STAT325 (Text Analytics) Moby-Dick Analysis

Adam Rogers (arogers24@amherst.edu)

2024-03-15

Background

Herman Melville's *Moby-Dick* (Melville 1851) is an unparalleled masterpiece in American literature, a work that combines deep philosophical inquiry with a compelling narrative of adventure and obsession. Initially met with mixed reviews and modest success upon its publication in 1851, the novel's true brilliance was not universally recognized until well after Melville's death. This posthumous recognition marks *Moby-Dick* as a seminal work, a testament to Melville's genius, which only time could unveil.

Melville's magnum opus is distinguished not only by its ambitious themes and complex characters but also by its striking variation in literary style. *Moby-Dick* oscillates between narrative storytelling, centered around the perilous journey of the Pequod and its diverse crew, and expansive expository sections that delve into the intricacies of whaling life and maritime culture. This stylistic duality reflects Melville's personal experiences aboard a whaling vessel, infusing the narrative with authenticity and a palpable sense of the sea and its perils. In one of the earliest and most renowned critiques of *Moby-Dick,* Walter Bezanson highlights the public's interest in *Moby-Dick* as "direct narrative, as moral analogue, as modern source, and as spiritual autobiography."(Hillway and Mansfield 1953)

The novel's narrative voice is another aspect of its complexity. While the story is ostensibly told through the eyes of Ishmael, a reflective and observant sailor, there are times when the narrative transcends his perspective. In many of the descriptive passages about whale anatomy, ship construction, and the philosophical underpinnings of whaling, the voice seems to shift from Ishmael to Melville himself. This transition suggests a blurring of lines between narrator and author, providing a window into Melville's own reflections and insights drawn from his time at sea.

The fluctuation between the detailed, almost encyclopedic exposition and the dynamic, character-driven narrative segments underscores the novel's thematic depth and complexity. It is this blend of narrative styles, driven by Melville's firsthand maritime experiences, that enriches *Moby-Dick* with its unique literary texture. The novel invites readers into a world where the boundaries between adventure narrative and meditative discourse are seamlessly intertwined.

This work aims to explore the intricate relationships between parts of speech and literary styles within *Moby-Dick,* facilitated by a Shiny App designed to allow users to navigate through the text seamlessly. By examining the stylistic shifts between Ishmael's storytelling and Melville's authoritative discourses, we seek to understand how these changes in narrative voice and style contribute to the novel's profound impact and lasting legacy in the landscape of American literature.

Wrangling

Literary scholars often assert that *Moby-Dick* contains diverse writing styles. However, substantiating this assertion with textual data prompts the question: what tangible evidence can we derive from the text to confirm this diversity? The segments that provide encyclopedic knowledge about whaling span a plethora of subjects. Melville's vivid descriptions range from various whale species and the critical role of the harpooner to the depictions of whales in art, the intricacies of blubber, the ship's components, whale anatomy, and other elements encapsulating the art and essence of whaling. While as readers we can discern these expository chapters from the narrative ones by virtue of literary flow, as data scientists, we must delve into identifying the distinct attributes of these divergent writing styles.

I posit that analyzing the frequency of various parts of speech could demarcate the descriptive chapters from the narrative ones. For instance, it stands to reason that chapters with more proper nouns likely delve more into character interactions and dialogues, given their portrayal of the Pequod's crew, encounters with different characters, and visits to new locations. Similarly, narrative chapters are expected to feature a higher verb count, resonating with the active engagement and interaction of characters.

To validate these hypotheses, we can extract the text of *Moby-Dick* from Project Gutenberg (ProjectGutenberg 2001) to create a data package. We then leverage the 'cleanNLP' package to parse the text of *Moby-Dick,* focusing on identifying parts of speech and named entities, such as "Person" and "Location." For this case, we are interested in the parts of speech. This part of the function appears more accurate than the NER (named-entity recognition).

```r
file_name <- "wrangling/anno_moby.Rds"
stopifnot(file.exists(file_name))
anno <- readRDS(file_name)

anno$entity[7,]
```

```
# A tibble: 1 x 6
```

```
  doc_id   sid   tid tid_end entity_type entity
   <int> <int> <int>   <int> <chr>       <chr>
1     24     1     7       8 PERSON      Sabbath afternoon
```

As you can see, NER recognizes the words “Sabbath afternoon” as a person, when it is a description of a time. Therefore, we will continue exploring parts of speech for our analysis. After loading and reading the `anno_moby.Rds` file, we can view the tokenized text.

```
head(anno$token)
```

```
# A tibble: 6 x 10
  doc_id   sid   tid token   token_with_ws lemma upos  xpos  tid_source relation
   <int> <int> <int> <chr>   <chr>         <chr> <chr> <chr>      <int> <chr>
1      1     1     1 CHAPTER "CHAPTER "    CHAP~ PROPN NNP            3 compound
2      1     1     2 I.      "I. "         I.    PROPN NNP            3 compound
3      1     1     3 LOOMIN~ "LOOMINGS"    LOOM~ PROPN NNP            0 root
4      2     1     1 Call    "Call "       call  VERB  VB             0 root
5      2     1     2 me      "me "         I     PRON  PRP            1 dobj
6      2     1     3 Ishmael "Ishmael"     Ishm~ PROPN NNP            1 oprd
```

This annotation maps every word and symbol of punctuation to a part of speech. We are interested in the `upos` variable, which tells us this information.

Since we are interested in exploring the parts of speech in each chapter, we need to add that variable. To do this, we take a cumulative sum of the number of times we encounter “CHAPTER” and `tid` = 1. We add this additional conditional statement to avoid the edge case where “CHAPTER” is in the text, but does not indicate a new chapter. Notice how we see this in chapter 32, where new sub-chapters begin.

```
chapter_example <- anno$token |>
  filter((token == "CHAPTER" | lag(token) == "CHAPTER") & doc_id >= 4164)

head(chapter_example)
```

```
# A tibble: 6 x 10
  doc_id   sid   tid token   token_with_ws lemma upos  xpos  tid_source relation
   <int> <int> <int> <chr>   <chr>         <chr> <chr> <chr>      <int> <chr>
1   4164     1     1 CHAPTER "CHAPTER "    chap~ NOUN  NN             2 compound
2   4164     1     2 XXXII   "XXXII"       XXXII PROPN NNP            0 root
3   4304     1     9 CHAPTER "CHAPTER "    CHAP~ PROPN NNP           10 compound
4   4304     1    10 I.      "I. "         I.    PROPN NNP            2 appos
```

```
5   4330     1     9 CHAPTER "CHAPTER "    chap~ NOUN  NN            10 compound
6   4330     1    10 II      "II"          II    PROPN NNP             2 appos
```

For this case, we still want those sub-chapters to be included in chapter 32, so we must condition.

```
token_with_chapters <- anno$token |>
  # Add chapter number to tokenized text
  # Use tid == 1 to avoid edge case where text "CHAPTER" is within a chapter
  mutate(chapter_number =
           cumsum(str_detect(token,
                             regex("^CHAPTER", ignore_case = FALSE)) & tid == 1))

head(token_with_chapters)
```

```
# A tibble: 6 x 11
  doc_id   sid   tid token   token_with_ws lemma upos  xpos  tid_source relation
   <int> <int> <int> <chr>   <chr>         <chr> <chr> <chr>      <int> <chr>
1      1     1     1 CHAPTER "CHAPTER "    CHAP~ PROPN NNP             3 compound
2      1     1     2 I.      "I. "         I.    PROPN NNP             3 compound
3      1     1     3 LOOMIN~ "LOOMINGS"    LOOM~ PROPN NNP             0 root
4      2     1     1 Call    "Call "       call  VERB  VB              0 root
5      2     1     2 me      "me "         I     PRON  PRP             1 dobj
6      2     1     3 Ishmael "Ishmael"     Ishm~ PROPN NNP             1 oprd
# i 1 more variable: chapter_number <int>
```

Subsequently, we aggregate relevant parts of speech per chapter from this tokenized text. This involves grouping the data by chapter, quantifying each part of speech, and determining their proportion relative to the chapter's total word count:

```
token_type_summary <- token_with_chapters |>
  group_by(chapter_number) |>
  summarize(total = n(),
            NOUN_count = sum(ifelse(upos == "NOUN", 1, 0)),
            NOUN_prop = NOUN_count/total,
            ADJ_count = sum(ifelse(upos == "ADJ", 1, 0)),
            ADJ_prop = ADJ_count/total,
            VERB_count = sum(ifelse(upos == "VERB", 1, 0)),
            VERB_prop = VERB_count/total,
            ADV_count = sum(ifelse(upos == "ADV", 1, 0)),
            ADV_prop = ADV_count/total,
            PROPN_count = sum(ifelse(upos == "PROPN", 1, 0)),
            PROPN_prop = PROPN_count/total,
            CONJ_count = sum(ifelse(upos == "CCONJ" | upos == "SCONJ", 1, 0)),
            CONJ_prop = CONJ_count/total,
            PRON_count = sum(ifelse(upos == "PRON", 1, 0)),
            PRON_prop = PRON_count/total,
            PUNCT_count = sum(ifelse(upos == "PUNCT", 1, 0)),
            PUNCT_prop = PUNCT_count/total,
            .groups = "drop")

head(token_type_summary)
```

```
# A tibble: 6 x 18
  chapter_number total NOUN_count NOUN_prop ADJ_count ADJ_prop VERB_count
           <int> <int>      <dbl>     <dbl>     <dbl>    <dbl>      <dbl>
1              1  2583        447     0.173       163   0.0631        270
2              2  1691        253     0.150       123   0.0727        172
3              3  7045       1083     0.154       465   0.0660        825
4              4  1921        295     0.154       121   0.0630        242
5              5   886        152     0.172        69   0.0779         84
6              6   981        197     0.201        71   0.0724         92
# i 11 more variables: VERB_prop <dbl>, ADV_count <dbl>, ADV_prop <dbl>,
#   PROPN_count <dbl>, PROPN_prop <dbl>, CONJ_count <dbl>, CONJ_prop <dbl>,
#   PRON_count <dbl>, PRON_prop <dbl>, PUNCT_count <dbl>, PUNCT_prop <dbl>
```

Then we can write this data to a CSV, which we will use in the Shiny App to explore our findings. Note that we save both the tokenized text with chapter numbers and the summary of totals and proportions per chapter. Both will be utilized in the UI of the Shiny App.

```
write_csv(token_with_chapters, "token_with_chapters.csv")
write_csv(token_type_summary, "token_type_summary.csv")
```

By adhering to this approach, we illuminate the distinct narrative and descriptive textures within *Moby-Dick,* thereby advancing both literary analysis and computational text examination.

Shiny App Exploration

Users can delve into the data relationships on their own through the Shiny App, enabling them to inspect various plots and statistics regarding the usage of different parts of speech across chapters. This feature facilitates comparisons between the proportions of specific parts of speech per chapter, allowing for an in-depth exploration of potential correlations with the text's content.

Upon accessing the Shiny App, users are greeted with options to select from parts of speech such as "Noun", "Adjective", "Verb", "Adverb", "Proper Noun", "Conjunction", "Pronoun", and "Punctuation." Following their selection, the app displays both a table and a plot.

The table lists chapter numbers alongside the total word count, the frequency of the chosen part of speech, and its chapter-wise proportion. By default, the plot orders these proportions in descending order, emphasizing the chapters where the selected part of speech is most prevalent.

We focus particularly on the proportions of verbs and proper nouns. Below is the R code segment that generates tables sorted by the descending proportion of these parts of speech:

```
# Generating and displaying tables for verbs and proper nouns
verb_table <- token_type_summary |>
      select(chapter_number,
             total,
             VERB_count,
             VERB_prop
             ) |>
      arrange(desc(VERB_prop))
propn_table <- token_type_summary |>
      select(chapter_number,
             total,
             PROPN_count,
             PROPN_prop
             ) |>
      arrange(desc(PROPN_prop))

head(verb_table)
```

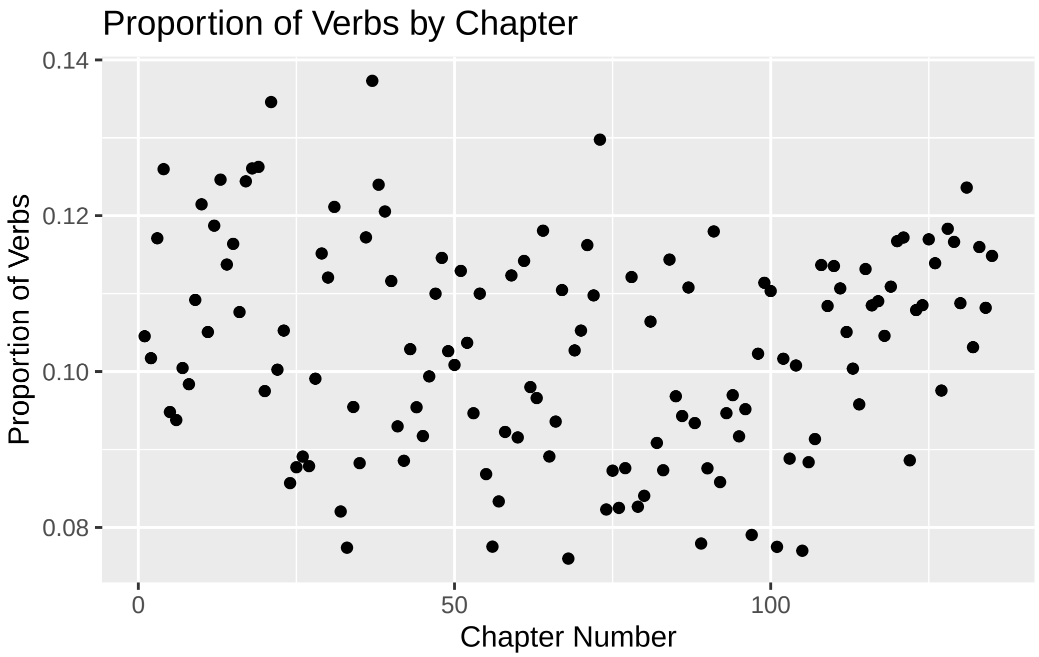
```
# A tibble: 6 x 4
  chapter_number total VERB_count VERB_prop
           <int> <int>      <dbl>     <dbl>
1             37   670         92     0.137
2             21  1434        193     0.135
3             73  2674        347     0.130
4             19  1584        200     0.126
5             18  1745        220     0.126
6              4  1921        242     0.126
```

```
head(propn_table)
```

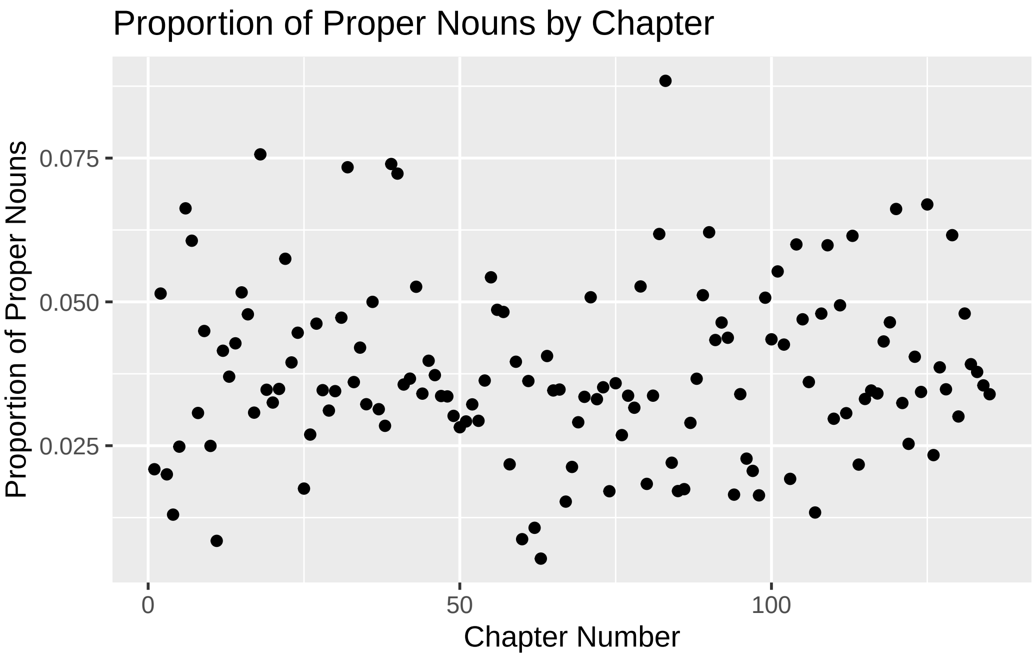
```
# A tibble: 6 x 4
  chapter_number total PROPN_count PROPN_prop
           <int> <int>       <dbl>      <dbl>
1             83   916          81     0.0884
2             18  1745         132     0.0756
3             39   365          27     0.0740
4             32  6253         459     0.0734
5             40  2213         160     0.0723
6            125  1479          99     0.0669
```

Below the table lies a plot generated by the `ggplot` package. The x- and y-axes of the plot are the chapter number and proportion of selected part of speech. We can display the plots for verbs and proper nouns. Within the app, the plots are displayed by the `plotly` package. The reason for inconsistency lies in our ability to render `plotly` outputs in a Quarto file as a PDF.

```
# Generating ggplot visuals for verbs and proper nouns
ggplot(data = token_type_summary, aes(x = chapter_number, y = VERB_prop)) +
  geom_point() +  # Use geom_line() if you prefer lines
  labs(title = 'Proportion of Verbs by Chapter',
       x = 'Chapter Number',
       y = 'Proportion of Verbs')
```



```r
ggplot(data = token_type_summary, aes(x = chapter_number, y = PROPN_prop)) +
  geom_point() +  # Use geom_line() if you prefer lines
  labs(title = 'Proportion of Proper Nouns by Chapter',
       x = 'Chapter Number',
       y = 'Proportion of Proper Nouns')
```



Within the app, the plot points are interactive: hovering over a point reveals the chapter number and its corresponding proportion of the selected part of speech. Clicking on a point displays the text from the respective chapter, highlighting occurrences of the chosen part of speech in bold.

The text is generated from the tokenized text with associated chapter numbers. Our goal is to produce this text in paragraph form. First, we need to annotate the text. We do this with the `mutate()` function by rewriting the `token` variable as either an annotated version of that token or the token itself, depending on whether the part of speech matches our desired part of speech. We annotate using the HTML technique of padding the front and back with `<b>` and `<\b>`, respectively. Then we would like to paste the text together. Our initial thought is to add a space after every word and paste them together. This, however, ignores the case of punctuations, which immediately follow words. Thus, we must create two new variables. First, we create a variable for a space after each word called `space_after`. If the next word is a punctuation, we do not add a space. Then we paste the word and the `space_after` to get the `word_with_space`. Then we paste the words together to get a sentence, and then paste the sentences together to get paragraphs.

We particularly examine chapters with notably high or low occurrences of the selected parts of speech to discern more about their relationship with the narrative style. For instance, chapters 37 and 21, which are rich in verbs, offer insights into character interactions aboard the Pequod.

```r
annotated_37 <- token_with_chapters |>
        # Get chapter by number
        filter(chapter_number == 37) |>
        # Add bold to specified part of speech
        mutate(token =
                 if_else(upos == "VERB",
                         paste0("<b>", token, "</b>"), token)) |>
        arrange(doc_id, sid, tid) |>
        group_by(doc_id, sid) |>
        # Add spaces after words, but not when next token is punctuation
        mutate(space_after = ifelse(lead(upos, default = " ") == "PUNCT",
                                    "", " "),
               word_with_space = paste(token, space_after, sep = "")) |>
        # Combine words to sentence
        summarise(sentence = paste(word_with_space, collapse = ""),
                  .groups = 'drop') |>
        group_by(doc_id) |>
        # Combine sentences to paragraphs
        summarise(paragraph = paste(sentence, collapse = " "),
                  .groups = 'drop')
HTML(annotated_37$paragraph[1:20])
```

```r
annotated_21 <- token_with_chapters |>
        # Get chapter by number
        filter(chapter_number == 21) |>
        # Add bold to specified part of speech
        mutate(token =
                 if_else(upos == "VERB",
                         paste0("<b>", token, "</b>"), token)) |>
        arrange(doc_id, sid, tid) |>
        group_by(doc_id, sid) |>
        # Add spaces after words, but not when next token is punctuation
        mutate(space_after = ifelse(lead(upos, default = " ") == "PUNCT",
                                    "", " "),
               word_with_space = paste(token, space_after, sep = "")) |>
        # Combine words to sentence
        summarise(sentence = paste(word_with_space, collapse = ""),
                  .groups = 'drop') |>
        group_by(doc_id) |>
        # Combine sentences to paragraphs
        summarise(paragraph = paste(sentence, collapse = " "),
                    .groups = 'drop')

HTML(annotated_21$paragraph[1:20])
```

In these chapters, verbs are displayed in **bold** text. As we can see, these chapters are about the crew of the Pequod and their character interactions. Now let's explore those chapters with the fewest verbs: chapters 68 and 105.

In our analysis, when delving into chapters like 68 and 105, which showcase a lesser frequency of verbs, we notice a focused exposition on the whale's anatomy and its various aspects. This observation reinforces our hypothesis: chapters with fewer verbs lean more towards a descriptive narrative, concentrating on detailed descriptions and informative content. This style is reflective of expository writing, marked by an emphasis on nouns, adjectives, and non-action verbs, crafted to provide a comprehensive picture to the reader.

Conversely, chapters teeming with verbs often pulsate with narrative energy, moving the story forward through vivid scenes of action and dialogue among characters. This variation in verb usage enlightens us to Melville's distinct literary styles (narrative vs. descriptive). High verb counts signal narrative character interaction and propel the plot, while a scarcity of verbs typically signifies a shift towards descriptive exposition, offering educational or background information.

Moreover, the usage of parts of speech as a literary device in *Moby-Dick* is not just a matter of stylistic preference but a deliberate technique to immerse readers into the dual facets of Melville's world: the tangible, action-driven life aboard the Pequod, and the expansive, contemplative realm of whaling knowledge.

In this context, the Shiny App serves as a powerful tool, empowering users to explore *Moby-Dick* with an analytical lens. It enables users to seamlessly navigate through the novel, fostering a deeper comprehension of the complex interplay between parts of speech and Melville's literary styles. By interacting with the app, users can visualize and understand how Melville weaves narrative drive with descriptive depth, thus enriching their appreciation of the novel's multifaceted composition.

Limitations and Moving Further

While this analysis uncovers some intriguing patterns within *Moby-Dick,* it's important to acknowledge certain limitations of the study. Firstly, the data preparation process is conducted externally, requiring some time to complete. Consequently, reproducing the results necessitates access to the `anno_moby.Rds` file, which might not be readily available. Secondly, the tokenized summaries utilized in this analysis are not included in our standard data package; instead, they are saved as CSV files and subsequently imported into the Shiny App. Should I extend this project, integrating this data directly into the `MobyDick` package would be a priority to streamline access.

Additionally, there's an aesthetic issue within the Shiny App where chapter texts are displayed as uninterrupted blocks of text rather than respecting the original paragraph breaks found in *Moby-Dick.* Correcting this to more accurately reflect the text's intended format—with appropriate new line breaks for paragraphs—would enhance readability and fidelity to the original work.

References

Hillway, Tyrus, and Luther S. Mansfield. 1953. *Moby-Dick: Centennial Essays*. Southern Methodist University Press.

Melville, Herman. 1851. *Moby-Dick; or, the Whale*. Vol. 1. Harper & Brothers, Publishers, New York.

ProjectGutenberg. 2001. "Moby Dick; or, the Whale by Herman Melville." https://www.gutenberg.org/ebooks/2701.