HarryPotter

C Dziewicki

12/21/2020

Introduction

This project was completed as an independent exploration of some R tools that I have not used yet. I have used code examples and got insipiration from the authors listed in the appendix.

```
#lists for books
titles <- c("philosophers_stone", "chamber_of_secrets", "prisoner_of_azkaban", "goblet_of_fire", "order_of_the_phoenix", "half_blood_prince", "deathly_hallows")
books<- list(philosophers_stone, chamber_of_secrets, prisoner_of_azkaban, goblet_of_fire, order_of_the_phoenix, half_blood_prince, deathly_hallows)

series <- tibble()
for(i in seq_along(titles)) {

  temp <- tibble(chapter = seq_along(books[[i]]),
                 text = books[[i]]) %>%
    unnest_tokens(word, text) %>%
    #tokenize each chapter into words
    mutate(book = titles[i]) %>%
    select(book, everything())

  series <- rbind(series, temp)
}

series$book <-factor(series$book, levels = rev(titles))
series
```

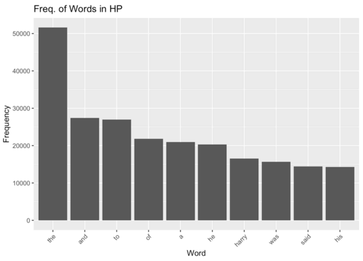
```
## # A tibble: 1,089,386 x 3
##    book              chapter word
##    <fct>               <int> <chr>
##  1 philosophers_stone      1 the
##  2 philosophers_stone      1 boy
##  3 philosophers_stone      1 who
##  4 philosophers_stone      1 lived
##  5 philosophers_stone      1 mr
##  6 philosophers_stone      1 and
##  7 philosophers_stone      1 mrs
##  8 philosophers_stone      1 dursley
##  9 philosophers_stone      1 of
## 10 philosophers_stone      1 number
## # … with 1,089,376 more rows
```

```
wordCount<- series %>% count(word, sort= TRUE)
```

Here we can see the frequency of all the words in the books. However words such as "the" are not too exciting. Next I will remove these filler words.

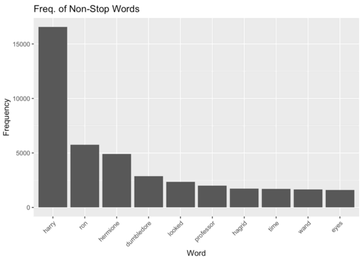
```
wordCount %>% top_n(10) %>% ggplot(mapping = aes(reorder(word, -n),n)) + geom_col() + theme(axis.text.x=element_text(angle=45, hjust=1)) + labs(x = "Word", y = "Frequency", title = "Freq. of Words in HP")
```

```
## Selecting by n
```




```
series$book <- factor(series$book, levels = rev(titles))
#creating a count without filler words
nonStop<- series %>%
  anti_join(stop_words) %>%
  count(word) %>% top_n(10)

nonStop %>%
  with(ggplot(mapping = aes(reorder(word, -n),n)) + geom_col() + labs(x= "Word", y = "Frequency", title = "Freq. of Non-Stop Words") + theme(axis.text.x=element_text(angle=45, hjust=1)))
```

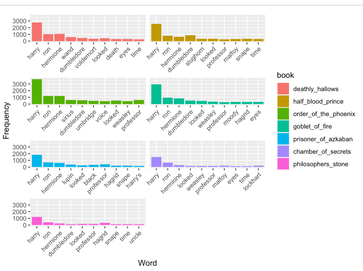


Now we can see that Harry and Ron are the most frequently used words. There is a phnomemna occuring in both this graph and the one preceeding. Zipf's Law states that given some corpus of natural language, the frequency of any word is inversly proportional to it's rank in the frequency table. The text of JK Rowling seems to be following this law. We can also see below that there is a similar relationship for each book.

```
series %>% group_by(book) %>% anti_join(stop_words) %>%
  count(word) %>% top_n(10) %>%
  ggplot(mapping = aes(reorder(word, -n),n, fill = book)) +
  geom_col() +
  labs(x= "Word", y = "Frequency") +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  facet_wrap(~book, ncol=2, nrow = 4, scales = "free_x") +
  theme(strip.background = element_blank(),
  strip.text.x = element_blank()
)
```

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

```
## Selecting by n
```



As you can see the words that are most frequent do vary across books, due to the story line and characters.

Next I will explore the sentiment of the words using the **SentimentAnalysis** library with the function *get_sentiments*.

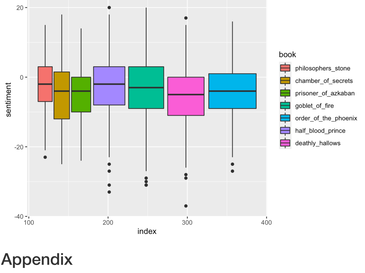
```
series %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE)
```

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

```
## # A tibble: 3,278 x 3
##    word      sentiment     n
##    <chr>     <chr>     <int>
##  1 dark      negative   1034
##  2 death     negative    757
##  3 magic     positive    606
##  4 hard      negative    466
##  5 fell      negative    460
##  6 top       positive    434
##  7 fudge     negative    433
##  8 moody     negative    422
##  9 magical   positive    380
## 10 dead      negative    378
## # … with 3,268 more rows
```

Here we can see the average sentiment of the words from each book. Positive words were given a 1 and negitive words a -1. This is interesting because we can see that the 6th book is the most overall positive book. I predict this is due to many light hearted moments throughout, even though there are many dark scenes.

```
series %>%
  group_by(book) %>%
  mutate(word_count = 1:n(),
         index = word_count %/% 500 + 1) %>%
  inner_join(get_sentiments("bing")) %>%
  count(book, index = index , sentiment) %>%
  ungroup() %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative,
         book = factor(book, levels = titles)) %>%
  ggplot(aes(index, sentiment, fill = book)) +
  geom_boxplot()
```



Appendix

https://en.wikipedia.org/wiki/Zipf%27s_law

https://rstudio-pubs-static.s3.amazonaws.com/300624_8260952d1f0346969e65f41a97006bf5.html

https://www.rpubs.com/dvdunne/reorder_ggplot_barchart_axis

https://moderndive.com/2-viz.html#geombar

https://github.com/bradleyboehmke/harrypotter

http://rstudio-pubs-static.s3.amazonaws.com/449570_fbd322569a664a139f38476a835492c1.html