerprint for that particular document.

A more interesting result is that of the significance Strip Punctuation with a negative parameter estimate, and the significance of Strip AlphaNumeric with a positive parameter estimate. This suggests that the punctuation that an author uses is important in identifying their authorial fingerprint. For example, an author who uses complex sentences with comma-splices versus an author who chooses to use simple sentences without comma-splices. One can see that the two authors' styles differ with one heavy with commas and the other with periods. The fact that Strip Numbers is significant and the parameter estimate being negative, also reinforces the fact that Strip AlphaNumeric helps in the attribution of a document. Future research should be conducted to determine the extent of how well the author's choice of punctuation describes his or her authorial fingerprint.

Punctuation separator failed to reject the null hypothesis of $\beta_i = 0$. Therefore, we can not say anything about the likelihood of Punctuation separator correctly attributing an unknown document versus using Normalize Whitespace.

5.2 Event Sets

The Event Set that had the highest estimate of β_i was POS with $\beta = 0.9997$ and significant at the $\alpha = 0.05$ level. The odds ratio (shown in table 4.4) for POS is 2.717 saying that POS is 2.717 times as likely to provide a correct attribution for an unknown document. POS is followed by Characters, with a significant parameter estimate of $\beta = 0.9091$ and odds ratio of 2.482. The next highest parameter estimate of $\beta = 0.8131$ belongs to 3-4 Letter Words with an odds ratio of 2.255.

An interesting result is that for all event sets that also have n-gram realizations of that event set, the higher n-gram had a lower parameter estimate. Further research should be conducted to explore the reason for this result. One reason could be that higher event n-grams can be useful in specific combinations of canonicizers, event cullers, and analysis drivers, but not in general.

5.3 Event Cullers

Most Common Events was significant and contributed to the correct authorship of a document since the parameter estimate was positive having an **OR** = 1.534 (shown in Table 4.6) saying that analyzing the most common events is 1.534 times more likely to correctly attribute an author to a document than the Least Common Events. Confirming that Most Common Events contributes to the overall attribution of a document is that Most Common Events/Least Common Events was significant in the model and having **OR** = 1.49. By applying the two sequentially we took the fifty least common events of the fifty most common events from the document set, leaving the exact same event set as if we took just the fifty most common events. Similarly by applying Most Common Events, Least Common Events, and X-treme Culler we are more likely than not to be left with just the fifty most common events in the document set. Therefore, since this permutation of event cullers came back significant and with a positive parameter estimate, it further solidifies that the most common events can be beneficial to the attribution of a document. Further, Most Common Events/X-treme Culler and X-treme Culler/Most Common Events were both significant and both had positive parameter estimates with **OR** > 1. This is shown in table 4.6.

Another interesting result is the significance of X-treme Culler/Least Common Events with a positive parameter estimate with **OR** = 1.138 and the lack of any other event culler permutation using Least Common Events as the first event culler failing to reject the null hypothesis. This suggests that by examining the least likely events that overlap the two authors event space can be beneficial in the attribution of a document. It also suggests that the order matters in which you apply Least Common Events and the other event cullers.

5.4 Analysis Drivers

Among the analysis drivers that were significant and had positive parameter estimates, Kendall Correlation Distance had the highest parameter estimate of $\beta = 0.5785$ and **OR** = 1.783 (see table 4.8). Cosine Distance had the next highest parameter estimate of $\beta = 0.4889$ and **OR** = 1.631 while also being significant at the $\alpha = 0.05$ level. Cosine distance assigns a value between 0 and 1, by taking the cosine of the angle between the event vectors in space. The rest of the predictors with positive estimates of β as well as being significant at the $\alpha = 0.05$ level are JW Cross Entropy and Intersection Distance with odds ratios of 1.165 and 1.203 respectively. Both JW Cross Entropy and Intersection Distance do not use traditional definitions of distance to assign an author to a document. Intersection Distance computes a ratio by taking the number of events in the intersection of the two event spaces, and dividing by the number of events in the union of the event space. These analysis drivers are among the top performers of analysis drivers studied.

The top methods all use a complex method to determine the "distance" between two documents. Suggesting that a more sophisticated approach of analyzing docu-

ments using complex correlations can better predict the true author of an unknown document. Therefore, further exploration of using complex means of computing the "distance" between a document is recommended.

The significant predictor at $\alpha = 0.05$ with the most negative parameter estimate of $\beta = -1.339$ was Markov Chain Analysis with $\mathbf{OR} = 0.262$ (see Table 4.8). The next most negative predictor that was significant was Naive Bayes Classifier with a parameter of $\beta = -0.8962$ and $\mathbf{OR} = 0.408$. Following these two are RN Cross Entropy, Levenshtein Distance, LZW Divergence, LDA, and Kullback Leibler Distance, all with parameter estimates $-0.1 < \beta < -0.5$ and significant at the $\alpha = 0.05$ level. Hence, the odds of correctly attributing an author to a document for each of these analysis drivers is the same as using Canberra Distance.

Chapter 6

Future Work

Future work would include a more comprehensive study, by sampling in such a way to isolate the top performing level in each variable, and testing to determine the single best level in each different variable. This will allow for us to recommend a single combination of an event set, event culler, canonicizer, and analysis driver for the attribution of English essays.

Similar to sampling to help refine the definition of the best attribution method, further sampling with different test corpora to see if the above suggested practices in authorship attribution translate to different languages, genres, or lengths of documents. Sampling in such a way can help define recommendations for the different types of authorship attribution problems. With a clear definition for multiple authorship attribution problem, a researcher can be prepared to handle problems on an ad-hoc basis.

Chapter 7

Conclusion

In light of the results described above, my recommendations for best practices are as follows. Studying the parts of speech, characters, and words beginning with a vowel are important in the describing an authorial fingerprint. Limiting the event set to the fifty most common events is also advised. Unifying the case of a document will help remove inconsistencies that can result in decreased accuracy of the attribution of a document. When analyzing a document, using complex pattern based "distance" methods such as, Kendall Correlation, Cosine Distance, and Juola-Wyner Cross entropy are useful methods.

This study gave an idea of the order of importance when studying the different means of conducting authorship attribution tasks. By identifying the rankings of the different levels inside these variable and controlling for the other confounders we can see how the levels perform while interacting with each other. However, just by combining the top performing event set, event culler, canonicizer, and analysis driver may not be the single best performing authorship attribution method for all corpora but it will be the method that is most likely to correctly attribute an English essay. Having a direction in which to move in authorship attribution studies will allow one to further refine the field and move towards the ultimate goal of being able to confidently

predict the author of an unknown text.

Bibliography

[1] M. Brennan and R. Greenstadt. Practical attacks against authorship recognition techniques. In *Innovative Applications of Artificial Intelligence (IAAI)*, Pasadena, Ca.

[2] D. I. Holmes. Authorship attribution. *Computers and the Humanities*, 28:87–106, 1994.

[3] Noecker J. Ryan M. Juola, P. and S. Speer. Jgaap 4.0- a revised authorship attribution tool. In *Digital Humanities 2009*, College Park, MD, 2009.

[4] P. Juola. Authorship atrribution for electronic documents. pages 119–130, 2006.

[5] P. Juola. Authorship attribution. *Foundations and trends in information Retrieval*, 1, 2006.

[6] Kupper L.L. Muller K.E. Kleinbaum, D.G.

[7] J. Rudman. The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities*, 31.

Appendix A

Distance Formulas

- Canberra Distance: $D(A,B) = \sum_{i=1}^{n} \frac{|a_i - b_i|}{|a_i| + |b_i|}$ Where $A = (a_1, a_2, \ldots, a_i)$ and $B = (b_1, b_2, \ldots, b_i)$
- Cosine Distance: $D(A,B) = \frac{A \cdot B}{\|A\| \|B\|}$
- Intersection Distance: $D(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|}$
- Kendall Correlation Distance:

$$D = 1 - \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

- Levenshtein Distance: The minimum number of edits to transform one string into another string.