Project 2

Yu Lu

2023-10-05

Question 1

Choose an author with over 5 distinct works (group_by gutenberg_works). All works should have text. Choose either all works or 5 random works. Only use public domain data.

The number of distinct works of Virginia Wolf is 5.

Question 2

Download the author's works using gutenberg_download and save the data as an RDS file, including title. Make sure strip = TRUE. The data should not be included in the repository (see usethis::use_git_ignore). Have if statements that check if the data exists, downloads the data if not available (e.g. if we clone the repo), and reads the data from the saved RDS if it is available.

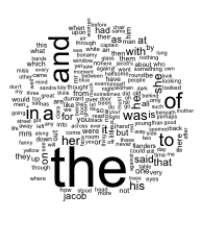
The downloaded data of Virginia Woolf's work is stored in woolf.rds

Question 3 Sentimental Analysis

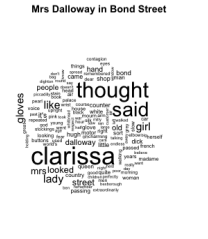
Produce a sentiment analysis by book by percentage completed in the book similar to the lecture. Create a plot showing cumulative sentiment over time.

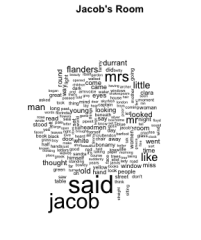
Remove or adjust the data depending on words that may have been mischaracterized (similar to the "miss" issue in Austen's works)

First, plot a wordcloud for each book of Virginia Woolf.



From the above plots, we noticed the top words are all stop words. Now we filter out all the stop words and plot out the word clouds again.





Now we conduct sentiment analysis.

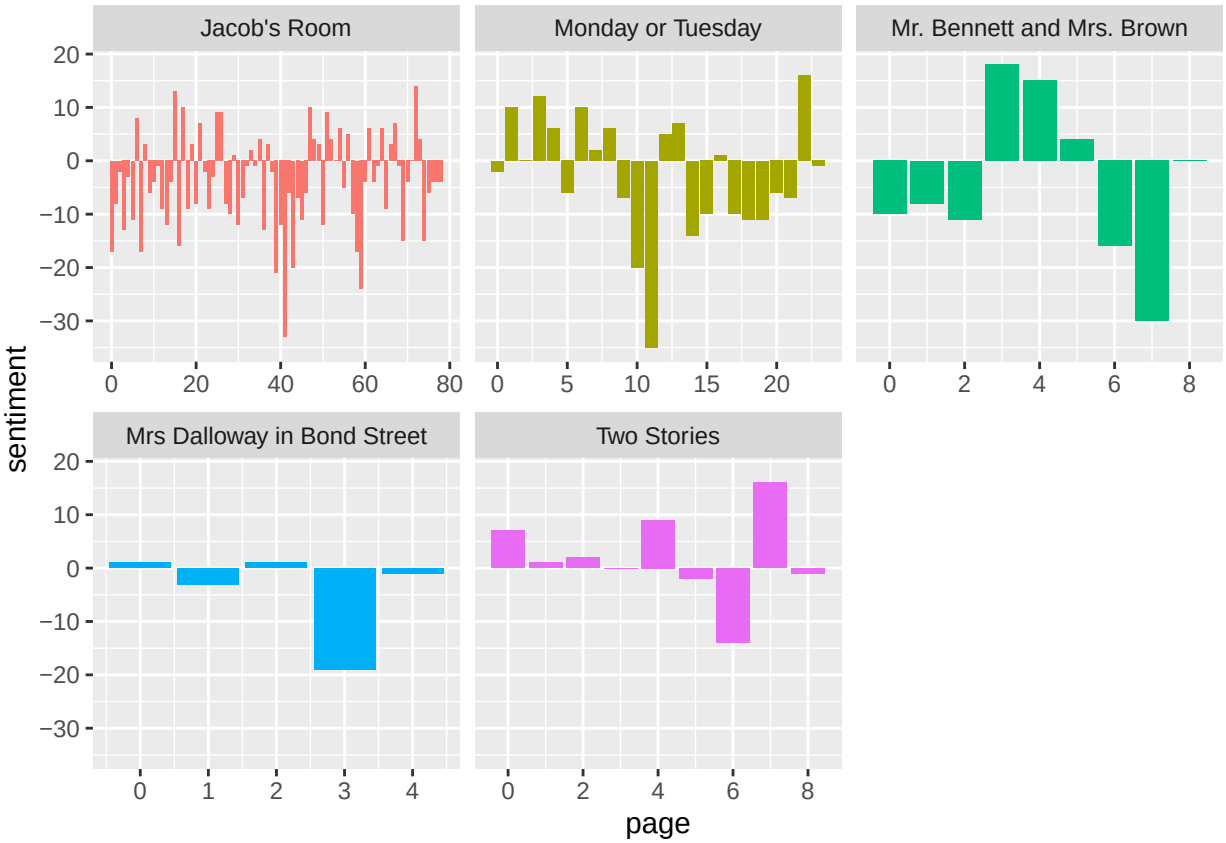
First we chech if there are words that may have been mischaracterized.

```
## # A tibble: 10 x 4
```

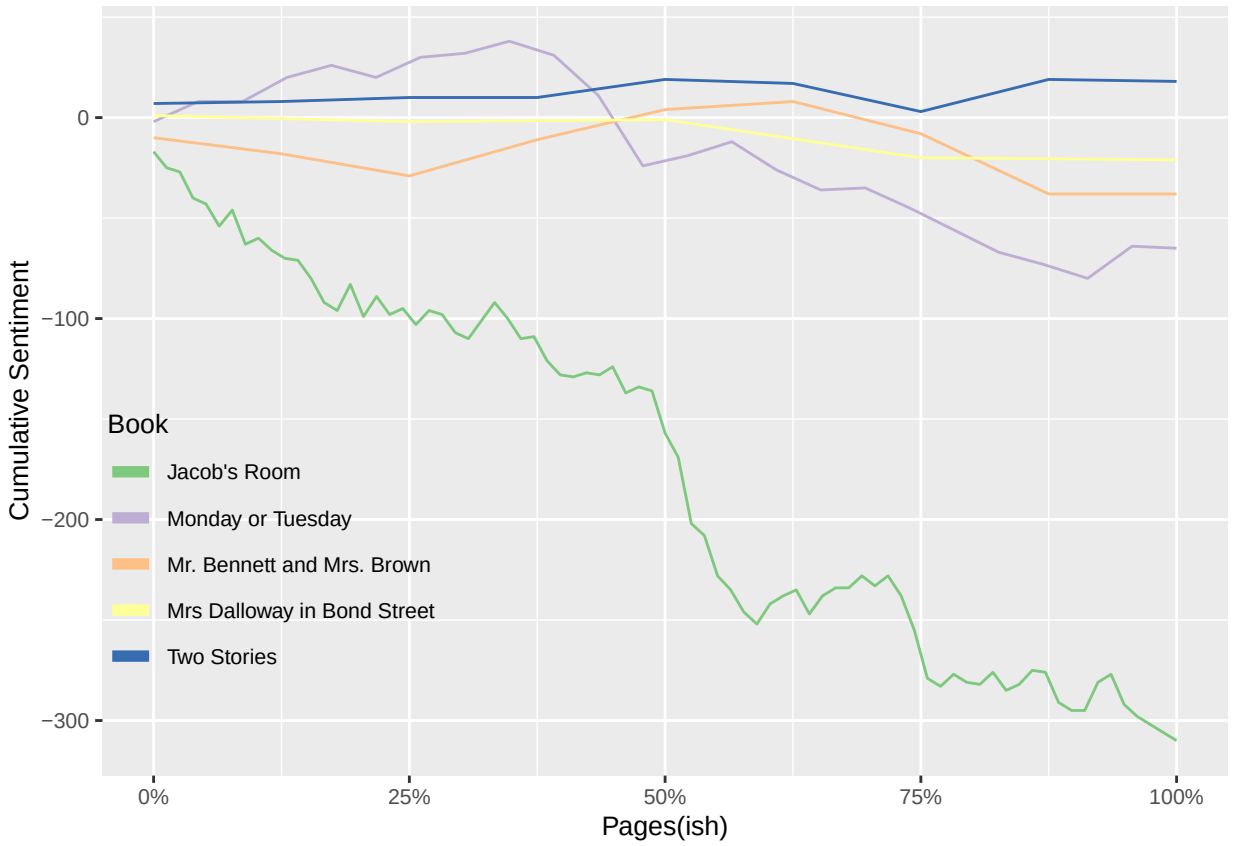
```
## # Groups:   title [5]
##    word  title                          n sentiment
##    <chr> <chr>                      <int> <chr>
##  1 like  Jacob's Room                 166 positive
##  2 miss  Jacob's Room                  73 negative
##  3 like  Monday or Tuesday             49 positive
##  4 good  Monday or Tuesday             26 positive
##  5 great Mr. Bennett and Mrs. Brown    13 positive
##  6 like  Mr. Bennett and Mrs. Brown    10 positive
##  7 like  Mrs Dalloway in Bond Street   16 positive
##  8 dick  Mrs Dalloway in Bond Street    7 negative
##  9 like  Two Stories                   24 positive
## 10 poor  Two Stories                   11 negative
```

There is still "miss" being classified as negative. I would also see the word "like" as misclassified as positive. With that in mind, we will do the sentiment analysis.

First, we have the sentiment trajectory:



Now we plot the cumulative sentiment with normalized book length



The first dicovery we can tell from the above plots is that most of Woolf's novels are pretty short expect Jacob's Room. This agree with the nature of Woolf's writings. As a nature of the stream of consciousness, there is no strong sign of positive or negative in Woolf's works. But the cumulative sentiments of Jacob's Room decreases overtime showed the underlying sadness from the void and emptiness writing itself. Also for Mrs. Dalloway, where the main character killed herself in the end, the cumulative sentiments also decreased, but very smoothly. Partly because this is a short book but may also because of the ambiguous writing style, with the slight mist of sadness.

Question 4

Randomly select 2 other authors with over 5 distinct works, download 5 random works from each. Combine these data with the original author and perform a topic model analysis with 3 topics and see how it breaks by author.

Discuss what topics load moist highly to each topic.

Repeat the analysis with more than 3 topics. Discuss how you chose the number of topics.

Create plots showing the topics by book and words by topic.

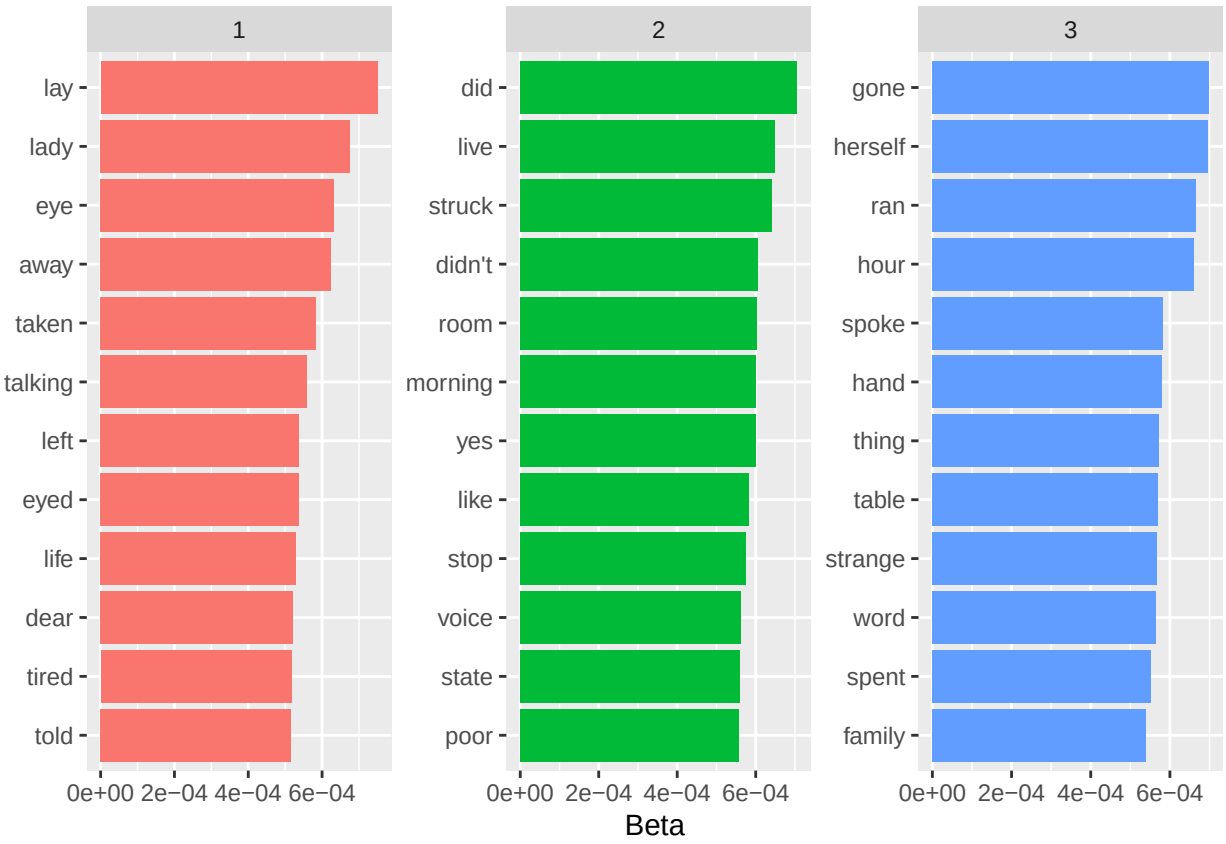
Here I first want to compare the topics of the stream of consciousness writers but found this dataset does not have enough books from them. So I turned to compare Virginia Woolf's work with other two famous female writers. Frances Hodgson Burnett and Louisa May Alcott.

The most used words in these three writers works.

word	title	n	sentiment
jacob	Jacob's Room	356	NA
little	Jacob's Room	137	NA
man	Jacob's Room	125	NA
it's	Monday or Tuesday	39	NA
oh	Monday or Tuesday	35	NA
man	Monday or Tuesday	33	NA
brown	Mr. Bennett and Mrs. Brown	39	NA
character	Mr. Bennett and Mrs. Brown	33	NA
bennett	Mr. Bennett and Mrs. Brown	27	NA
clarissa	Mrs Dalloway in Bond Street	42	NA
thought	Mrs Dalloway in Bond Street	36	NA
girl	Mrs Dalloway in Bond Street	16	NA
don't	Two Stories	28	NA
looked	Two Stories	18	NA
believe	Two Stories	17	NA

word	title	n	sentiment
polly	An Old-Fashioned Girl	1182	NA
tom	An Old-Fashioned Girl	614	NA
n't	An Old-Fashioned Girl	586	NA
little	Aunt Jo's Scrap-Bag, Volume 3 Cupid and Chow-chow, etc.	261	NA
old	Aunt Jo's Scrap-Bag, Volume 3 Cupid and Chow-chow, etc.	130	NA
chow	Aunt Jo's Scrap-Bag, Volume 3 Cupid and Chow-chow, etc.	114	NA
rose	Eight Cousins; Or, The Aunt-Hill	660	NA
little	Eight Cousins; Or, The Aunt-Hill	319	NA
aunt	Eight Cousins; Or, The Aunt-Hill	223	NA
little	Kitty's Class Day and Other Stories	269	NA
old	Kitty's Class Day and Other Stories	188	NA
amy	Kitty's Class Day and Other Stories	169	NA
jasper	The Abbot's Ghost, or Maurice Treherne's Temptation: A Christmas Story	101	NA
treherne	The Abbot's Ghost, or Maurice Treherne's Temptation: A Christmas Story	93	NA
snowdon	The Abbot's Ghost, or Maurice Treherne's Temptation: A Christmas Story	73	NA

word	title	n	sentiment
m	"Le Monsieur de la Petite Dame"	65	NA
villefort	"Le Monsieur de la Petite Dame"	60	NA
bertha	"Le Monsieur de la Petite Dame"	56	NA
editha	Editha's Burglar: A Story for Children	59	NA
burglar	Editha's Burglar: A Story for Children	47	NA
little	Editha's Burglar: A Story for Children	46	NA
clelie	Esmeralda	46	NA
mother	Esmeralda	36	NA
esmeraldy	Esmeralda	32	NA
ye	Lodusky	55	NA
n	Lodusky	49	NA
rebecca	Lodusky	41	NA
yer	The Dawn of a To-morrow	101	NA
dart	The Dawn of a To-morrow	78	NA
thing	The Dawn of a To-morrow	66	NA



The three topics showed somewhat characters of these three ladies writing and their usage of femine words. But there is no clear distinctions