Stat.653 Homework 2

Jiayi Liu

April 8, 2019

Read: Chapter 2 Problems:

# Run the R code from Chapter2

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

sentiments

## # A tibble: 27,314 x 4  
## word sentiment lexicon score  
## <chr> <chr> <chr> <int>  
## 1 abacus trust nrc NA  
## 2 abandon fear nrc NA  
## 3 abandon negative nrc NA  
## 4 abandon sadness nrc NA  
## 5 abandoned anger nrc NA  
## 6 abandoned fear nrc NA  
## 7 abandoned negative nrc NA  
## 8 abandoned sadness nrc NA  
## 9 abandonment anger nrc NA  
## 10 abandonment fear nrc NA  
## # ... with 27,304 more rows

get\_sentiments("afinn")

## # A tibble: 2,476 x 2  
## word score  
## <chr> <int>  
## 1 abandon -2  
## 2 abandoned -2  
## 3 abandons -2  
## 4 abducted -2  
## 5 abduction -2  
## 6 abductions -2  
## 7 abhor -3  
## 8 abhorred -3  
## 9 abhorrent -3  
## 10 abhors -3  
## # ... with 2,466 more rows

get\_sentiments("bing")

## # A tibble: 6,788 x 2  
## word sentiment  
## <chr> <chr>   
## 1 2-faced negative   
## 2 2-faces negative   
## 3 a+ positive   
## 4 abnormal negative   
## 5 abolish negative   
## 6 abominable negative   
## 7 abominably negative   
## 8 abominate negative   
## 9 abomination negative   
## 10 abort negative   
## # ... with 6,778 more rows

get\_sentiments("nrc")

## # A tibble: 13,901 x 2  
## word sentiment  
## <chr> <chr>   
## 1 abacus trust   
## 2 abandon fear   
## 3 abandon negative   
## 4 abandon sadness   
## 5 abandoned anger   
## 6 abandoned fear   
## 7 abandoned negative   
## 8 abandoned sadness   
## 9 abandonment anger   
## 10 abandonment fear   
## # ... with 13,891 more rows

# 02-sentiment-analysis.Rmd

library(knitr)  
opts\_chunk$set(message = FALSE, warning = FALSE, cache = TRUE)  
options(width = 100, dplyr.width = 100)  
library(ggplot2)  
theme\_set(theme\_light())

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)

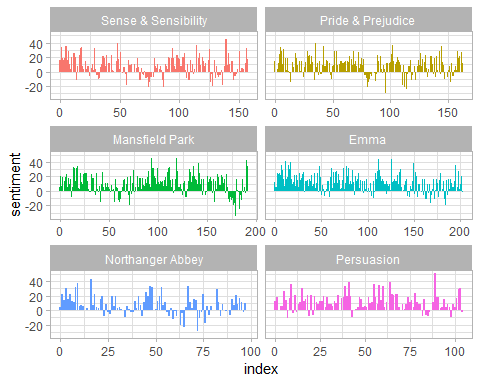
nrc\_joy <- get\_sentiments("nrc") %>%   
 filter(sentiment == "joy")  
  
tidy\_books %>%  
 filter(book == "Emma") %>%  
 inner\_join(nrc\_joy) %>%  
 count(word, sort = TRUE)

## # A tibble: 303 x 2  
## word n  
## <chr> <int>  
## 1 good 359  
## 2 young 192  
## 3 friend 166  
## 4 hope 143  
## 5 happy 125  
## 6 love 117  
## 7 deal 92  
## 8 found 92  
## 9 present 89  
## 10 kind 82  
## # ... with 293 more rows

library(tidyr)  
  
jane\_austen\_sentiment <- tidy\_books %>%  
 inner\_join(get\_sentiments("bing")) %>%  
 count(book, index = linenumber %/% 80, sentiment) %>%  
 spread(sentiment, n, fill = 0) %>%  
 mutate(sentiment = positive - negative)

library(tidyr)  
janeaustensentiment <- tidy\_books %>%   
 inner\_join(get\_sentiments("bing")) %>%   
 count(book, index = linenumber %/% 80, sentiment) %>%   
 spread(sentiment, n, fill = 0) %>%   
 mutate(sentiment = positive - negative)

library(ggplot2)  
ggplot(janeaustensentiment, aes(index, sentiment, fill = book)) +   
 geom\_col(show.legend = FALSE) +   
 facet\_wrap(~book, ncol = 2, scales = "free\_x")

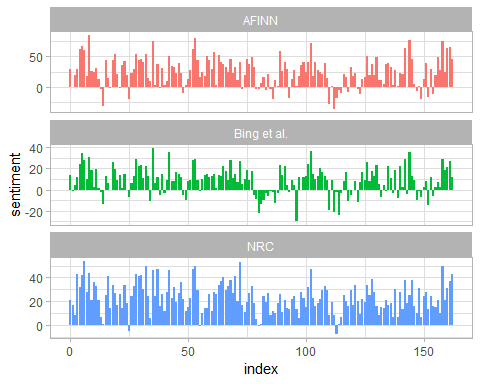


pride\_prejudice <- tidy\_books %>%   
 filter(book == "Pride & Prejudice")  
  
pride\_prejudice

## # A tibble: 122,204 x 4  
## book linenumber chapter word   
## <fct> <int> <int> <chr>   
## 1 Pride & Prejudice 1 0 pride   
## 2 Pride & Prejudice 1 0 and   
## 3 Pride & Prejudice 1 0 prejudice  
## 4 Pride & Prejudice 3 0 by   
## 5 Pride & Prejudice 3 0 jane   
## 6 Pride & Prejudice 3 0 austen   
## 7 Pride & Prejudice 7 1 chapter   
## 8 Pride & Prejudice 7 1 1   
## 9 Pride & Prejudice 10 1 it   
## 10 Pride & Prejudice 10 1 is   
## # ... with 122,194 more rows

afinn <- pride\_prejudice %>%   
 inner\_join(get\_sentiments("afinn")) %>%   
 group\_by(index = linenumber %/% 80) %>%   
 summarise(sentiment = sum(score)) %>%   
 mutate(method = "AFINN")  
  
bing\_and\_nrc <- bind\_rows(pride\_prejudice %>%   
 inner\_join(get\_sentiments("bing")) %>%  
 mutate(method = "Bing et al."),  
 pride\_prejudice %>%   
 inner\_join(get\_sentiments("nrc") %>%   
 filter(sentiment %in% c("positive",   
 "negative"))) %>%  
 mutate(method = "NRC")) %>%  
 count(method, index = linenumber %/% 80, sentiment) %>%  
 spread(sentiment, n, fill = 0) %>%  
 mutate(sentiment = positive - negative)

bind\_rows(afinn,   
 bing\_and\_nrc) %>%   
 ggplot(aes(index, sentiment, fill = method)) +   
 geom\_col(show.legend = FALSE) +   
 facet\_wrap(~method, ncol = 1, scales = "free\_y")



get\_sentiments("nrc") %>%   
 filter(sentiment %in% c("positive",   
 "negative")) %>%   
 count(sentiment)

## # A tibble: 2 x 2  
## sentiment n  
## <chr> <int>  
## 1 negative 3324  
## 2 positive 2312

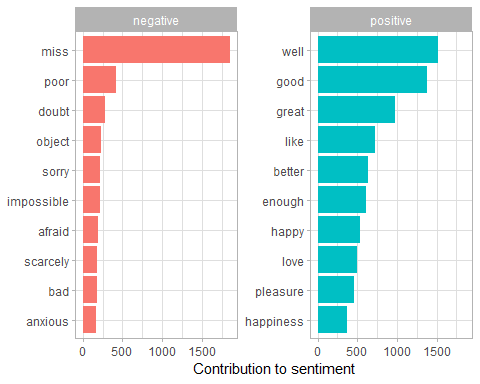
get\_sentiments("bing") %>%   
 count(sentiment)

## # A tibble: 2 x 2  
## sentiment n  
## <chr> <int>  
## 1 negative 4782  
## 2 positive 2006

bing\_word\_counts <- tidy\_books %>%  
 inner\_join(get\_sentiments("bing")) %>%  
 count(word, sentiment, sort = TRUE) %>%  
 ungroup()  
  
bing\_word\_counts

## # A tibble: 2,585 x 3  
## word sentiment n  
## <chr> <chr> <int>  
## 1 miss negative 1855  
## 2 well positive 1523  
## 3 good positive 1380  
## 4 great positive 981  
## 5 like positive 725  
## 6 better positive 639  
## 7 enough positive 613  
## 8 happy positive 534  
## 9 love positive 495  
## 10 pleasure positive 462  
## # ... with 2,575 more rows

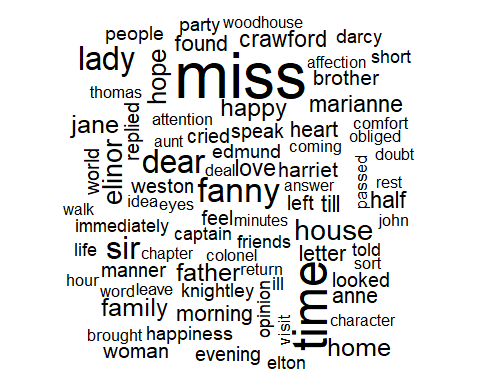
bing\_word\_counts %>%  
 group\_by(sentiment) %>%  
 top\_n(10) %>%  
 ungroup() %>%  
 mutate(word = reorder(word, n)) %>%  
 ggplot(aes(word, n, fill = sentiment)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~sentiment, scales = "free\_y") +  
 labs(y = "Contribution to sentiment",  
 x = NULL) +  
 coord\_flip()



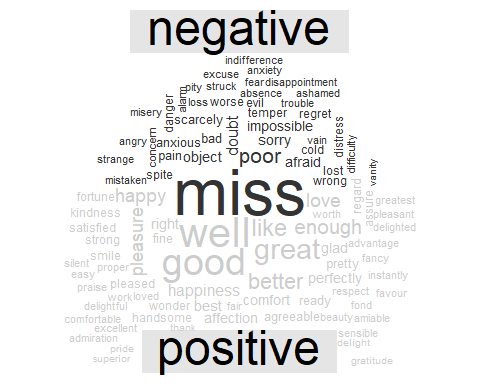
custom\_stop\_words <- bind\_rows(tibble(word = c("miss"),   
 lexicon = c("custom")),   
 stop\_words)  
  
custom\_stop\_words

## # A tibble: 1,150 x 2  
## word lexicon  
## <chr> <chr>   
## 1 miss custom   
## 2 a SMART   
## 3 a's SMART   
## 4 able SMART   
## 5 about SMART   
## 6 above SMART   
## 7 according SMART   
## 8 accordingly SMART   
## 9 across SMART   
## 10 actually SMART   
## # ... with 1,140 more rows

library(wordcloud)  
  
tidy\_books %>%  
 anti\_join(stop\_words) %>%  
 count(word) %>%  
 with(wordcloud(word, n, max.words = 100))



library(reshape2)  
  
tidy\_books %>%  
 inner\_join(get\_sentiments("bing")) %>%  
 count(word, sentiment, sort = TRUE) %>%  
 acast(word ~ sentiment, value.var = "n", fill = 0) %>%  
 comparison.cloud(colors = c("gray20", "gray80"),  
 max.words = 100)



PandP\_sentences <- tibble(text = prideprejudice) %>%   
 unnest\_tokens(sentence, text, token = "sentences")

PandP\_sentences$sentence[2]

## [1] "however little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters."

austen\_chapters <- austen\_books() %>%  
 group\_by(book) %>%  
 unnest\_tokens(chapter, text, token = "regex",   
 pattern = "Chapter|CHAPTER [\\dIVXLC]") %>%  
 ungroup()  
  
austen\_chapters %>%   
 group\_by(book) %>%   
 summarise(chapters = n())

## # A tibble: 6 x 2  
## book chapters  
## <fct> <int>  
## 1 Sense & Sensibility 51  
## 2 Pride & Prejudice 62  
## 3 Mansfield Park 49  
## 4 Emma 56  
## 5 Northanger Abbey 32  
## 6 Persuasion 25

bingnegative <- get\_sentiments("bing") %>%   
 filter(sentiment == "negative")  
  
wordcounts <- tidy\_books %>%  
 group\_by(book, chapter) %>%  
 summarize(words = n())  
  
tidy\_books %>%  
 semi\_join(bingnegative) %>%  
 group\_by(book, chapter) %>%  
 summarize(negativewords = n()) %>%  
 left\_join(wordcounts, by = c("book", "chapter")) %>%  
 mutate(ratio = negativewords/words) %>%  
 filter(chapter != 0) %>%  
 top\_n(1) %>%  
 ungroup()

## # A tibble: 6 x 5  
## book chapter negativewords words ratio  
## <fct> <int> <int> <int> <dbl>  
## 1 Sense & Sensibility 43 161 3405 0.0473  
## 2 Pride & Prejudice 34 111 2104 0.0528  
## 3 Mansfield Park 46 173 3685 0.0469  
## 4 Emma 15 151 3340 0.0452  
## 5 Northanger Abbey 21 149 2982 0.0500  
## 6 Persuasion 4 62 1807 0.0343

# Try out the sentimentr package on a sent of Tweets you find on Twitter. Just use the Twitter Search.

library(tidyverse)  
library(tidytext)  
library(janeaustenr)  
library(stringr)  
library(reshape2)  
  
library(sentimentr)

mytext <- c(  
 'In the early modern period',  
 'If you were keeping a book full of really useful stuff',  
 'youâd chuck everything in it.',  
 'It really is a the-more-the-merrier philosophy,',  
 'says \"A Discovery of Witches\" author'  
)

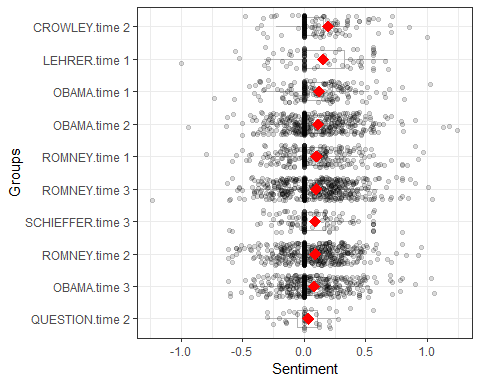
mytext <- get\_sentences(mytext)  
sentiment(mytext)

## element\_id sentence\_id word\_count sentiment  
## 1: 1 1 5 0.2683282  
## 2: 2 1 11 0.2713602  
## 3: 3 1 5 0.0000000  
## 4: 4 1 9 0.0000000  
## 5: 5 1 6 0.2449490

sentiment\_by(mytext) # for the whole twitters

## element\_id word\_count sd ave\_sentiment  
## 1: 1 5 NA 0.2683282  
## 2: 2 11 NA 0.2713602  
## 3: 3 5 NA 0.0000000  
## 4: 4 9 NA 0.0000000  
## 5: 5 6 NA 0.2449490

out <- with(  
 presidential\_debates\_2012,   
 sentiment\_by(  
 get\_sentences(dialogue),   
 list(person, time)  
 )  
)  
  
plot(out)



head(out)

## person time word\_count sd ave\_sentiment  
## 1: OBAMA time 1 3599 0.2535006 0.12256892  
## 2: OBAMA time 2 7477 0.2509177 0.11217673  
## 3: OBAMA time 3 7243 0.2441394 0.07975688  
## 4: ROMNEY time 1 4085 0.2525596 0.10151917  
## 5: ROMNEY time 2 7536 0.2205169 0.08791018  
## 6: ROMNEY time 3 8303 0.2623534 0.09968544