Solution slides Part 2
Introduction to R & Data for Humanities

Afternoon session

Text-mining with Tidyverse

Exercise 9

9a.

```r
#Exercise 9a. Can you call on the tidytext package and then use the function mentioned above to get tibbles of the three lexicons mentioned on the slide? Press 'enter' twice and start coding!

library(tidytext)

get_sentiments("afinn")
get_sentiments("bing")
get_sentiments("nrc")
```

spec_tbl_df 2477 x 2 | tbl_df 6786 x 2 | tbl_df 13901 x 2

word <chr>	sentiment <chr>
abacus	trust
abandon	fear
abandon	negative
abandon	sadness
abandoned	anger
abandoned	fear
abandoned	negative
abandoned	sadness
abandonment	anger
abandonment	fear

1-10 of 13,901 rows

Previous 1 2 3 4 5 6 ... 100 Next

See next slide for a characteri-zation of how these lexicons score sentiment

9a. (resumed)

All three of these lexicons are based on unigrams, i.e., single words. These lexicons contain many English words and the words are assigned scores for positive/negative sentiment, and also possibly emotions like joy, anger, sadness, and so forth. The nrc lexicon categorizes words in a binary fashion (“yes”/“no”) into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The bing lexicon categorizes words in a binary fashion into positive and negative categories. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

9b.

```
153
154 # Sentiment analysis of Emma I
155 ```{r}
156
157 # Exercise 9b.Let's ask ourselves: What are the most common joy words in Austen's novel Emma? Run this code in order to make
    your data tidy first and do some real code work in the next exercise!
158
159 library(janeaustenr)
160 library(dplyr)
161 library(stringr)
162
163 tidy_books <- austen_books() %>%
164   group_by(book) %>%
165   mutate(
166     linenumber = row_number(),
167     chapter = cumsum(str_detect(text,
168                                 regex("^chapter [\\divxlc]",
169                                       ignore_case = TRUE)))) %>%
170   ungroup() %>%
171   unnest_tokens(word, text)
172 ```
```

170:16 Chunk 12 R Markdown

Console Terminal Jobs

C:/WINDOWS/system32/

```
> library(janeaustenr)
> library(dplyr)
> library(stringr)
>
> tidy_books <- austen_books() %>%
+   group_by(book) %>%
+   mutate(
+     linenumber = row_number(),
+     chapter = cumsum(str_detect(text,
+                                 regex("^chapter [\\divxlc]",
+                                       ignore_case = TRUE)))) %>%
+   ungroup() %>%
+   unnest_tokens(word, text)
>
```

We only run this code to make sure that our data is tidy; there is no visible output you need to take into account.

9c.

```r
# Exercise 9c.We want to know what the most common joy words in Emma are. Can you complete the code and run the script based on the pointers on the slide?

nrc_joy <- get_sentiments("nrc") %>%
  filter(sentiment == "joy")

tidy_books %>%
  filter(book == "Emma") %>%
  inner_join(nrc_joy) %>%
  count(word, sort = TRUE)
```

R Console

tbl_df
303 x 2

word <chr>	n <int>
good	359
young	192
friend	166
hope	143
happy	125
love	117
deal	92
found	92
present	89
kind	82

1-10 of 303 rows

Previous 1 2 3 4 5 6 ... 31 Next

9d.

```r
# Exercise 9d. We can also examine how sentiment changes throughout each of Austen's novels. We can do this with just a handful
of lines that are mostly dplyr functions. Can you complete the code and run the script based on the pointers on the slide?


library(tidyr)

jane_austen_sentiment <- tidy_books %>%
  inner_join(get_sentiments("bing")) %>%
  count(book, index = linenumber %/% 80, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
  mutate(sentiment = positive - negative)
```

```
package 'tidyr' was built under R version 4.0.5Joining, by = "word"
```

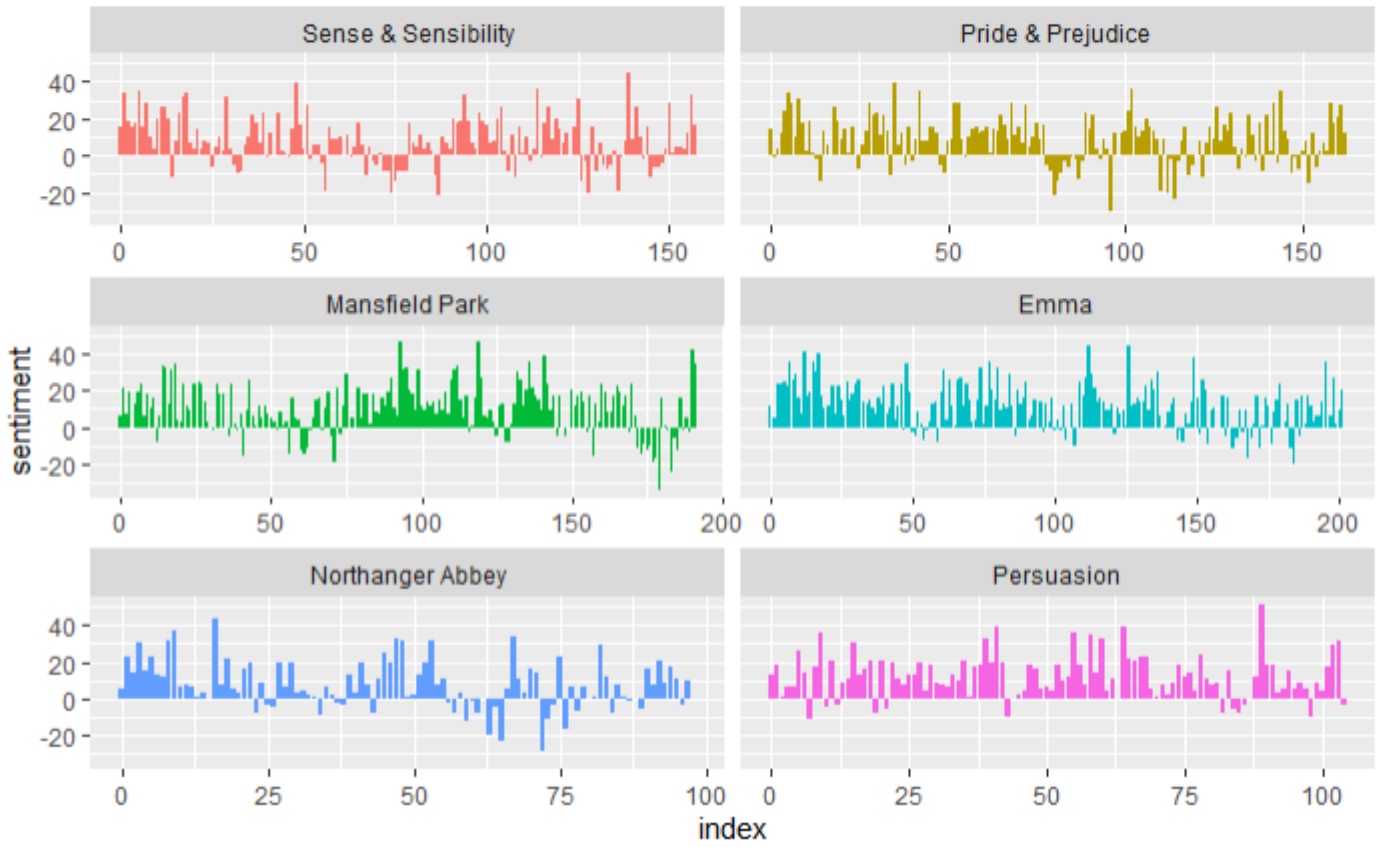
We run this code as a precursor to visualizing how sentiment changes throughout each of Austen's novels, so there is no visible output you need to take into account right now.

If you see the warning in red, you can safely ignore it.

9e.

```r
#Exercise 9e. Call on the ggplot2 package (do you remember how to call on a package? You have done so numerous times before!)
and then run the following code:


library(ggplot2)
ggplot(jane_austen_sentiment, aes(index, sentiment, fill = book)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~book, ncol = 2, scales = "free_x")
```



9e. (resumed)

Based on these graphs, we can begin to explore trends or differences in the novels' sentiment structures. For example, how the plot of each novel changes toward more positive or negative sentiment over the trajectory of the story. Based on your observations of the visualization you might want to start close reading certain passages of the novels, in order to analyze the specific language used in specific sections. You could also use these graphs as a starting point to look into how Austen's writing changes over time when it comes to the sentiment character of her novels.