A Stylometric Analysis of the Book of Mormon Through PCA and Random Forests

Mitchell Pudil

December 14, 2019

1 Introduction

The Book of Mormon is a religious text used by members of The Church of Jesus Christ of Latter-day Saints in tandem with the Bible. This book is believed by members of that church to be a 19th century translation from Hebrew and Ancient Egyptian of writings by ancient American inhabitants[1]. Critics of the Book of Mormon, however, claim that the supposed translator, Joseph Smith, was primarily responsible for writing (rather than translating) the book. To discredit Joseph Smith's description of the book's origin, skeptics started proposing theories about how it was written even before it was published in 1830 (Midgley, 1997).

This analysis compares the sentence structures and phraseology of major characters throughout the book with other characters in the book, and then compares these speakers with the writings of Smith. Specifically, the questions that will be answered here are:

- How distinguishable are the individual characters in the Book of Mormon?
- How distinguishable is the translator, Joseph Smith, from the voices in the Book of Mormon?

Previous research has varied in both findings and quality of analysis. Since 1980, five major stylometric analyses of the Book of Mormon have been published: three by researchers at Brigham Young University[2], another by a doctoral student at Bristol Polytechnic[3], and another by researchers at Stanford University. The most recent studies have used techniques such as stylometry, nearest shrunken centroid, and linear discriminate analysis and have come to the conclusion that the authors in the book are distinguishable from each other and from Joseph Smith[4].

The purpose of this paper is to analyze the claims of authorship from a slightly different perspective. Specifically, I focus not only on the specific writings, but also take into account the speaker, if any, that the writer was referencing.[5] Further, my methodology differs in that I use principal component analysis (PCA) and random forests that make use of part of speech bi-grams and common translated phrases in Hebrew to understand the nature of authorship. These methods are powerful enough to give us a good chance at separating the authors, but also allow us to interpret the results.

[1] Title Page, Book of Mormon

[2] See Larsen (1980), Hilton (1990), Roper (2012)

[3] See Holmes (1992)

[4] And distinguishable from other individuals Smith knew during that time

[5] For example, if an author quoted someone else directly, then the person quoted would be used.

The next two sections will explain more specifically the data and methodology used to perform the analysis. Following that, I will explain the results, and the final section will detail the strengths, weaknesses, and implications of the study alongside opportunities for further research.

2 Data

The data used in these analyses are primarily random samples from the Book of Mormon. In order to create these samples, the Book of Mormon was separated by speaker and by author. Noting that some characters speak much more in the Book of Mormon than others (oftentimes differing by a factor of more than 100), only a few speakers with the most words were chosen to be analyzed. The text was then broken down to create several subcorpora [6].

In order to produce a subcorpora of the writings of Joseph Smith, I sampled from various selections of the Doctrine and Covenants[7] his journal entries[8], and several of his revelations, letters, and speeches from History of the Church. All of these writings of Joseph Smith were crucial to use in the analyses because it gives us a broad scope of his writing style over various subject matters. That is, since the Book of Mormon is comprised of narratives, sermons, and other types of writings, we need more than simply one type of writing from Smith to effectively compare his style with the authors in the Book of Mormon. Table 1 below specifically shows the decomposition of the text used by author for this study.

Speaker	Alma	Helaman	Mormon	Moroni	Nephi	Joseph Smith
Tokens	17,517	6,030	12,522	9,614	17,040	20,293
Subcorpora	10	4	9	3	7	10

Table 1: Corpus Composition Table

These tokens were then further broken down to analyze parts of speech of each word using common n-grams of tokens and of parts of speech used throughout the Book of Mormon and Joseph Smith's writings. Table 2 below describes the variables used in the analysis, alongside their respective descriptions. Figure 1 then shows the distributions of each of the features from Table 2. It is apparent from Figure 1 that there are many outliers in the frequencies of various n-grams. This is likely due to two reasons: the length of the subcorpora, and the differences in writing styles across authors. Specifically,the probability of identifying specific bigrams or trigrams is greatly reduced as the length of the text decreases. Also, if it is to be the case that the writing styles of the authors vary, then we may expect to see a wide array of bigrams or trigrams for each author.

[6] Each file consisted of around 1,500-2,000 words

[7] A text that largely had been written by Joseph Smith, without debate

[8] See www.josephsmithpapers.org

Feature	Description
Author	Person to which the writing and voice is contributed
JS	"Joseph" if the author is Joseph Smith, "BOM" otherwise
Sentence Length	Average sentence length of the document
Verbs	Proportion of verbs used in the excerpt as a fraction of all words, excluding spaces and punctuation
Adjectives	Proportion of adjectives used in the excerpt as a fraction of all words, excluding spaces and punctuation
Adverbs	Proportion of adverbs used in the excerpt as a fraction of all words, excluding spaces and punctuation
Nouns	Proportion of nouns used in the excerpt as a fraction of all words, excluding spaces and punctuation
Conjunction	Proportion of conjunctions used in the excerpt as a fraction of all words, excluding spaces and punctuation
Determinants	Proportion of determinants used in the excerpt as a fraction of all words, excluding spaces and punctuation
Det + Noun	Proportion of bigrams that are a determinant followed by a noun This tended to be a popular part of speech bigram among all authors, including Joseph Smith
Adp + Det	Proportion of bigrams that are an adposition followed by a determinant
Pron + Verb	Proportion of bigrams that use a pronoun followed by a verb
Did + verb	Proportion of bigrams that use the word "did" followed by a verb. This is commonly found in narrative-style writings
Cause	Proportion of tokens that include the word "cause", i.e. "cause", "causes", "caused". This tended to be more common for sermons
Things + which	Proportion of bigrams that are the word "things" followed by either "that" or "which". This tended to be more common for narrative-style writing
Passed + away	Proportion of bigrams that are the phrase "passed away."
Holy + Ghost	Proportion of bigrams that are the phrase "Holy Ghost."
More + Part	Proportion of bigrams that are the phrase "more part" More common with sermons than narratives
Say + Unto + You	Proportion of trigrams that are the phrase "more part" More common with sermons than narratives
Save + it + be	Proportion of trigrams that are the phrase "save it be" or "save it were"
By + the + power	Proportion of trigrams that are the phrase "by the power."
These + are + they	Proportion of trigrams that are the phrase "these are they"
Came + to + pass	Proportion of trigrams that are the phrase "came to pass." The Hebrew word for this is "wayehi," which is oftentimes used to connect two ideas together.

Table 2: Variables Used Throughout Analysis

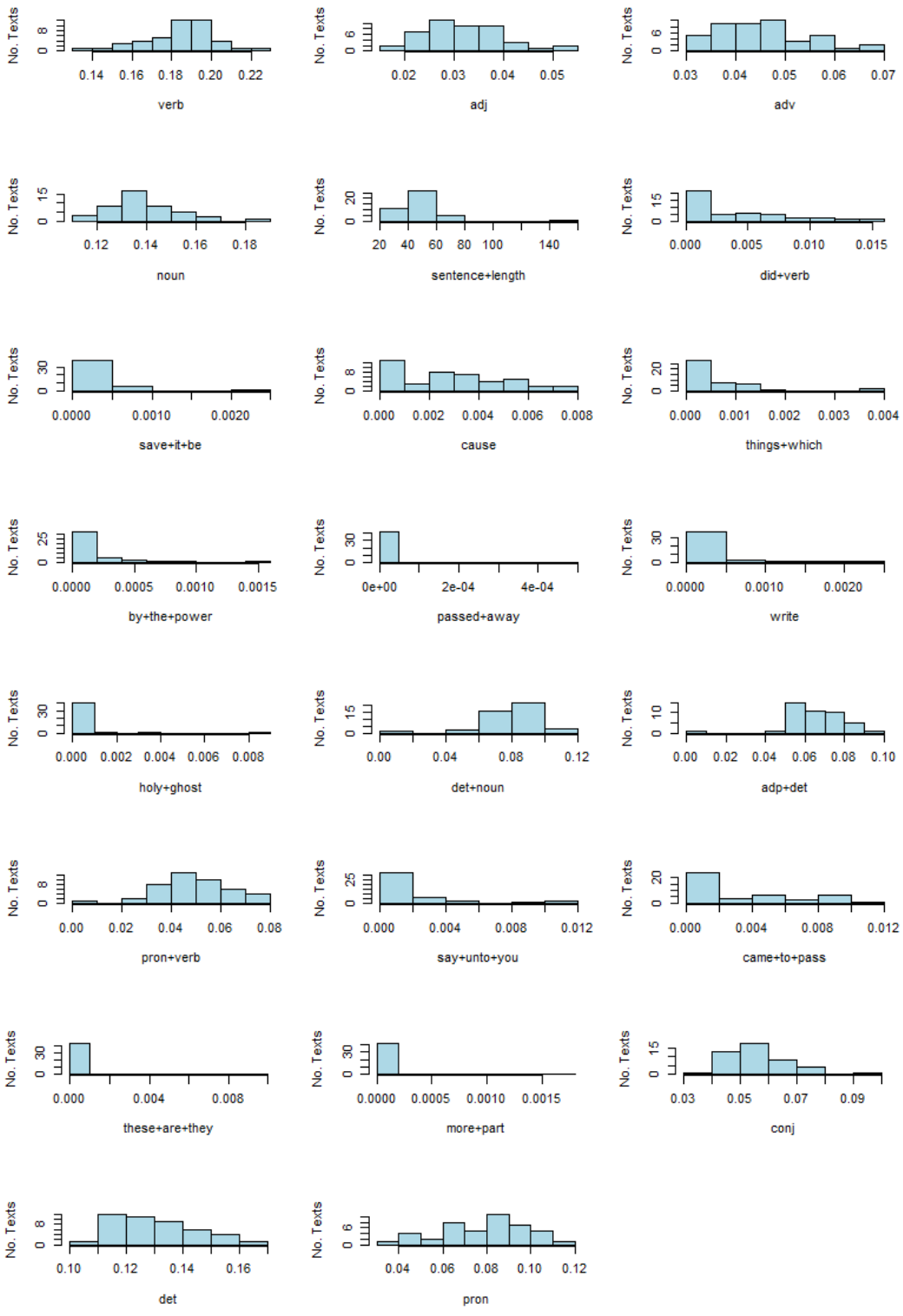

Figure 1: Frequencies of features in Book of Mormon and Joseph Smith's writings

Next, we examine the dispersions of various words. This allows us to understand how common the word is across our corpus, giving us more insight into how particular words were used to classify the author. Figure 2 below shows how the tokens found in the corpus vary in dispersion.

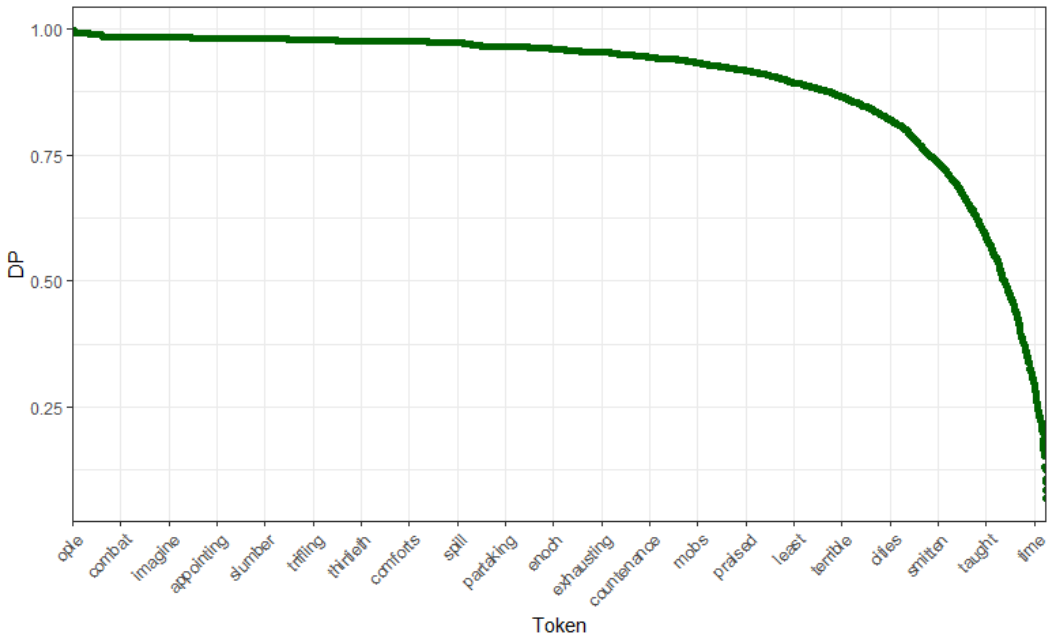

Figure 2: Dispersion of tokens across Book of Mormon & Joseph Smith Corpus

As shown in Figure 2, many of the tokens found across the corpus have high levels of dispersion, suggesting that they are only found in a few of the texts used. However, other words such as "taught" or "time" have a smaller DP score, suggesting that they are found in many of the texts.

We now transition to explaining the methods that were used to analyze the questions from the introduction.

3 Methods

The first question of interest is: how distinguishable are the writings of the authors in the Book of Mormon? The explanation behind this question is that it would be much easier for someone to write in one different style than to write in three or four (or more) unique styles. Thus, if we can effectively differentiate the writings of the individuals in the Book of Mormon, it gives weight to the notion that Joseph Smith was a translator, rather than the author, of the Book of Mormon[9]. The analysis for this question will use a random forest to determine the distinguishability . The model will be trained using the variables discussed in the first question and subcorpora from each author. Following this, other subcorpora not used in the prediction phase (i.e. the test set) will be ran through the model and will predict which author best matches the style of writing. Again, I will only be using the work of the most prolific authors and therefore will be unable to say anything about the distinguishability of any other author in the Book of Mormon[10].

[9] Or at the very least, that he was not the author of the Book of Mormon

[10] Credit is given to dozens of authors or speakers in the Book of Mormon

The second question asks how the writings of Joseph Smith differ from the authors written in the Book of Mormon. In order to explicitly test this, we use another random forest in which we collapse all Book of Mormon authors into a single "Book of Mormon" author category and use the features from Table 2 to test how accurate of a prediction we receive.

Both the second and the third questions will be partially answered through the construction of a cluster plot created through PCA. Specifically, it will compare the Euclidean distances of the parts of speech used in each subcorpora of the Book of Mormon and Joseph Smith's writings.

4 Results

This section is dedicated to carrying out the analyses sufficient to answering the main questions of interest. While many of the figures initially shown in section 4.1 include information regarding the writing style of Joseph Smith, discussion of his writing style will be saved until section 4.2.

4.1 Comparisons of Authors Within the Book of Mormon

We begin the analysis by exploring how similar the author are within the Book of Mormon. To begin, we compare the proportion of each part of speech across the major authors. This is shown in Figure 3.

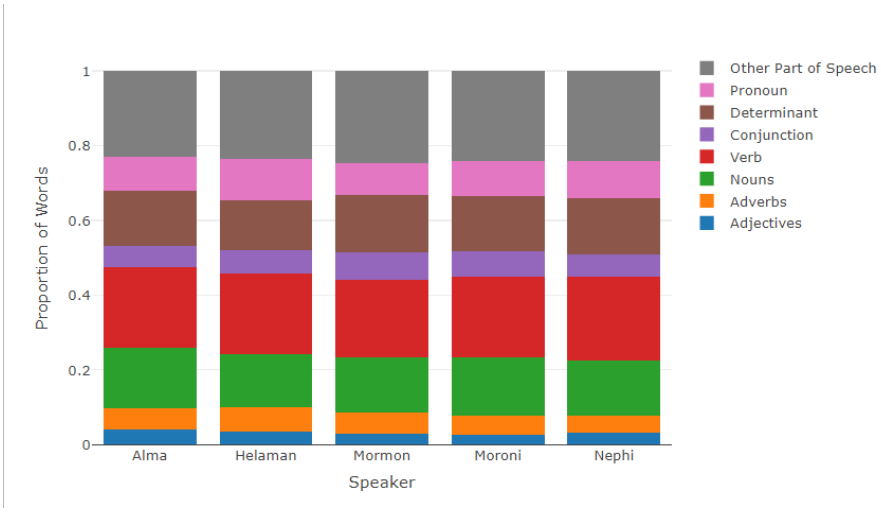

Figure 3: Part of speech by speaker in the Book of Mormon

From this plot, we notice that there is much similarity among the authors in the Book of Mormon in terms of their relative usage of various parts of speech. Simultaneously, however, there is also variability between speakers in the Book of Mormon in terms of the exact proportion of each part of speech used. Overall, we notice that nouns and verbs are used most frequently for each of the authors, whereas adjectives and adverbs are used much less frequently.

Next, we consider the most common multi-word expressions used among authors of the Book of Mormon. Table 3 below shows the most common phrase used by each of the authors[11].

[11] Along with Joseph Smith's most common phrase, as will be discussed in subsection 4.2

Author	4-gram Token
Alma	"I say unto you", 0.40%
Helaman	"It came to pass", 0.71%
Mormon	"It came to pass", 0.79%
Moroni	"The brother of Jared", 0.32%
Nephi	"It came to pass", 0.47%
Joseph Smith	"These are they who", 0.14%

Table 3: Most popular phrase by Book of Mormon author, and Joseph Smith

It is interesting to note that the 4-gram "It came to pass" is the most common 4-gram among several of the Book of Mormon authors. This should not come as a surprise, however, since the Hebrew word for this is "wayehi," which is oftentimes used to connect two ideas together. Thus, while an initial look at these popular n-grams may initially be surprising, it is actually more in line with the theory that the books were not authored by Joseph Smith, even though at this stage, it is difficult to determine exactly how distinguishable the writing styles of the authors of the Book of Mormon are. We can tell, however, that the usage of this phrase is similar between Helaman and Mormon, but different than Nephi, and even more different from other authors[12]

Up until now, the descriptive analyses in this paper have largely been a factor of individual words or phrases: we have not yet examined how the combinations of such features can help classify the works of the authors. Interestingly, this is where much of the earlier work (1980-1990) ended and would have given the authors reason to claim that the works are quite similar based off of the frequency of parts of speech and most common phrases or words. However, using more advanced algorithms and models such as PCA and Random Forest, we show that accurate classification is quite simple, even when simply using parts of speech.

To exemplify this notion, I used Principal Component Analysis of the Euclidean distances among the parts of speech for each subcorpora. Figure 4 below shows the classification. It is obvious to notice from Figure 4 how separate each of the authors are from each other. With regards to the Book of Mormon authors alone, it is apparent that each author has a unique writing style that, in general, is straightforwardly separable from any other. One potential exception to this is the work of Helaman, which happens to be the subcorpus for which the least amount of text was extracted.

[12]Since their most popular 4-gram is not "It came to pass"

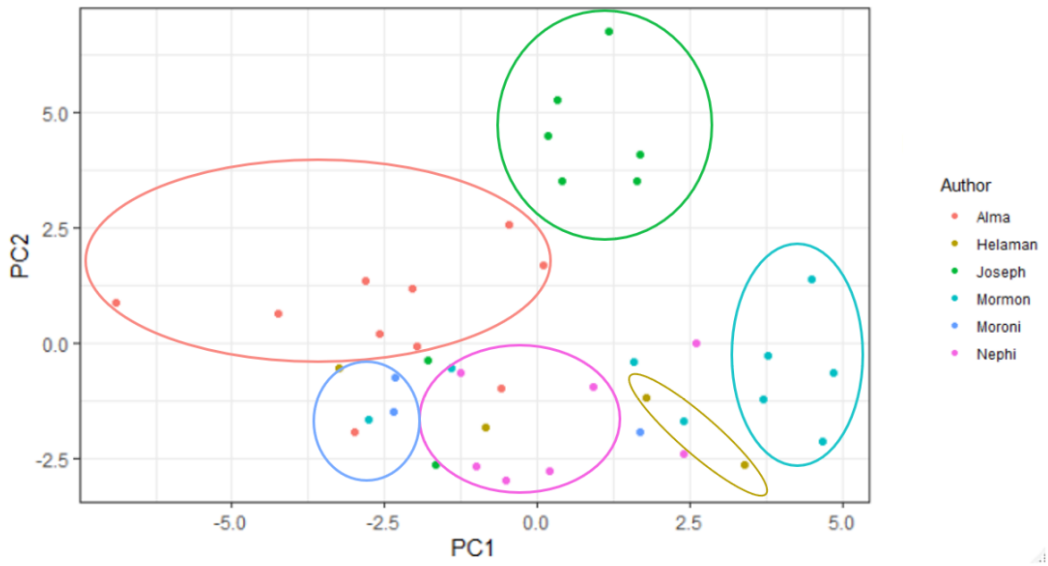

Figure 4: Clusters by part of speech usage among Book of Mormon authors and Joseph Smith via PCA of Euclidean distances

To further the analysis, I constructed a random forest that uses all of the features shown in Table 2, including n-grams of parts of speech, n-grams of common phrases used throughout the Book of Mormon and the Hebrew language, when translated to English. Due to the small number of subcorpora, I used two versions of the random forest. The first used approximately 67% of all texts, randomly chosen from within the Book of Mormon, as training data, and left the other 23% as test data. This provided us with figure 5, which shows the most important variables used in determining which author within the Book of Mormon.

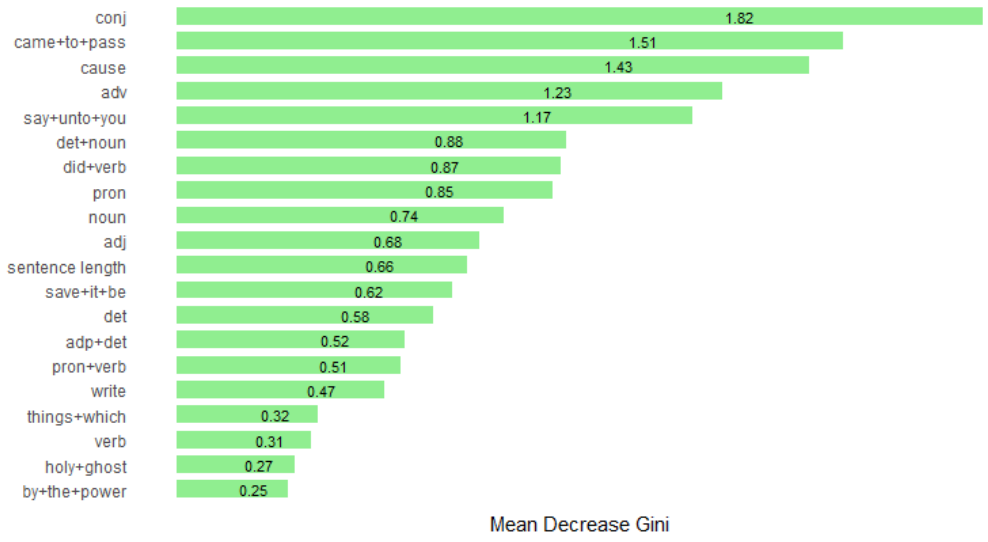

Figure 5: Top 20 most important features for classification of Book of Mormon authors

It makes sense here that certain phrases such as "came to pass" and "say unto you" would be highly influential in classifying the author. First, they are common phrases when translated from Hebrew to English. Additionally, some authors in the Book of Mormon use more narratives (such as Nephi) whereas others use sermons instead. Since "came to pass" is used much more frequently with narratives, and "say unto you" is much more common with sermons, these phrases are highly valued in classification.

The second method of random forest I used was derived from the leave-one-out cross validation method. That is, for each text, I constructed a random forest where the training dataset was all subcorpora that was not that specific text and then used the random forest to predict the most likely author for the particular passage. Seeing that there was a limited number of each subcorpora for each author, it would initially be expected that without the use of unique proper pronouns or other "give-aways"[13], it would be difficult to build a classifier that accurately predicted which author wrote which passage. Surprisingly, however, the random forest predicted quite well for the amount of data it was given. Figure 6 shows the confusion matrix, which compares the prediction of the author of each text when left out of the model against the actual author.

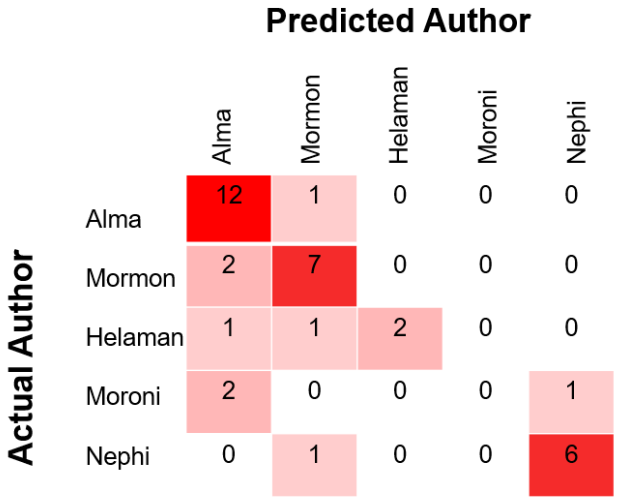

Figure 6: Top 20 most important features for classifying authors within the Book of Mormon

As evident from Figures 4, 5, and 6, although our corpus only contains 36 samples of text from the Book of Mormon and there are 5 possible authors, the models (PCA and Random Forest) do quite well at classification. Specifically, the random forest has a misclassification rate of only 25%.

4.2 Joseph Smith v. Book of Mormon Authors

The second question we are interested in answering is whether there is a distinct difference in the writing style of Joseph Smith and the writers in the Book of Mormon. In order to answer that question, we first analyze Table 3 and Figure 4 from the perspective of Joseph Smith versus the authors in the Book of Mormon. From Table 3, we notice that Joseph Smith has a different 4-gram token than the others. Upon

[13]The term "give-away" here is used to express specific words or phrases that usually appear in a particular text from an author but never appear in text from other authors.

closer observation, we find that Joseph Smith's most common phrase and Moroni's are the only ones whose phrase does not have a one-word Hebraic translation. Also, Joseph's most common 4-gram occurs much less frequently than any other author's 4-gram (about 3.8 times less frequently). From Figure 4, we notice that Joseph's writings are easily separable from the other authors in the Book of Mormon. The vast majority of Joseph's writings are clustered alone, with no other author inside the cluster.

Along with these figures, I also constructed random forests in a similar fashion as I did to separate authors within the Book of Mormon. The training data for all random forests used to distinguish between Joseph and the Book of Mormon had as its predictor whether the text was written by Joseph or if it was claimed to be written by someone in the Book of Mormon. For the first random forest, I sampled 6 of the 10 texts from Joseph Smith and 28 out of 33 remaining texts from the Book of Mormon to use as training data. Figure 7 below shows the most important features used to distinguish between the works of Joseph Smith and that of the Book of Mormon.

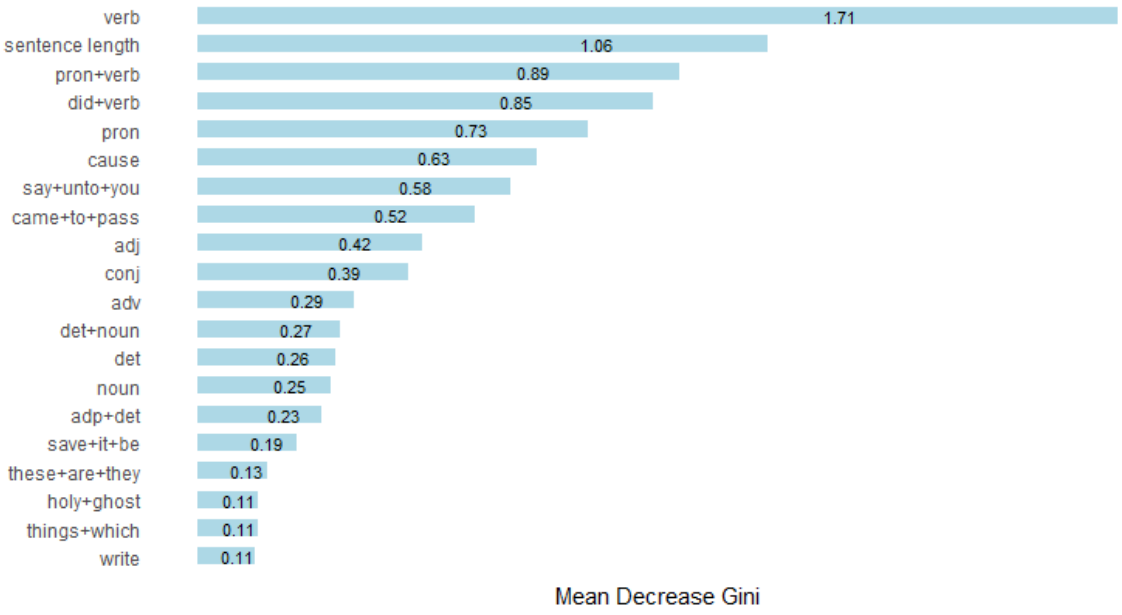

Figure 7: Top 20 most important features for classifying Smith's writings from Book of Mormon authors

From comparing Figures 5 and 7, we can see that the most important variables in determining whether a work was written by Joseph Smith or by an author in the Book of Mormon are a bit different than the most important variables in determining which author in the Book of Mormon wrote a text. Specifically, the within-Book comparison weighted token n-grams more highly, but for Joseph v. Book of Mormon, part of speech and sentence length were more important on average.

The second random forest created to test the difference between the writings of Joseph and the Book of Mormon authors was the leave-one-out random forest/cross-validation method. Figure 8 below shows the confusion matrix for that random forest.

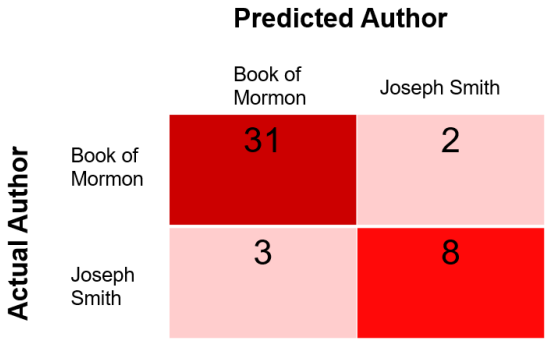

Figure 8: Random forest confusion matrix: Joseph Smith v. Book of Mormon authors

As shown in Figure 8, the random forest was able to correctly predict with 88.6% accuracy whether the text was written by Joseph Smith or if it was written by an author in the Book of Mormon. Overall, we conclude that it is quite simple to determine whether an excerpt was written by Joseph Smith or if it was written by someone in the Book of Mormon.

5 Discussion

The figures and analyses shown in the results section suggest that there are clear differences among the writers in the Book of Mormon, and perhaps even more distinguishable is the writings of Joseph Smith against the writings of authors in the Book of Mormon. In either case, sentence structure alone[14] allows us to clearly distinguish between the writings of the respective authors. However, when using random forests that take into consideration additional features such as n-grams of tokens or parts of speech, the accuracy increases to around 75% for determining which author within the Book of Mormon wrote a passage. Similarly, the accuracy of the random forest with the same data is able to accurately predict whether Joseph Smith or an author in the Book of Mormon wrote a passage around 88% of the time. Overall, these analyses provide additional evidence to the claim that the Book of Mormon is ancient scripture that was translated by Joseph Smith.

This being said, there are still some caveats to the analyses. First, if this text was indeed translated, rather than authored, by Joseph Smith, then it is likely that much of the sentence structure would have been changed in order to be readable in English. For example, periods are English constructions that do not exist in Hebrew. Additionally, some single words in Hebrew may take three or four (or more) words to express in English. While we are still able to accurately make predictions, this could have major implications for the initial exploratory data analysis, especially Figures 2 and 3.

We also cannot discount the possibility based solely on this analysis that there were not other writers who may have helped Joseph write the Book of Mormon. However, the fact that forms of poetry and phraseology such as chiasmus[15] that were popular among ancient inhabitants who spoke and wrote Hebrew was not yet discovered until over a hundred years after the death of Joseph Smith and after the Book of Mormon was released, placing serious doubt on this alternative theory (Welch 1969).

[14] Even without using specific words

[15] A style of writing in which words, grammatical constructions, or concepts are repeated in reverse order, in the same or a modified form; e.g. 'Poetry is the record of the best and happiest moments of the happiest and best minds.'.

Finally, it is unfortunately difficult to use all of the Book of Mormon effectively. There are dozens of speakers in the Book of Mormon who wrote a page or less, so any attempt to analyze their style of writing would be largely uninformative. Additionally, it is difficult to process the entire Book of Mormon and all the works of Joseph Smith computationally, and thus we do not have the entirety of the works of any particular individual.

While the original records of the Book of Mormon are no longer, there are several opportunities for further research in this area. For example, it may be worthwhile to attempt to re-translate the Book of Mormon into Hebrew or Egyptian and determine the distinguishability of the Book of Mormon authors. It could also be interesting to compare the writings of the friends of Joseph Smith to the authors in the Book of Mormon, although again, the notions of certain forms of Hebraic writings undiscovered during Joseph's time places doubt on that theory.

References

Church of Jesus Christ of Latter-Day Saints (1999). The Book of Mormon: Another Testament of Jesus Christ; The Doctrine and Covenants of the Church of Jesus Christ of Latter-Day Saints.

David I. Holmes (1992). "A Stylometric Analysis of Mormon Scripture and Related Texts," Journal of the Royal Statistical Society A 155, part 1: 91–120.

Hilton, John (1990). "On Verifying Wordprint Studies: Book of Mormon Authorship," BYU Studies 30/3: 89–108; reprinted in Reynolds, Book of Mormon Authorship Revisited, 225–53.

Jessee, Dean C., Ronald K. Esplin, and Richard Lyman Bushman, eds. The Joseph Smith Papers. Salt Lake City: Church Historian's Press, 2008-.

Jockers, Matthew, et.al (2008). "Reassessing Authorship of the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification," Literary and Linguistic Computing 23/4: 465–91.

Larsen, Wayne, et. al (1980). "Who Wrote the Book of Mormon? An Analysis of Wordprints," BYU Studies 20/3: 225–51; reprinted by Wayne A. Larsen and Alvin C. Rencher in Book of Mormon Authorship: New Light on Ancient Origins, ed.

Midgley, Louis (1997). "Who Really Wrote the Book of Mormon? The Critics and Their Theories," 101–39.

R Core Team (2017), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roper, Matthew, et. al (2012). "Stylometric Analyses of the Book of Mormon: A Short History," Journal of Book of Mormon Studies Vol 21.1 Article 4 : 28-45.

Smith, Joseph (1902–32). History of the Church, 7 volumes; Deseret Book Company; ISBN 0-87579-486-6 (1902–1932; Paperback, 1991).

Welch, John W (1969). "Chiasmus in the Book of Mormon." Brigham Young University Studies, vol. 10, no. 1, pp. 69–84. JSTOR, www.jstor.org/stable/43041878.

Code Appendix

```
# Book of Mormon Voices: EDA


library(quanteda)
library(tidyverse)
library(spacyr)
library(cluster)
library(factoextra)
library(readtext)
library(plotly)
library(tau)
library(data.table)
library(randomForest)

setwd("C:/Users/Mitchell Pudil/Documents/textstat_tools/BookMormon/Voices/Subset/")
authors <- c("alma", "helaman", "joseph", "mormon", "moroni", "nephi")



# Question 1  ------------------------------------------------------------


pos_by_author <- data.frame(matrix(nrow=length(authors), ncol=10, data=NA))
colnames(pos_by_author) <- c("author", "total_words", "verb",
"adj", "adv", "noun", "sentence_length")
pos_by_author$author <- authors
# We'll get the paths to the book of mormon subsets
bom_paths <- list.files(full.names = TRUE, pattern = "*.txt")

for(i in 1:length(authors)){

  # And we'll subset out the author (voice)
  paths_sub <- bom_paths %>% str_detect(authors[i]) %>%
  keep(bom_paths, .)

  # Finally, we'll create our data.frame of texts and doc_ids
  sub_df <- readtext(paths_sub)

  # We going to use regex to remove all periods including those between words
  # which will help with the accuracy of our parsing.
```

```
  sub_df$text <- as.character(sub_df$text) %>%
  gsub("\\.", " ", .) %>% gsub("\\;", " ", .) %>%
    gsub("\\,", " ", .)

  # And create a corpus object.
  sub_corpus <- corpus(sub_df)

  # Next, wee're going to use spacy to parse the corpus.
  sub_prsd <- spacy_parse(sub_corpus, pos = T, tag = T,
  dependency = T, entity = F)

  # Remove any spaces, and then combine our pos and tag columns.
  sub_prsd <- sub_prsd  %>% filter(pos != "SPACE") %>%
    unite("pos", pos:tag)

  # Calculate percentage of each major part of speech by author
  sub_prsd_no_punct <- filter(sub_prsd, dep_rel!="punct")
  pos_by_author$total_words[i] <- nrow(sub_prsd_no_punct)
  pos_by_author$verb[i] <- nrow(filter(sub_prsd, grepl(pattern = "VERB",
  x=sub_prsd$pos)))/nrow(sub_prsd_no_punct)
  pos_by_author$adj[i] <- nrow(filter(sub_prsd, grepl(pattern = "ADJ",
  x=sub_prsd$pos)))/nrow(sub_prsd_no_punct)
  pos_by_author$adv[i] <- nrow(filter(sub_prsd, grepl(pattern = "ADV",
  x=sub_prsd$pos)))/nrow(sub_prsd_no_punct)
  pos_by_author$noun[i] <- nrow(filter(sub_prsd, grepl(pattern = "NOUN",
  x=sub_prsd$pos)))/nrow(sub_prsd_no_punct)
  pos_by_author$conj[i] <- nrow(filter(sub_prsd, grepl(pattern = "CONJ",
  x=sub_prsd$pos)))/nrow(sub_prsd_no_punct)
  pos_by_author$det[i] <- nrow(filter(sub_prsd, grepl(pattern = "DET",
  x=sub_prsd$pos)))/nrow(sub_prsd_no_punct)
  pos_by_author$pron[i] <- nrow(filter(sub_prsd, grepl(pattern = "PRON",
  x=sub_prsd$pos)))/nrow(sub_prsd_no_punct)
  pos_by_author$sentence_length[i] <- nrow(sub_prsd) / sum(sub_prsd$token==".")
}



pos_by_author$author <- c("Alma", "Helaman", "Joseph Smith", "Mormon", "Moroni",
"Nephi")
pos_by_author$otherpos <- 1 - pos_by_author$verb -
pos_by_author$adj - pos_by_author$adv - pos_by_author$noun -
  pos_by_author$conj - pos_by_author$det - pos_by_author$pron

# Plot stacked bar chart of parts of speech
plot_ly(subset(pos_by_author, author!="Joseph Smith"),
```

```
x = ~author, y = ~adj,
type = 'bar', name = 'Adjectives') %>%
  add_trace(y = ~adv, name = 'Adverbs') %>%
  add_trace(y = ~noun, name = 'Nouns') %>%
  add_trace(y = ~verb, name = 'Verb') %>%
  add_trace(y = ~conj, name = 'Conjunction') %>%
  add_trace(y = ~det, name = 'Determinant') %>%
  add_trace(y = ~pron, name = 'Pronoun') %>%
  add_trace(y = ~otherpos, name = 'Other Part of Speech') %>%
  layout(yaxis = list(title = 'Proportion of Words'),
  xaxis = list(title='Speaker'), barmode = 'stack')



plot_ly(pos_by_author, x = ~author, y = ~sentence_length, type = 'bar',
name = 'Average Sentence Length') %>%
  layout(yaxis = list(title = 'Average Sentence Length'))




# N-gram Frequencies


# given a string vector and size of ngrams this function returns word ngrams
# with corresponding frequencies
createNgram <-function(stringVector, ngramSize){

  ngram <- data.table()

  ng <- textcnt(stringVector, method = "string", n=ngramSize, tolower = FALSE)

  if(ngramSize==1){
    ngram <- data.table(w1 = names(ng), freq = unclass(ng), length=nchar(names(ng)))
  }
  else {
    ngram <- data.table(w1w2 = names(ng), freq = unclass(ng), length=nchar(names(ng)))
  }
  return(ngram)
}

# Most popular ngrams by book

for(i in c("nephi.txt", "alma.txt", "helaman.txt", "mormon.txt", "moroni.txt",
```

```
"joseph.txt")){
  ngrams <- createNgram(readtext(i), 4)
  ordered_ngrams <- ngrams[order(ngrams$freq, decreasing = TRUE),]
  total <- sum(ordered_ngrams$freq)
  print(paste0(i, ", ", ordered_ngrams$w1w2[1], ", ",
  (ordered_ngrams$freq[1] / total)*100))
}


# Most popular POS ngrams by book

sub_df <- readtext("mormon.txt")
sub_df$text <- gsub("\\.", " ", sub_df$text)
sub_corpus <- corpus(sub_df)
sub_prsd <- spacy_parse(sub_corpus, pos = T, tag = T, dependency = T,
entity = F)$pos

ngrams <- createNgram(sub_prsd, 2)
ordered_ngrams <- ngrams[order(ngrams$freq, decreasing = TRUE),]
ordered_ngrams$prop <- ordered_ngrams$freq/sum(ordered_ngrams$freq)
ordered_ngrams[1:3]







# Cluster Plot 2 ------------------------------------------------------------

source("functions/helper_functions.R")
setwd("C:/Users/Mitchell Pudil/Documents/textstat_tools/BookMormon/Voices/CrossVal/")

# We'll get the paths to the BOM corpus
bom_paths <- list.files(full.names = TRUE, pattern = "*.txt")

# Create our data.frame of texts and doc_ids
sub_df <- readtext(bom_paths)

# We going to use regex to remove all punctuation
# which will help with the accuracy of our parsing.
sub_df$text <- gsub('[[:punct:] ]+',' ',sub_df$text, perl = T)

# And create a corpus object.
sub_corpus <- corpus(sub_df)
```

```
# Next, wee're going to use spacy to parse the corpus.
# Note that we can add a dependency column to our parsing.
sub_prsd <- spacy_parse(sub_corpus, pos = T, tag = T, dependency = T,
entity = F)


# We're going to remove any spaces, and then combine our pos
#and tag columns.
sub_prsd <- sub_prsd  %>% filter(pos != "SPACE") %>%
  unite("pos", pos:tag)

# Next we create a named list from the new, concatenated column.
sub_tokens <- split(sub_prsd$pos, sub_prsd$doc_id)

# See what the result looks like.
sub_tokens

# Now, we'll use that as our tokens object.
sub_tokens <- as.tokens(sub_tokens)

# From that, we'll generate a dfm.
sub_dfm <- dfm(sub_tokens)

# We'll weight the raw counts.
sub_dfm <- dfm_weight(sub_dfm, scheme = "prop")

# And convert the result to a data.frame.
sub_dfm <- convert(sub_dfm, to = "data.frame") %>%
  rename(doc_id = document)

# Finally, we're going to convert the first row (doc_id) into row names.
# And, for fun, we'll order our columns alphabetically.
sub_dfm <- sub_dfm %>% column_to_rownames("doc_id")



df <- data.frame(scale(sub_dfm))



# Create cluster plot

m <- sapply(1:nrow(df), function(i) sapply(1:nrow(df),
                                           function(j) dist(rbind(df[i,], df[j,]))))
```

```
cmdeuc <- cmdscale(m) %>% data.frame
a <- gsub("[0-9]+.txt", "", bom_paths ) %>% basename
cmdeuc$Author <- paste0(toupper(substr(a, 1, 1)), substr(a, 2, nchar(a)),
sep="")
colnames(cmdeuc)[1:2] <- c("PC1", "PC2")
ggplot(cmdeuc, aes(PC1, PC2,color=Author)) + geom_point() +
  theme_bw() +  ggtitle("Euclidean Distance")



# Frequencies and Distributions of All Files ------------------------------------

# Create training dataset
columns <- c("author", "total_words", "verb", "adj", "adv", "noun", "sentence_length",
  "did_verb", "save_it_be", "cause", "things_which", "by_the_power", "passed_away",
  "write", "holy_ghost", "det_noun", "adp_det", "pron_verb", "say_unto_you",
  "came_to_pass", "these_are_they")
bomtrain <- data.frame(matrix(nrow=length(bom_paths), ncol=length(columns), data=NA))
colnames(bomtrain) <- columns
bomtrain$author <- cmdeuc$Author


# We'll get the paths to the book of mormon subsets
bom_paths <- list.files(full.names = TRUE, pattern = "*.txt")

# Our clustering will be based on pos counts, so we need to initialize our spcacy model.
for(i in 1:nrow(bomtrain)){

  # Finally, we'll create our data.frame of texts and doc_ids
  sub_df <- readtext(bom_paths[i])


  # And create a corpus object.
  sub_corpus <- corpus(sub_df)

  # Next, we're going to use spacy to parse the corpus.
  sub_prsd <- spacy_parse(sub_corpus, pos = T, tag = T, dependency = T, entity = F)


  for(j in 1:nrow(sub_prsd)){
    sub_prsd$nextpos[j] <- ifelse(j==nrow(sub_prsd), NA, sub_prsd$pos[j+1])
    sub_prsd$nextword[j] <- ifelse(j==nrow(sub_prsd), NA, sub_prsd$token[j+1])
    sub_prsd$nextnextword[j] <- ifelse(j >= nrow(sub_prsd)-1, NA, sub_prsd$token[j+2])
  }
```

```
# Common phrases
bomtrain$did_verb[i] <- sum(sub_prsd$token=="did" &
sub_prsd$nextpos == "VERB")/(nrow(sub_prsd) -1)
bomtrain$more_part[i] <- sum(sub_prsd$token=="more" &
sub_prsd$nextword == "part")/(nrow(sub_prsd) -1)
bomtrain$save_it_be[i] <- sum(sub_prsd$token=="save" &
sub_prsd$nextword == "it" & sub_prsd$nextnextword %in% c("were", "be"))/(nrow(sub_prsd) -2)
bomtrain$cause[i] <- sum(grepl("cause", sub_prsd$token))/nrow(sub_prsd)
bomtrain$things_which[i] <- sum(sub_prsd$token=="things" &
sub_prsd$nextword == "which")/(nrow(sub_prsd) -1)
bomtrain$by_the_power[i] <- sum(sub_prsd$token=="by" &
sub_prsd$nextword == "the" & sub_prsd$nextnextword == "power")/(nrow(sub_prsd) -2)
bomtrain$passed_away[i] <- sum(sub_prsd$token=="passed"
& sub_prsd$nextword == "away")/(nrow(sub_prsd) -1)
bomtrain$write[i] <- sum(sub_prsd$token=="write") / nrow(sub_prsd)
bomtrain$holy_ghost[i] <- sum(sub_prsd$token %in% c("Holy", "holy")
& sub_prsd$nextword %in% c("Ghost", "ghost"))/(nrow(sub_prsd) -1)
bomtrain$say_unto_you[i] <- sum(sub_prsd$token == "say" &
sub_prsd$nextword == "unto" & sub_prsd$nextnextword == "you")/(nrow(sub_prsd) -2)
bomtrain$came_to_pass[i] <- sum(sub_prsd$token == "came" &
sub_prsd$nextword == "to" & sub_prsd$nextnextword == "pass")/(nrow(sub_prsd) -2)
bomtrain$these_are_they[i] <- sum(sub_prsd$token == "these" &
sub_prsd$nextword == "are" & sub_prsd$nextnextword == "they")/(nrow(sub_prsd) -2)


# Common bigrams
bomtrain$det_noun[i] <- sum(sub_prsd$pos=="DET" &
sub_prsd$nextpos == "NOUN")/nrow(sub_prsd)
bomtrain$adp_det[i] <- sum(sub_prsd$pos=="ADP" &
sub_prsd$nextpos == "DET")/nrow(sub_prsd)
bomtrain$pron_verb[i] <- sum(sub_prsd$pos=="PRON" &
sub_prsd$nextpos == "VERB")/nrow(sub_prsd)

# Calculate percentage of each major part of speech by author
sub_prsd_no_punct <- filter(sub_prsd, dep_rel!="punct")
bomtrain$total_words[i] <- nrow(sub_prsd_no_punct)
bomtrain$verb[i] <- nrow(filter(sub_prsd, grepl(pattern = "VERB",
x=sub_prsd$pos)))/nrow(sub_prsd)
bomtrain$adj[i] <- nrow(filter(sub_prsd, grepl(pattern = "ADJ",
x=sub_prsd$pos)))/nrow(sub_prsd)
bomtrain$adv[i] <- nrow(filter(sub_prsd, grepl(pattern = "ADV",
x=sub_prsd$pos)))/nrow(sub_prsd)
bomtrain$noun[i] <- nrow(filter(sub_prsd, grepl(pattern = "NOUN",
x=sub_prsd$pos)))/nrow(sub_prsd)
```

```
  bomtrain$conj[i] <- nrow(filter(sub_prsd, grepl(pattern = "CONJ",
  x=sub_prsd$pos)))/nrow(sub_prsd)
  bomtrain$det[i] <- nrow(filter(sub_prsd, grepl(pattern = "DET",
  x=sub_prsd$pos)))/nrow(sub_prsd)
  bomtrain$pron[i] <- nrow(filter(sub_prsd, grepl(pattern = "PRON",
  x=sub_prsd$pos)))/nrow(sub_prsd)
  bomtrain$sentence_length[i] <- nrow(sub_prsd) / sum(sub_prsd$token==".")

}

bomtrain[is.na(bomtrain)] <- 0
#bomtrain <- bomtrain[,-which(is.na(colnames(bomtrain)))]

#bomtrain <- bomtrain[,-14]
# Plot distributions of each part of speech
par(mfrow=c(3,3))
for(i in 3:ncol(bomtrain)){
  hist(bomtrain[,i], xlab=gsub("_", "+", colnames(bomtrain)[i]), main="", col="light blue")
}
```

```
# Dispersions & Distributions ----------------------------------------------
```

```
# Look at dispersions and word frequencies of words
sub_df <- readtext(bom_paths)
bom_corpus <- corpus(sub_df)
summary(bom_corpus)
docvars(bom_corpus) <- bomtrain

bom_tokens <- tokens(bom_corpus, include_docvars=TRUE, remove_punct = TRUE,
                         remove_numbers = TRUE, remove_symbols = TRUE, what = "word")

bom_dfm <- dfm(bom_tokens)
topfeatures(bom_dfm)
bom_disp <- dispersions_all(bom_dfm) # All dispersions



word_freq <- textstat_frequency(bom_dfm)
```

```
ggplot(word_freq[1:100,], aes(x = rank, y = frequency)) +
  geom_point(shape = 1, alpha = .5) +
  theme_classic()


ggplot(bom_disp, aes(x=freq, y=DP)) + geom_point(col="dark green")

bom_d <- bom_disp %>% rownames_to_column("token") %>% arrange(DP)
bom_d$token <- factor(bom_d$token, levels=bom_d[order(bom_d$DP,decreasing=T),]$token)


every_nth = function(n) {
  return(function(x) {x[c(TRUE, rep(FALSE, n - 1))]})
}

#bom_d <- bom_d[-which(grepl("\\.", as.character(bom_d$token))),]

ggplot(bom_d, aes(x=as.factor(token))) +
  geom_point(aes(y=DP), col="dark green") +
  theme_bw() +
  scale_x_discrete(breaks=every_nth(250)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x="Token")




# JS and BOM word usage Plot ----------------------------------------------------
js_files <- which(grepl("joseph", bomfiles))

bom_only <- readtext(bomfiles[-js_files])
bom_only$text <- gsub('[[:punct:] ]+',' ',bom_only $text, perl = T)

js_only <- readtext(bomfiles[js_files])
js_only$text <- gsub('[[:punct:] ]+',' ',js_only $text, perl = T)

bom_only_corpus <- corpus(bom_only)
js_only_corpus <- corpus(js_only)

bom_tokens_count <- tokens(bom_only_corpus) %>% unlist %>% tolower %>%
table %>% data.frame
colnames(bom_tokens_count) <- c("Token", "Frequency_bom")
js_tokens_count <- tokens(js_only_corpus) %>% unlist %>% tolower %>%
```

```
table %>% data.frame
colnames(js_tokens_count) <- c("Token", "Frequency_js")

all_tokens_count <- merge(bom_tokens_count, js_tokens_count,
                          by="Token", all=TRUE)
all_tokens_count[is.na(all_tokens_count)] <- 0
all_tokens_count$NFbom <- (all_tokens_count$Frequency_bom /
sum(all_tokens_count$Frequency_low))*100
all_tokens_count$NFjs <- (all_tokens_count$Frequency_js /
sum(all_tokens_count$Frequency_high))*100


all_tokens_count$NFbom <- all_tokens_count$Frequency_bom/sum(all_tokens_count$Frequency_bom)
all_tokens_count$NFjs <- all_tokens_count$Frequency_js/sum(all_tokens_count$Frequency_bom)

all_tokens_count_subset <- subset(all_tokens_count, NFbom < 0.02 &
NFjs < 0.02)



ggplot(data=all_tokens_count_subset, mapping=aes(x=NFbom, y=NFjs),
label=Token) +
  geom_point() +
  theme_bw() +
  theme(axis.title=element_text(size=15), axis.text=element_text(size=12)) +
  geom_abline(slope=0.2, color="red", linetype="dashed") +
  labs(x="Book of Mormon", y="Joseph Smith",
       caption = "Figure 1: Normalized Frequency of Tokens in Book
       of Mormon vs. Joseph Smith") +
  geom_text(aes(label=ifelse(all_tokens_count_subset$NFbom > 0.013 |
  all_tokens_count_subset$NFjs > 0.0017 &
                               !all_tokens_count_subset$Token %in% c("the", "in"),
                             as.character(Token), '')) , hjust=-0.2, vjust=0)



# Random Forest JS vs BOM  ------------------------------------------------------------
set.seed(12)
bomtrain$js <- ifelse(bomtrain$author=="Joseph", "Joseph", "BOM")
bomall <- bomtrain

# Sample separately from Joseph Smith and other authors to make sure
# we have some
# of Smith's work to go off of
```

```
# Regular Random Forest
js_rows <- sample(which(bomtrain$js=="Joseph"), 6, replace=FALSE)
bom_rows <- sample(which(bomtrain$js=="BOM"), 28, replace=FALSE)

train_rows <- c(js_rows, bom_rows)
bom_train <- bomall[train_rows,]
bom_test <- bomall[-train_rows,]

bom_m1 <- randomForest(formula = as.factor(js) ~ . -author -total_words,
data = bom_train)


pred_df <- data.frame(matrix(nrow=nrow(bomall[-train_rows,]), ncol=2, data=NA))
colnames(pred_df) <- c("actual", "pred")
pred_df$actual <- bomall[-train_rows,]$js

pred_df$pred <- predict(bom_m1, bom_test) %>% as.character

imp <- bom_m1$importance %>%
  data.frame() %>%
  rownames_to_column("feature") %>%
  dplyr::arrange(desc(MeanDecreaseGini)) %>%
  dplyr::top_n(20)
imp$feature <- gsub("_", "+", imp$feature)
imp$feature[which(imp$feature=="sentence+length")] <- "sentence length"
imp %>%
  ggplot(aes(x = reorder(feature, MeanDecreaseGini), y = MeanDecreaseGini)) +
  geom_col() +
  coord_flip() +
  labs(x = "", y = "Mean Decrease Gini") +
  ggtitle("Top 20 important variables (JS vs. BOM)") +
  theme_classic()

# Leave one out

pred_df <- data.frame(matrix(nrow=nrow(bomall), ncol=2, data=NA))
colnames(pred_df) <- c("actual", "pred")
pred_df$actual <- bomall$js

for(i in 1:nrow(bomall)){
  bom_train <- bomall[-i,]
  bom_test <- bomall[i,]
  bom_model <- randomForest(formula = as.factor(js) ~ . -author -total_words, data = bom_train)
  pred_df$pred[i] <- predict(bom_model, bom_test) %>% as.character
```

```
}

pred_df


# RF Book of Mormon

bom_only_all <- bomall[-which(bomall$author=="Joseph"),]
bom_train_rows <- sample(1:nrow(bom_only_all), 22, replace=FALSE)
bom_train_nojs <- bom_only_all[bom_train_rows,]
bom_test_nojs <- bom_only_all[-bom_train_rows,]

bom_m2 <- randomForest(formula = as.factor(author) ~ . -total_words -js,
data = bom_train_nojs)


pred_bom <- predict(bom_m2, bom_test_nojs)
pred_bom %>% data.frame()


imp <- bom_m2$importance %>%
  data.frame() %>%
  rownames_to_column("feature") %>%
  dplyr::arrange(desc(MeanDecreaseGini)) %>%
  dplyr::top_n(20)

imp$feature <- gsub("_", "+", imp$feature)
imp$feature[which(imp$feature=="sentence+length")] <- "sentence length"

imp %>%
  ggplot(aes(x = reorder(feature, MeanDecreaseGini), y = MeanDecreaseGini)) +
  geom_col() +
  coord_flip() +
  labs(x = "", y = "Mean Decrease Gini") +
  ggtitle("Top 20 important variables (Within BOM)") +
  theme_classic()


# Leave One Out - BOM only

pred_df <- data.frame(matrix(nrow=nrow(bom_only_all), ncol=2, data=NA))
colnames(pred_df) <- c("actual", "pred")
pred_df$actual <- bom_only_all$author

for(i in 1:nrow(bom_only_all)){
```

```
  bom_train <- bom_only_all[-i,]
  bom_test <- bom_only_all[i,]
  bom_model <- randomForest(formula = as.factor(author) ~ . -js -total_words, data = bom_train)
  pred_df$pred[i] <- predict(bom_model, bom_test) %>% as.character
}

pred_df
```