sentiment_analysis

2023-01-20

```
defaultW <- getOption("warn")
options(warn = -1)
library(textdata)
library(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(tidyr)
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
# package giving the textual data in the form of books by Jane Austen
library(janeaustenr)
# helping to work with string
library(stringr)
options(warn=defaultW)
```

Including Plots

austen_books has text and book column

```
formated_data <- austen_books() %>%
  group_by(book)%>%
  mutate(linenumber= row_number())%>%
  ungroup()%>%unnest_tokens(word,text)
#adding a column named 'linenumber' to keep track of which line
# unnest_tokens() convert the text of the novels to tidy format
#we used 'word' for the output as sentiment lexicons have same name column
# hence performing inner-join will be easy
```

```
# Use NRC lexicon for sentiment analysis
nrc_analysis<- get_sentiments(lexicon = "nrc")
# we see some of the sentiments of NRC: trust, fear, negative, sadness, surprise, positive
```

```
# Let us take a look at the surprise sentiment from the book Northanger Abbey
nrc_surprise <-nrc_analysis%>% filter(sentiment== 'surprise')
north_abbey<-formated_data%>% filter(book=="Northanger Abbey")%>%
            inner_join(nrc_surprise)%>%
            count(word,sort = TRUE)
```

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

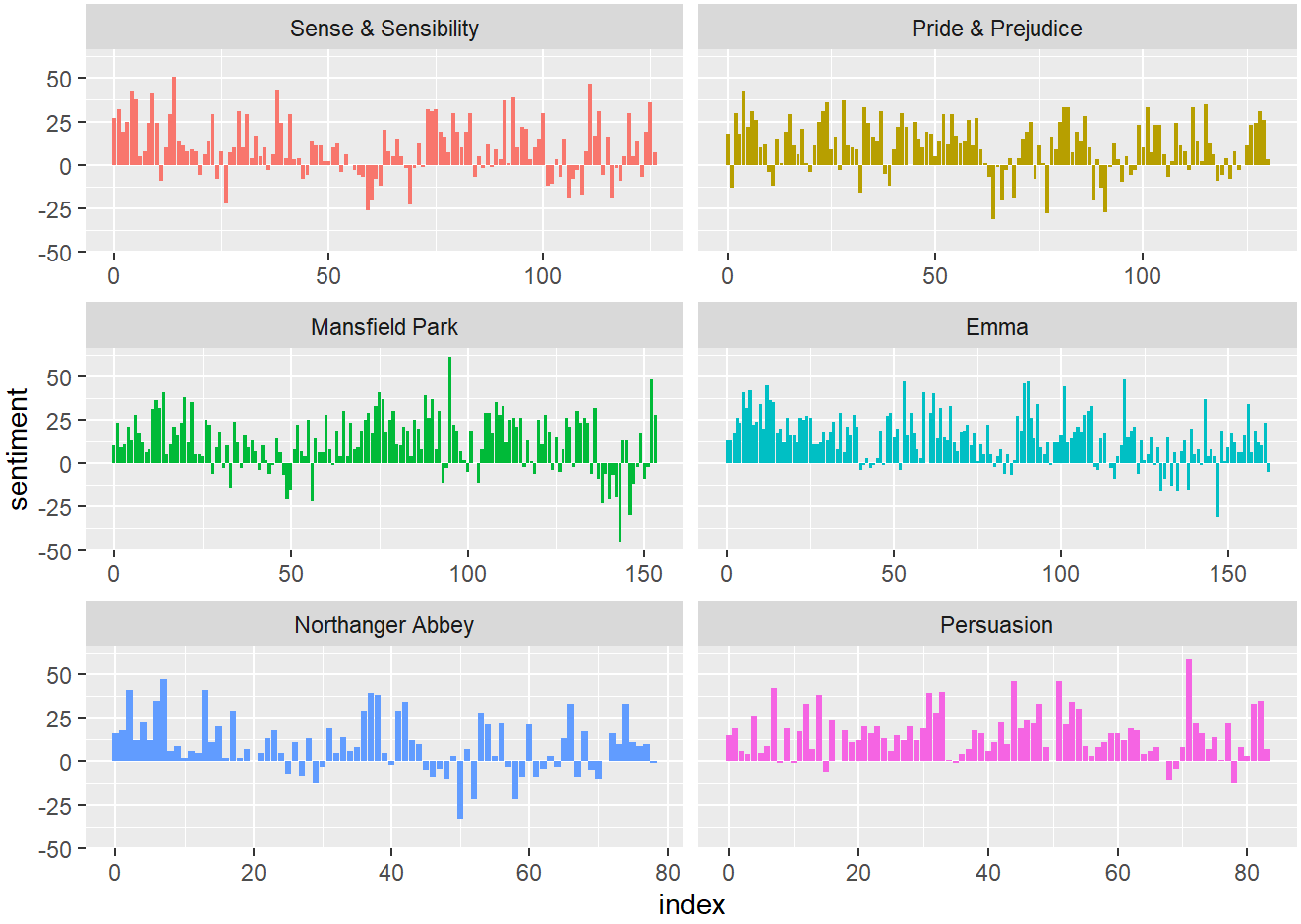
```
# here we are using count() to find the most common surprise word in Northanger Abbey
# by setting sort = TRUE we get highest count to lowest

## here we see mostly good, hope, spirits (positive words) in the list.
```

```
# further analysis: how sentiment changes throughout each of the 6 novels.
jane_sentiment <- formated_data %>%
  inner_join(get_sentiments(lexicon= "bing")) %>%
  count(book,index=linenumber%/% 100, sentiment)%>%
  pivot_wider(names_from = sentiment, values_from = n)%>%
  mutate(sentiment=positive - negative)
```

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

```
# here we inner_join() with 'bing' lexicon, which gives us positive/negative sentiment for words.
# next we count how many positive/negative words are there
# using a large section for analysis to have enough words to get a good estimate
# hence using 100 lines in each section
#using pivot_wider() to have separate columns of positive and negative
#finally added a column of net sentiment (positive - negative)
```

```
#plot og sentiment analysis of 6 novels
library(ggplot2)
ggplot(jane_sentiment, aes(index, sentiment, fill=book))+
  geom_col(show.legend = F)+
  facet_wrap(~book, ncol=2, scale='free_x')
```




```
# from the plot we can see mostly all of her novels have positive sentiment throughout the story.

#finally let us create a wordcloud with the most occuring positive and negative words in the 6 novels

formated_data %>% inner_join(get_sentiments(lexicon="bing"))%>%
  count(word,sentiment,sort=TRUE)%>%
  acast(word~sentiment, value.var="n",fill=0)%>%
  comparison.cloud(colors = c("red","dark blue"), max.words = 100)
```

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

negative

positive