Searching for Structural Differences between Fiction and Nonfiction

Marshall Pauley

December 7 2022

Introduction

Fiction and Nonfiction differ on their subject matter, by definition. But I'm curious whether there's an underlying difference in the structure or the prose that separates fiction and nonfiction as well. The data is the text of 12 books, six from two authors, pulled from Project Gutenberg with the gutenbergr package.

Dependencies

You will need to have the following packages installed:

- dplyr
- forcats
- ggplot2
- gutenbergr
- scales
- stringr
- textdata
- tidyr
- tidytext

```
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(forcats)
library(ggplot2)
library(gutenbergr)
library(scales)
library(stringr)
library(textdata)
library(tidyr)
library(tidytext)
```

Import Data

```
raw_books <- gutenberg_download(gutenberg_id = c("2530", "2539", "50289", "39928", "56506"
```

```
Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest

Using mirror http://aleph.gutenberg.org
```

I had some trouble with the gutenbergr package, as ebook number 69454 is one of A. R. Wallace's books, but gutenbergr silently failed to download it. I side-stepped this by switching to a different book by the same author.

Cleaning the data involves removing the stop words, which aren't useful for most analyses. We also need to preserve the line numbers, so that the words remain ordered.

```
# code based off code from _Text Mining with R: a Tidy Approach_ by Julia Silge & David Ro
clean_books <- raw_books |>
  mutate(linenumber = row_number()) |>
  unnest_tokens(word, text) |>
  anti_join(stop_words)
```

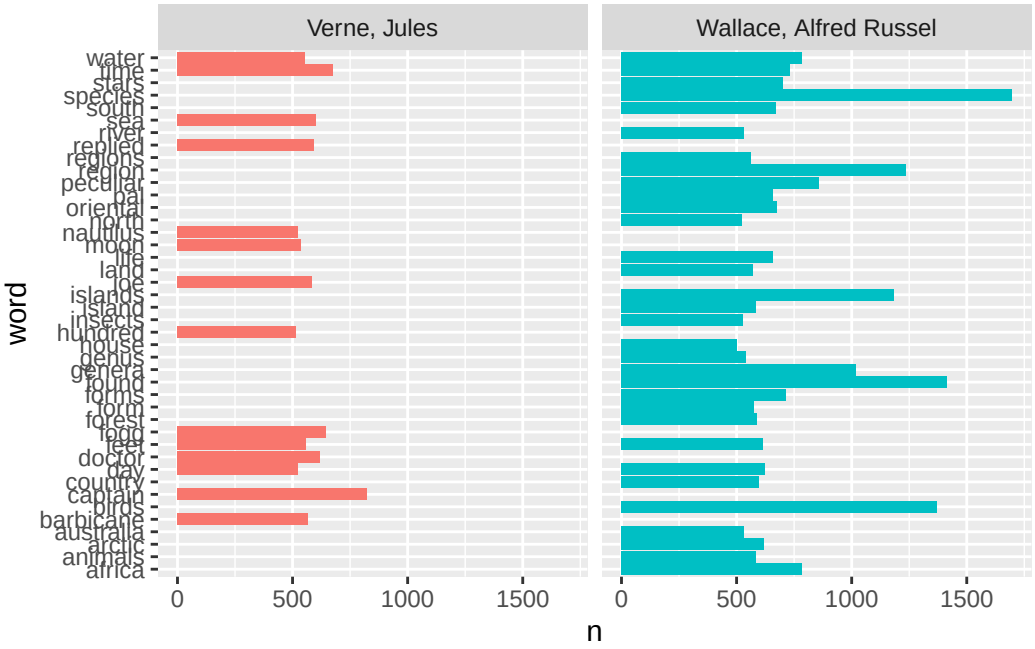
```
Joining, by = "word"
```

This leaves us with workable data.

EDA & basic vis

I began by seeing the highest frequency words for both authors. Here I've filtered for words that appear at least 500 times.

```
clean_books |>
  mutate(word = str_extract(word, "[a-z']+")) |>
  count(author, word, sort = TRUE) |>
  group_by(author) |>
  filter(n > 500) |>
  na.omit() |>
  ggplot(mapping = aes(n, word, fill = author)) +
  geom_col() +
  facet_wrap(~author, ncol = 2) +
  theme(legend.position = "none")
```



The fact that Wallace has more words above 500 uses is likely an artifact of the selected works being longer than those selected from Verne. That said, this visual does give us a good sense

for the words (and types of words) that each writer uses most. A possible outlier is "species", which an earlier exploratory analysis showed was used by Wallace more than some stop words, but the high frequency of technical terminology could be a differentiating trait of nonfiction, so it was decided to leave the word in.

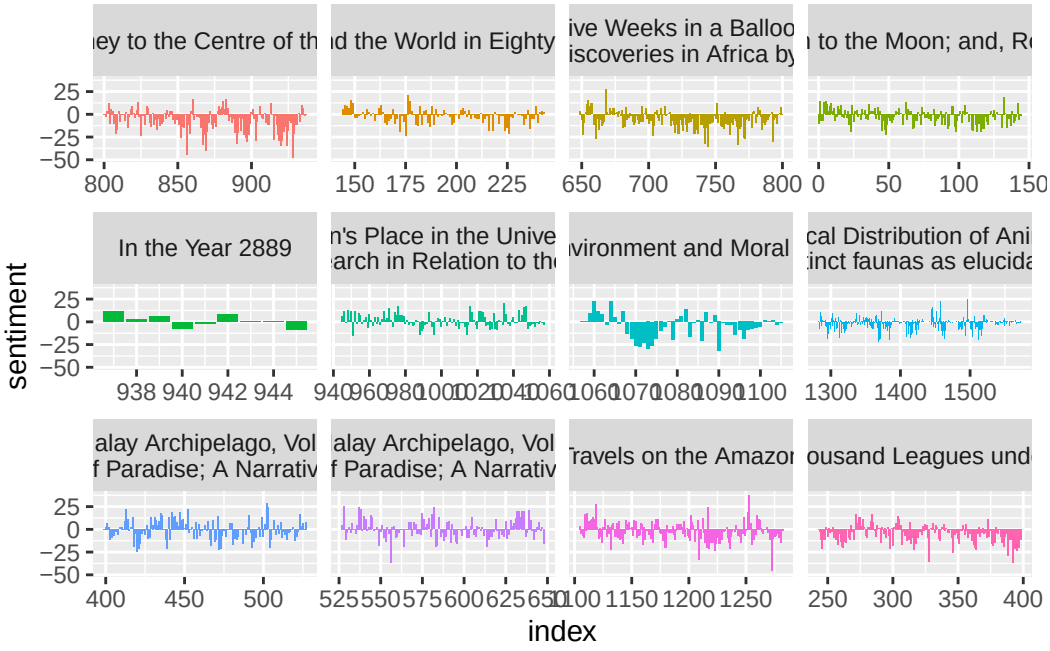
Analysis

Wallace's most frequent words are more technical than Verne's. Not just "species", but "genera" and "genus" too, despite the fact that Wallace was often writing for a general audience, not just fellow scientists. Technical jargon tends to lack strong sentiment, so I ran a sentiment analysis.

```
bing_sent <- get_sentiments("bing")

# code based off code from _Text Mining with R: a Tidy Approach_ by Julia Silge & David Ro
book_sentiment <- clean_books |>
  inner_join(bing_sent) |>
  count(title, index = linenumber %/% 80, sentiment) |>
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) |>
  mutate(sentiment = positive - negative)

ggplot(book_sentiment, aes(index, sentiment, fill = title)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~title, ncol = 4, scales = "free_x")
```



Verne goes consistently into the negative towards the end, the climax. This is fitting for drama, which tends to have a "darkest hour" moment before the conflict resolves. Wallace is more variable, with his most technical works sampled, *Geographic Distribution of Animals…* and *Man's Place in the Universe*, wobbling near the center while his more public-oriented essays and travelogues show stronger positive and negative sentiments, but those sentiments don't group up into runs of negative or positive as long as Verne's (excepting *Moral Progress*, which is an outlier since it was more of a persuasive essay.)

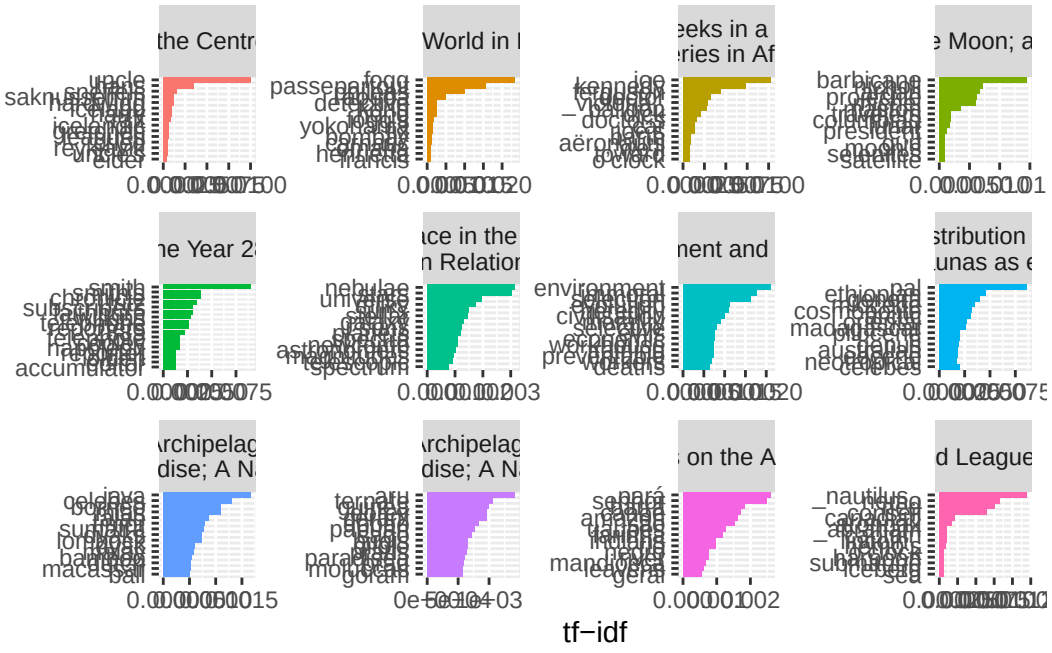
```
# code based off code from _Text Mining with R: a Tidy Approach_ by Julia Silge & David Ro
book_words <- raw_books |>
  unnest_tokens(word, text) |>
  count(title, word, sort = TRUE)

total_words <- book_words |>
  group_by(title) |>
  summarise(total = sum(n))

book_words <- left_join(book_words, total_words)

book_tf_idf <- book_words |>
  bind_tf_idf(word, title, n)
```

```
book_tf_idf |>
  group_by(title) |>
  slice_max(tf_idf, n = 15) |>
  ungroup() |>
  ggplot(aes(tf_idf, fct_reorder(word, tf_idf), fill = title)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~title, ncol = 4, scales = "free") +
  labs(x = "tf-idf", y = NULL)
```



Next I used inverse document frequency to find words that are common within each text, but not too common. This approach aims to find the words that define a text. Wallace's texts are built around nouns that clearly pertain to the book's particular subject. Verne's books are dominated by character names, character roles, hazards (icebergs, for example), and vehicles.

Conclusion

My question isn't entirely answered. Characters being absent from some nonfiction is hardly a surprise, but the pattern of sentiment Verne's fiction shows (and Wallace's nonfiction mostly lacks) suggests a major difference. Several literature and creative writing teachers have insisted that reversals in fortune, and the accompanying shift in sentiment, are a fundamental of writing,

without which a narrative's energy dries up. I had assumed that this applied to all writing, nonfiction included. Visualizing the sentiment of Wallace's books suggests that nonfiction might be free of this dramatic necessity, or at least nonfiction can safely ignore traditional structures of fiction, like 3-Act or Hero's Journey.

I suffered a recent disappointment at Verne's hands, which may have biased me against him. To compensate for this, I picked a nonfiction writer who was his contemporary (Wallace was born and died within a few years of Verne) and whose writing had good chance to overlap Verne's: Verne wrote many adventures of discovery, Wallace lived two.

Still, it has come to my attention more recently that Verne suffered many poor translations into English, and I do not yet know whether the translation Project Gutenberg hosts is considered a good or poor one. A more fair future analysis should compare authors who wrote in the same language.