Analyzing the Differences in How Fiction and Nonfiction Authors Describe Characters and Settings

Aaron Kruchten

Introduction

The reasons fiction and nonfiction authors write are many and varied. Fiction works contain genres such as fantasy, children's fiction, crime fiction, and historical fiction. Non-fiction works span the areas of scientific works, biographies, philosophy, and many others. While some fiction works are intended simply to entertain their audience and provide a passtime, many fiction works will also serve an additional purpose. Some examples of these purposes include changing the reader's perspective and critiquing our society. Similarly, some non-fiction texts are made simply to serve as a passtime or to be a collection of facts. Some examples of these would be books debunking popular myths and encyclopedias. However, works such as a biographies or historical works will also often attempt to change the reader's perspective and offer critiques of society and our world. Because nonfiction works primarily revolve around facts, real events, and real people, the way in which authors of fiction and nonfiction works serve their purpose is likely different.

We hypothesize that fiction writers will often attempt to change their reader's persepctive by creating a fictional world that is an exaggeration or hyperbole of our own. For example, the book *1984* depicts a near dystopian future where an entire police force was dedicated to fighting something called thoughtcrime which was essentially any set of beliefs against or differing from those who held political power. Because fiction writers create imaginary worlds, we hypothesize their writing will be different from nonfiction writing. We hypothesize that fiction writers may put more effort into creating and describing the world, characters, and setting they are describing than nonfiction writers. Therefore, in this analysis we will attempt to understand the differences between how fiction and nonfiction authors write. Specifically, we will attempt to investigate the differences in how fiction authors describe the characters and settings of their writings. Although this analysis has no clear direct importance or applications, it may increase our fundamental understanding of both fiction and nonfiction writing.

Data

For this analysis we have chosen to analyze 10 texts, 5 fiction and 5 nonfiction. The titles of these texts, the category they are in (fiction or non-fiction), and a quick description is available in the table below[1].

Title	Category	Description
Pride and Prejudice	F	A romantic novel of manners, a work of fiction which recreates a social world, published in 1813
1984	F	A dystopian novel written in 1949 that depicts the government overreach and totalitarianism
Fahrenheit 451	F	A dystopian novel that depicts a world where books are banned and any that are found are burned
To Kill a MockingBird	F	Depicts the life in a small town during the trial of a black man accuses of raping a white woman
A Tale of Two Cities	F	A historical fiction novel depicting France and England just before and during the French Revolution
A Room of One's Own	NF	An important early feminist text published in 1929
Outwtting the Hun	NF	A historical account of a an American soldier being captured and escaping from German soldiers in WWI
Frederick Douglass	NF	A former American Slave's account of growing up in slavery and then eventually escaping to the North
The Footprints of Time	NF	An analysis of the American system of Government
The Profits of Religion	NF	A snapshot of the religious movements in the United States just before WW1

We chose these texts subject to several different criteria. First, our choices needed to be freely available and easily accessibile in a text file format. Second, we wanted texts which probably have a purpose besides simply serving as a passtime. The fiction texts we chose are all very well know, often read in a high school curriculum, and are well known to possess themes and motives besides serving as a passtime. We chose the nonfiction texts specifically because some of them explore similar themes as the nonfiction texts. We felt that trying to choose nonfiction books which explore

[1] 1.The 8th book in the table has its title listed as "Frederick Douglas". The full title is "Narrative of the Life of Frederick Douglass, an American Slave", but we abbreviated because of space constraints

similar themes to our fiction texts may reduce the noise of our analysis especially when finding collocates. For example, *The footprints of time* analyzes the American Government. We decided to include this story because two of our fiction novels 1984 and *Fahrenheit* 451 both primarily serve the purpose of critiquing the government.

I have omitted a table which reports the number of tokens (as is convention) for each text in this section, because we have included a very similar table in our vocabulary size analysis.

Methods

In this analysis, we would like to study how fiction authors describe their characters and setting and how it differs from nonfiction writing if it does differ. In previous work, (Coffee Break Experiment One), I looked at the differences in adjective use between nonfiction and fiction writing. Although we were able to find adjectives which occurred more frequently in fiction than nonficition, it was difficult to find any clear differences in the adjectives that occurred more frequently in one group over the other. Therefore, even though analyzing the differences in adjectives seems like a clear first step in analyzing the differences between how fiction and nonfiction writers write and describe their characters and setting, we will not be doing this in this analysis because I have done it in previous work.

Our first step in this analysis will be to look at the differences in the size of vocabulary of fiction and nonfiction texts. In this part of the analysis, we are trying to better understand how fiction and nonfiction authors create and describe their settings. Because we are primarily interested in investigating how authors describe their settings, it may seem that we should only look at the vocabulary size of descriptive words such as adverbs and adjectives. Even though we will do this, it will also be useful to look at all words because authors can be descriptive with their verbs and nouns. For example, names such as *Thought Police* are descriptive of what they are naming, and it may be true that fiction writers create and use more specific and descriptive names than nonfiction writers.

This approach of looking at vocabulary size has been used in previous computational approaches to stylometric work (Le et. al 2011), where they studied vocabulary size to look for signs of alzheimers and dementia in aging authors. We hypothesize that because fiction authors tend to create imaginary worlds, they may tend to use more descriptive and varied language and therefore their vocabulary may be larger. We will look at both their entire vocabulary size and the vocabulary size of just descriptive words such as adjectives and adverbs. Vocabulary size will be measured as $\frac{Unique\ word\ tokens\ used}{Total\ word\ tokens\ used}$.

Second, we are also concerned with how the authors describe characters and people. We hypothesize that fiction authors will tend to create more eccentric and varied characters than nonfiction authors. In order to study this, we will look at which words collocate with some of the very common personal pronouns such as 'he' and 'she'. In order to do this, we will find the most common collocates for both 'he' and 'she' as measured by mutual information. We will then select a large threshold for mutual information and select a small number of collocates to visualize graphically.

Total Vocabulary Size

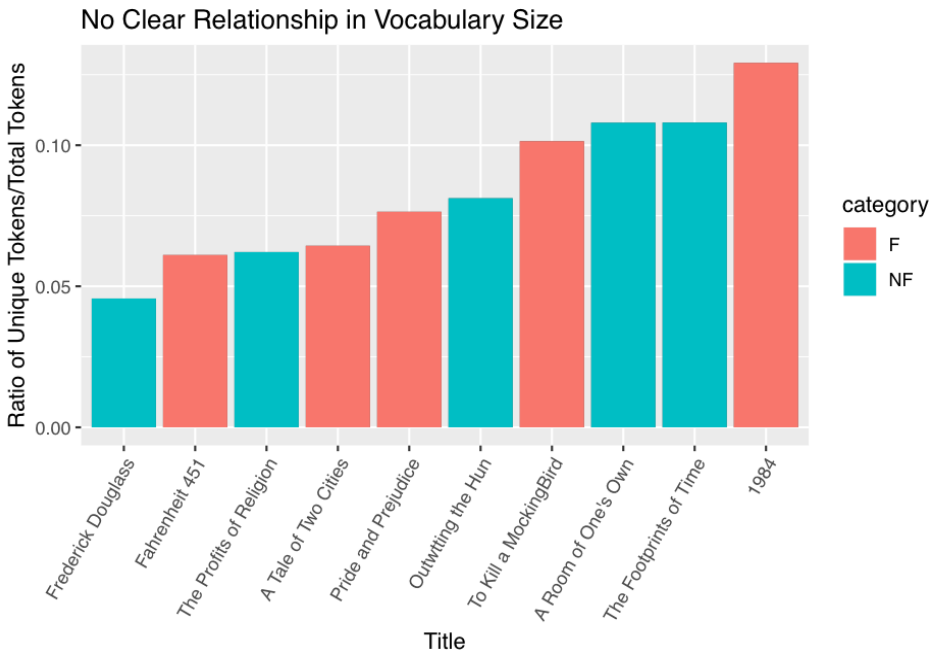
Table 1 contains the number of tokens for each text, the number of unique tokens, and the ratio of unique tokens over total tokens. We tokenized by words. We also plotted the ratio column in table one on the next page in order to better visualize this data. Table 2 contains the total number of tokens and unique tokens for fiction and nonfiction texts. Looking at table 2 and the plot on the next page, it appears that fiction texts do not have a vocabulary size larger than nonfiction texts.

Table 1: Vocabulary Size of Each Work in Our Corpus

Title	Author	Category	Unique.Tokens	Total.Tokens	Ratio
Pride and Prejudice	Jane Austen	F	10035	131327	0.0764123
1984	Geroge Orwell	F	6138	47514	0.1291830
Fahrenheit 451	Ray Bradbury	F	11948	195625	0.0610760
To Kill a MockingBird	Harper Lee	F	5778	56971	0.1014200
A Tale of Two Cities	Charles Dickens	F	12306	191184	0.0643673
A Room of One's Own	Virginia Woolf	NF	6009	55654	0.1079707
Outwtting the Hun	Pat O'Brien	NF	5849	72026	0.0812068
Frederick Douglass	Frederick Douglass	NF	7466	163499	0.0456639
The Footprints of Time	Charles Bancroft	NF	12033	111392	0.1080239
The Profits of Religion	Upton Sinclair	NF	9276	149323	0.0621204

Table 2: Vocabulary Size of Fiction and Nonfiction

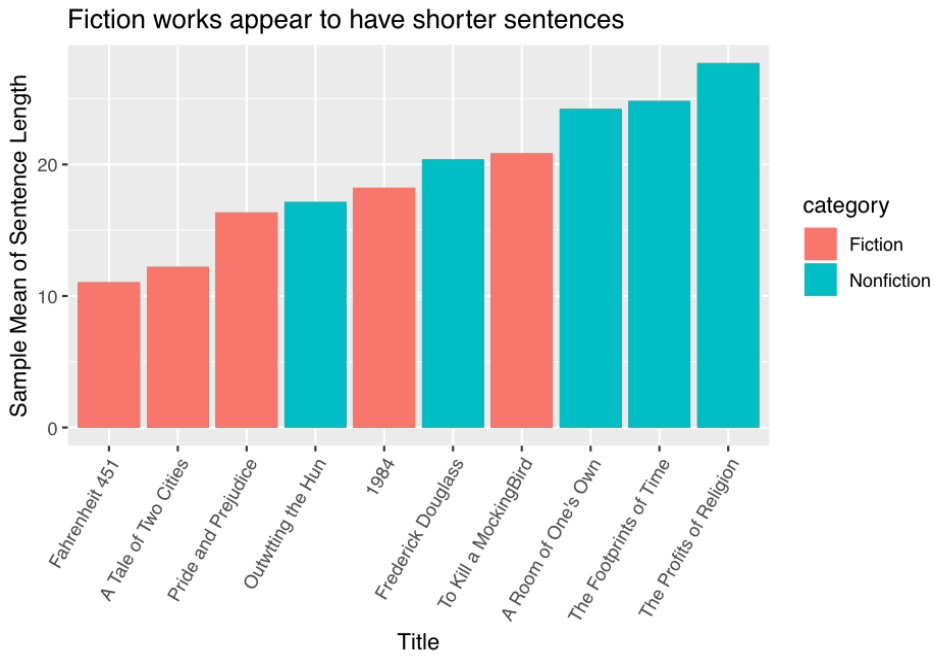
Category	Unique.Token	Total.Tokens	Ratio
Fiction	46205	622621	0.0742105
Nonfiction	40633	551894	0.0736246



Descriptive Word Vocabulary Size

From the previous section, it appears that there is not a substantial difference in vocabulary size between the fiction and nonfiction texts in our corpus. However, it may be true that fiction writers tend to use more descriptive words, such as adjectives and adverbs, than nonfiction writers. Before we continue with our descriptive word vocabulary size analysis, we will first address a token imbalance we will have in this analysis.

Because of the time it takes to part of speech tag these large texts, we first take a random sample of 1000 sentences from each of the fiction and non-fiction texts. When we randomly sample theses sentences and then further tokenize these sentences into words, we find that we have substantially fewer fiction tokens than nonfiction tokens. This implies that on average fiction texts have fewer words in a sentence than nonfiction texts. The following graph displays the average sentence length (measured as the number of nonunique words in a sentence) of each of the books in our corpus.



We hypothesize this difference is largely due to a focus on dialogue in fiction writing over nonfiction writing. Fiction writing will often contain many very short sentences which are simply one of the characters saying something such as "Hello". We considered formally testing this phenomenon to see if there is a difference in the mean sentence length of fiction and nonfiction writing, but decided not to for several reasons. First, although this test may give us some insight into the differences between fiction and nonfiction writing, we don't believe it is pertinent to the questions we are studying in this analysis. Second, although we have large sample sizes, and we would be considering the differences in mean sentence length between two groups of 5000 sentences, the sentences that occur in these groups are not independent. They come from only 5 authors for each group. In order to properly study this difference, it would likely be more wise to select a large sample of fiction and nonfiction texts, and draw only a few sentences from each.

Now, that we have addressed the issue of token imbalance, we will look at the vocabulary size of descriptive words. The below table displays the total tokens, unique tokens, and ratio as we did in the previous section.

Table 3: Descriptive Word Vocabulary Size in Fiction and Nonfiction

Category	Total.Tokens	Unique.Tokens	Ratio
Fiction	10559	3500	0.3314708
Nonfiction	13484	4423	0.3280184

We find that the ratio of $\frac{Unique\ Tokens}{Total\ Tokens}$ is slightly larger for fiction than it is for nonfiction, however this difference is very small. Even though it seems likely that this difference is not statistically significant, we would still like to formally test this. There are two possible ways we could do this.

- One way we could do this is predict total tokens from unique tokens as well as a factor with the levels of fiction and nonfiction. Although it is difficult to tell from a histogram the distribution of total tokens, because total tokens is

a count with no upper limit, poisson regression is likely appropriate for this task (Shalizi, C.R., pg 240). However, the mean of our total tokens is 2369.8 while the variance is 441298.6. This implies that the poission distribution is *Overdispersed*. Therefore, the result of this test would likely not be trustworthy.

- The second way we could perform this test is to do the same prediction as above but with a linear model. Even though the relationship between total tokens and unique tokens is likely not linear, the other assumptions of linear regression such as independence of errors and no heteroskedasticity seem to be met.

Therefore, in order to perform this test, we will use the second approach, and use a linear model to predict total tokens from unique tokens. A summary of our model is listed in the table below.

Table 4: Linear Model Summary

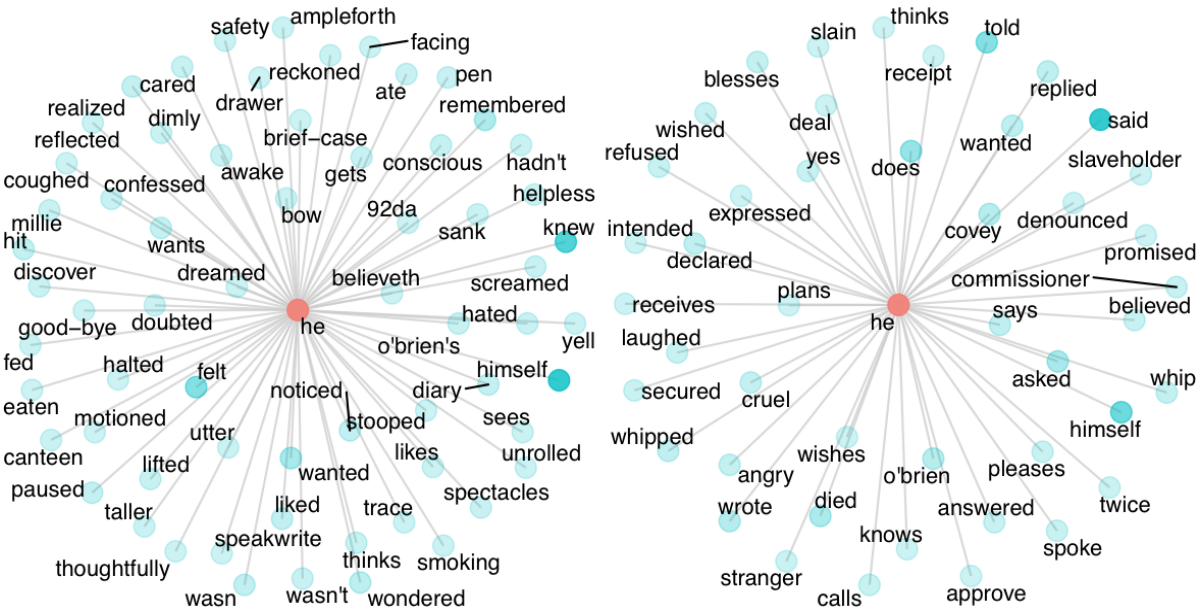
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	304.881696	723.411919	0.4214496	0.6860666
Unique	2.581312	0.990801	2.6052778	0.0351541
factorNonfiction	108.489830	343.583786	0.3157595	0.7613913

We see that the p-value for the coefficient for nonfiction is approximately .761. This implies that there is no evidence of a difference between the number of descriptive words used by fiction and nonfiction writers.

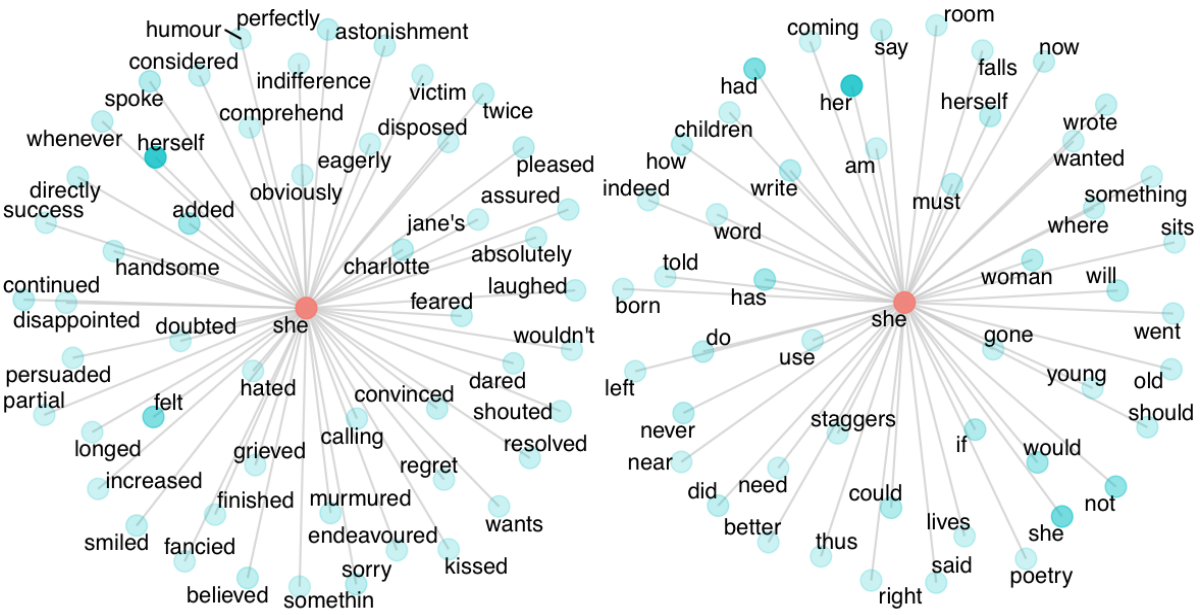
Pronoun Collocations

In this section, we will look at the collocates of the pronouns he and she. By looking at the collocates of these personal pronouns, we will be able to see how the authors describe the characters in their stories. We formed collocate plots for the pronouns he and she for both fiction texts and nonfiction texts.

The collocate plots for fiction are on the **left side** of the page below. The collocate plots for nonfiction are on the **right side**[2].



[2]These plots were formed with the ggraph package. I wanted to put a caption or title on each but for some reason this causes issues when knitting to pdf



It appears that the function and type of words that collocate with 'he' are mostly similar in both fiction and nonfiction texts. Additionally, it appears that that the collocates with 'she' in our fiction sample seem to be more descriptive of emotional states than nonfiction. Words such as astonishment, indifference, and feared collocate with she in fiction while there are very few words in the nonfiction collocates with 'she' that are likely used to describe emotional states. This could be due to commonly held gender biases such as that women are much more emotional than men, however this analysis is definitely not conclusive. Because we have only included ten texts in this analysis, this is likely not representative of all fiction and nonfiction writing. Many of our nonfiction texts likely do not have many female characters. For example, *Outwitting The Hun* is mostly an account of the frontlines of World War I. Additionally, *A Room of One's Own* is a famous and influential feminist text, and therefore the way this text writes about women is likely not representative of all nonfiction texts.

Conclusion

In this analysis, we took two main approaches to identifying and understanding the differences between how fiction and nonfiction writers describe their characters and settings. Our first approach was an analysis of vocabulary size where we looked at both the total vocabulary size as well as vocabulary size of just descriptive words. In both cases, we did not see a clear difference in vocabulary size. This is contrary to our hypothesis that fiction writers would have a large vocabulary size because they would put more effort into describing their characters and setting than nonfiction writers.

In the second part of our analysis, we formed collocate graphs for the pronouns 'he' and 'she' for both fiction and nonfiction texts. We then compared these plots visually. We found that the function of collocates for he in both fiction and nonfiction texts seemed to be similar. However, the collocates for 'she' in fiction and nonfiction appeared to be different. There appeared to be many more words that focused on describing an emotional state than in the nonfiction text. However, we are hesitant to believe that this is reflective of all fiction and nonfiction texts.

In conclusion, we find that our hypotheses were mostly incorrect and not supported by this analysis. There does not appear to be a significant difference in the ways that fiction and nonfiction writers describe their characters and settings. And, it does not appear to be true that fiction writers put more effort towards describing their characters and setting than nonfiction writers. However, further analysis with a larger and likely more representative sample of fiction and nonfiction texts is needed to confirm these conclusions.

References

Xuan Le, Ian Lancashire, Graeme Hirst, Regina Jokel, Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists, Literary and Linguistic Computing, Volume 26, Issue 4, December 2011, Pages 435–461, https://doi.org/10.1093/llc/fqr013

Shalizi C.R., Advanced Data Analysis from an Elementary Point of View (Cambridge University Press), http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV

Code Appendix

```
#chunk one
library(knitr)
library(tidyverse)
library(kableExtra)
title <- c("Pride and Prejudice","1984","Fahrenheit 451","To Kill a MockingBird","A Tale of Two Citi
category <- c("F","F","F","F","F","NF","NF","NF","NF","NF")
author <- c("Jane Austen","Geroge Orwell","Ray Bradbury","Harper Lee","Charles Dickens","Virginia Wo
Description <- c("A romantic novel of manners, a work of fiction which recreates a social world, pub
                 "A dystopian novel written in 1949 that depicts the government overreach and totali
                 "A dystopian novel that depicts a world where books are banned and any that are fou
                 "Depicts the life in a small town during the trial of a black man accuses of raping
                 "A historical fiction novel depicting France and England just before and during the
                 "An important early feminist text published in 1929",
                 "A historical account of a an American soldier being captured and escaping from Ger
                 "A former American Slave's account of growing up in slavery and then eventually esc
                 "An analysis of the American system of Government",
                 "A snapshot of the religious movements in the United States just before WW1")
frame <- data.frame("Title" = title,"Category" = category,"Description" = Description)
kable(frame) %>%  kable_styling(latex_options=c("scale_down","HOLD_POSITION"))

#chunk two
set.seed(1)
source("/Users/aaronkruchten/Desktop/Text Analysis/textstat_tools-master 2/functions/helper_functio
library(quanteda)
library(tidyverse)
library(ggplot2)
library(knitr)
library(kableExtra)
files <- list.files(path  = "/Users/aaronkruchten/Desktop/Final project ",full.names = TRUE)
files_lst <- readtext_lite(files)
corpus <- corpus(files_lst)
tokens_vector <- c(131327,47514,195625,56971,191184,55654,72026,163499,111392,149323)
tokens_words <- tokens(corpus, include_docvars=TRUE, remove_punct = TRUE,
                       remove_numbers = TRUE, remove_symbols = TRUE, what = "word")
token_lengths <- tokens_words %>% lapply(FUN = unique) %>% lapply(FUN = length)
token_len_vector <- unlist(token_lengths,use.names = FALSE)
ratio_vector <- token_len_vector/tokens_vector

title <- c("Pride and Prejudice","1984","Fahrenheit 451","To Kill a MockingBird","A Tale of Two Citi
author <- c("Jane Austen","Geroge Orwell","Ray Bradbury","Harper Lee","Charles Dickens","Virginia Wo
category <- c("F","F","F","F","F","NF","NF","NF","NF","NF")
kable_frame <- data.frame("Title" = title,"Author" = author,"Category" = category,"Unique Tokens" =
```

```
#kable(kable_frame)  %>%  kable_styling(latex_options="scale_down")

total_fiction_tokens <- sum(tokens_vector[c(1:5)])
total_nonfiction_tokens <- sum(tokens_vector[c(6:10)])
total_unique_tokens <- sum(token_len_vector[c(1:5)])
total_unique_nonfiction_tokens <- sum(token_len_vector[c(6:10)])
ratio_vector_combo <- c(total_unique_tokens/total_fiction_tokens,total_unique_nonfiction_tokens/tota
kable_frame_two <- data.frame("Category" = c("Fiction","Nonfiction"),"Unique Token" = c(total_uniqu


#chunk three
library(knitr)
new_title <- c("Pride and Prejudice","1984","Fahrenheit 451","To Kill a MockingBird","A Tale of Two


kable_frame <- data.frame("Title" = new_title,"Author" = author,"Category" = category,"Unique Token
kable(kable_frame,caption = "Vocabulary Size of Each Work in Our Corpus")  %>%  kable_styling(latex

kable(kable_frame_two,caption = "Vocabulary Size of Fiction and Nonfiction") %>% kable_styling(posi

ggplot(data = kable_frame,aes(x = reorder(new_title,ratio_vector),y  = ratio_vector)) + geom_bar(st

#chunk four
set.seed(123)
source("/Users/aaronkruchten/Desktop/Text Analysis/textstat_tools-master 2/functions/helper_functio
library(spacyr)
library(quanteda)
library(tidyverse)
library(knitr)
files <- list.files(path  = "/Users/aaronkruchten/Desktop/Final project ",full.names = TRUE)
files_lst <- readtext_lite(files)
fiction_files <- files_lst[c(1,3,4,8,10),]
nonfiction_files <- files_lst[c(2,5,6,7,9),]

fiction_corpus <- corpus(fiction_files)
nonfiction_corpus <- corpus(nonfiction_files)



fiction_sent_toks <- tokens(fiction_corpus, what = "sentence")
idx <- seq(from = 1, to = length(fiction_sent_toks))
fiction_sample <- lapply(idx, function(i) sample(unlist(fiction_sent_toks[i]), 1000))
fiction_sample <- lapply(idx, function(i) paste(unlist(fiction_sample[i]), collapse = " "))
fiction_sample <- data.frame(text = do.call(rbind, fiction_sample), stringsAsFactors = F)
fiction_sample <- bind_cols(doc_id = rownames(fiction_files$doc_id),fiction_sample)


nonfiction_sent_toks <- tokens(nonfiction_corpus, what = "sentence")
idx <- seq(from = 1, to = length(nonfiction_sent_toks))
nonfiction_sample <- lapply(idx, function(i) sample(unlist(nonfiction_sent_toks[i]), 1000))
nonfiction_sample <- lapply(idx, function(i) paste(unlist(nonfiction_sample[i]), collapse = " "))
nonfiction_sample <- data.frame(text = do.call(rbind, nonfiction_sample), stringsAsFactors = F)
nonfiction_sample <- bind_cols(doc_id = rownames(nonfiction_files$doc_id),nonfiction_sample)
```

```
fiction_corpus <- corpus(fiction_sample)
nonfiction_corpus <- corpus(nonfiction_sample)

fiction_prsd <- spacy_parse(fiction_corpus,pos = TRUE,tag = TRUE)
nonfiction_prsd <- spacy_parse(nonfiction_corpus,pos = TRUE,tag = TRUE)

fiction_tokens <- as.tokens(fiction_prsd,include_pos = "tag",concatenator = "_")
nonfiction_tokens <- as.tokens(nonfiction_prsd,include_pos = "tag",concatenator = "_")

fiction_tokens <- tokens_select(fiction_tokens, "_[A-Z]", selection = "keep", valuetype = "regex",

# This filters any that don't have a word or digit chacter before the underscore.
fiction_tokens <- tokens_select(fiction_tokens, "\\W_", selection = "remove", valuetype = "regex")

# And lastly any tokeans with a digit immediate before the underscore.
fiction_tokens <- tokens_select(fiction_tokens, "\\d_", selection = "remove", valuetype = "regex")

nonfiction_tokens <- tokens_select(nonfiction_tokens, "_[A-Z]", selection = "keep", valuetype = "re

# This filters any that don't have a word or digit chacter before the underscore.
nonfiction_tokens <- tokens_select(nonfiction_tokens, "\\W_", selection = "remove", valuetype = "re

# And lastly any tokeans with a digit immediate before the underscore.
nonfiction_tokens <- tokens_select(nonfiction_tokens, "\\d_", selection = "remove", valuetype = "re

fiction_individual_descriptive_tokens_unique <- c()
nonfiction_individual_descriptive_tokens_unique <- c()
fiction_individual_descriptive_tokens <- c()
nonfiction_individual_descriptive_tokens <- c()
fiction_descriptive_tokens <- tokens_select(fiction_tokens, pattern = c("*_JJ*","*_RB*"))
nonfiction_descriptive_tokens <- tokens_select(nonfiction_tokens,pattern = c("*_JJ*","*_RB*"))
for(i in 1:5){
  fiction_individual_descriptive_tokens_unique[i] = length(unique(fiction_descriptive_tokens[[i]]))
  nonfiction_individual_descriptive_tokens_unique[i] = length(unique(nonfiction_descriptive_tokens[
  fiction_individual_descriptive_tokens[i] = length(fiction_descriptive_tokens[[i]])
  nonfiction_individual_descriptive_tokens[i] = length(nonfiction_descriptive_tokens[[i]])
}


total_fiction_descriptive = sum(fiction_individual_descriptive_tokens)
total_nonfiction_descriptive = sum(nonfiction_individual_descriptive_tokens)
total_fiction_unique_descriptive = sum(fiction_individual_descriptive_tokens_unique)
total_nonfiction_unique_descriptive = sum(nonfiction_individual_descriptive_tokens_unique)
ratio_fiction = total_fiction_unique_descriptive/total_fiction_descriptive
ratio_nonfiction = total_nonfiction_unique_descriptive/total_nonfiction_descriptive

kable_frame <- data.frame("Category" = c("Fiction","Nonfiction"),"Total Tokens" = c(total_fiction_d

#chunk five
new_title <- c("Pride and Prejudice","1984","Fahrenheit 451","To Kill a MockingBird","A Tale of Two
mean_sentence_length_frame <- data.frame("Titles" = new_title,"s_length" = c(fiction_tokens_counts/
ggplot(data = mean_sentence_length_frame,aes(x = reorder(Titles,s_length),y  = s_length )) + geom_b

#chunk six
```

```r
kable(kable_frame,caption = "Descriptive Word Vocabulary Size in Fiction and Nonfiction") %>% kable_
unique_counts <- c(fiction_individual_descriptive_tokens_unique,nonfiction_individual_descriptive_to
total <- c(fiction_individual_descriptive_tokens,nonfiction_individual_descriptive_tokens)
fiction_factor <- rep("Fiction",5)
nonfiction_factor <- rep("Nonfiction",5)
factor <- c(fiction_factor,nonfiction_factor)

data_frame <- data.frame("Unique" = unique_counts,"Total" = total,"Type" = factor,"Title" = title)
model <- glm(Total ~ Unique + factor,data = data_frame )

#chunk seven
summary_linear <- summary(model)
kable(summary_linear$coefficients,caption = "Linear Model Summary") %>% kable_styling(position = "c


library(tidygraph)
library(ggraph)
library(ggplot2)
library(tidyverse)
library(knitr)
source("/Users/aaronkruchten/Desktop/Text Analysis/textstat_tools-master 2/functions/collocations_fu

files <- list.files(path  = "/Users/aaronkruchten/Desktop/Final project ",full.names = TRUE)
files_lst <- readtext_lite(files)
fiction_files <- files_lst[c(1,3,4,8,10),]
nonfiction_files <- files_lst[c(2,5,6,7,9),]

fiction_corpus <- corpus(fiction_files)
nonfiction_corpus <- corpus(nonfiction_files)
fiction_tokens <- tokens(fiction_corpus,what = "word",remove_punct = T)
nonfiction_tokens <- tokens(nonfiction_corpus,what = "word",remove_punct = T)

fiction_he_collocates <- collocates_by_MI(fiction_tokens,"he",5,5)
fiction_he_collocates <- fiction_he_collocates %>% filter(col_freq >= 5.2 & MI_1 > 4.9)
fiction_he_collocates <- fiction_he_collocates[-19,]
net_fiction_he <- col_network(fiction_he_collocates)
ggraph(net_fiction_he) +

  geom_edge_link(color = "gray80", alpha = .75) +

  geom_node_point(aes(alpha = node_weight, size = 3, color = n_intersects)) +

  geom_node_text(aes(label = label), repel = T) +

  scale_alpha(range = c(0.2, 0.8)) +

  theme_graph() +

  theme(legend.position="none")

nonfiction_he_collocates <- collocates_by_MI(nonfiction_tokens,"he",5,5)
nonfiction_he_collocates <- nonfiction_he_collocates %>% filter(col_freq >= 5.3 & MI_1 > 5.4)
nonfiction_net_he <- col_network(nonfiction_he_collocates)
ggraph(nonfiction_net_he) +
```

```
  geom_edge_link(color = "gray80", alpha = .75) +

  geom_node_point(aes(alpha = node_weight, size = 3, color = n_intersects)) +

  geom_node_text(aes(label = label), repel = T) +

  scale_alpha(range = c(0.2, 0.8)) +

  theme_graph() +

  theme(legend.position="none")
par(mfrow = c(2,1))




library(knitr)
fiction_she_collocates <- collocates_by_MI(fiction_tokens,"she",5,5)
fiction_she_collocates <- fiction_she_collocates %>% filter(col_freq >= 5.4 & MI_1 > 5)
fiction_she_collocates = fiction_she_collocates[-9,]
fiction_she_collocates = fiction_she_collocates[-16,]
net_fiction_she <- col_network(fiction_she_collocates)
ggraph(net_fiction_she) +

  geom_edge_link(color = "gray80", alpha = .75) +

  geom_node_point(aes(alpha = node_weight, size = 3, color = n_intersects)) +

  geom_node_text(aes(label = label), repel = T) +

  scale_alpha(range = c(0.2, 0.8)) +

  theme_graph() +

  theme(legend.position="none")

nonfiction_she_collocates <- collocates_by_MI(nonfiction_tokens,"she",5,5)
nonfiction_she_collocates <- nonfiction_she_collocates %>% filter(col_freq >= 5 & MI_1 > 4.5)
nonfiction_net_she <- col_network(nonfiction_she_collocates)
ggraph(nonfiction_net_she) +

  geom_edge_link(color = "gray80", alpha = .75) +

  geom_node_point(aes(alpha = node_weight, size = 3, color = n_intersects)) +

  geom_node_text(aes(label = label), repel = T) +

  scale_alpha(range = c(0.2, 0.8)) +

  theme_graph() +

  theme(legend.position="none")
par(mfrow = c(2,1))
```