Romeo and Juliet Wordcloud

Jerrin Joe Varghese

November 7, 2017

Abstract

In this article we are constructing a wordcloud using the tidytext R package. Here we will be taking the words from the famous book Romeo and Juliets written by shakespear. We will extract all the words and convert it to a cloud of words in different shape and colurs using the package wordcloud.

Romeo and Juliet is a tragic love story written by William Shakespeare early in his career.[1]

1 The Gutenberg Package

This is a relatively new package for R,Gutenberg , that gives one access to all of the novels written by different authors.

```
library(janeaustenr)
library(dplyr)
library(tidytext)
library(gutenbergr)
library(wordcloud)
library(wordcloud2)
library(ggplot2)
library(stringr)
```

Let's find the book for Romeo and Juliets by shakespear using gutenberg.

```
gutenberg_works(str_detect(title,'Romeo'))
```

```
## # A tibble: 3 x 8
##   gutenberg_id
##          <int>
## 1         1513
## 2        22274
```

[1]The date of the publish is not available.

```
## 3        47960
## # ... with 7 more variables: title <chr>, author <chr>,
## #   gutenberg_author_id <int>, language <chr>, gutenberg_bookshelf <chr>,
## #   rights <chr>, has_text <lgl>
```

Let's now download and store it into a data frame.

```
Romeo<-gutenberg_download(1513)
```

2 Some Data Cleaning

Adding a new column of line to get the line numbers and clean the id to Null.

```
Romeo$line<-1:5268
Romeo$gutenberg_id<-NULL
```

3 The wordcloud

To make the word cloud, we first need to break up the lines into words. We can use the function from package tidytext for this.

```
Romeo_words<-Romeo%>%
  unnest_tokens(word,text)
```

We can remove common, unimportant words with the stop_words from the data frame with dplyr.

```
Romeo_words<-Romeo_words%>%
  filter(!(word %in% stop_words$word))
```

Since this is a tragic love story, so lets only take the joy and sadness word sentiments out of it.

```
nrc<-get_sentiments('nrc')

joy_sad<-nrc%>%
  filter(sentiment == 'joy' | sentiment == 'sadness')
```

Lets takeout all the joy and sadness words to a dataframe.

```
Romeo_joy_sad<-inner_join(Romeo_words,joy_sad)
```

Now we need to calculate frequencies of the words in the novel. Again, we can use standard techniques for this:

```
Romeo_frequency<-Romeo_joy_sad%>%
  group_by(word)%>%
  summarise(frequency = n())
```

Finally lets view our wordcloud.

```
wordcloud(Romeo_frequency$word,Romeo_frequency$frequency)
```

4 Sentiment Afinn

There is also another package afinn, which has 2 columns in its data frame, one with the sentiments and the other with the score for those words. Let's look into it.

```
affin<-get_sentiments('afinn')
```

Next, we can also divide the lines into groups or chuncks.That we can do by using "

```
Romeo_words$groups<-Romeo_words$line%/%80
```

We can use inner$_j oin to get our desired words from the sentiments, this is a way of cleaning.$

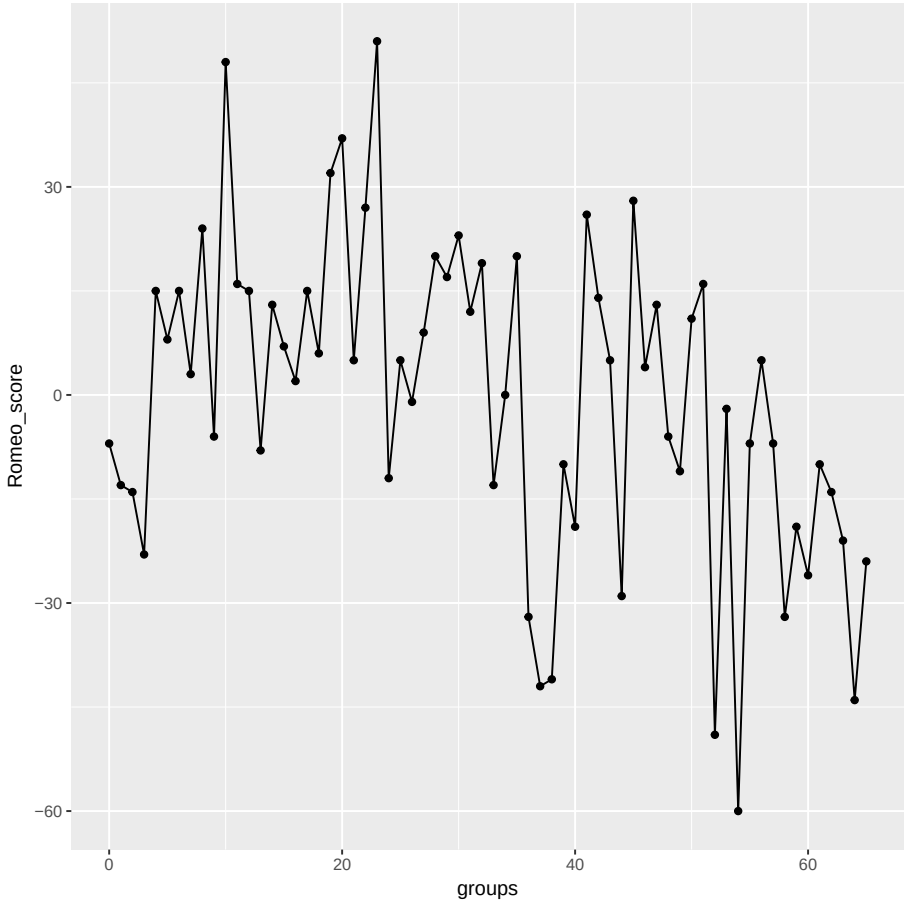
```
Romeo_words<-inner_join(Romeo_words,affin)
```

Now, lets score our words in groups of 80line that we have divided.

```
Romeo_senti_score<-Romeo_words%>%
  group_by(groups)%>%
  summarise(Romeo_score=sum(score))
```

Let us plot the graph for this score.

```
ggplot()+
  geom_point(data=Romeo_senti_score,aes(x=groups,y=Romeo_score))+
  geom_line(data=Romeo_senti_score,aes(x=groups,y=Romeo_score))
```



References

Feinerer, I. and Hornik, K. (2017). *tm: Text Mining Package*. R package version 0.7-1.

Fellows, I. (2014). *wordcloud: Word Clouds*. R package version 2.5.

Robinson, D. and Silge, J. (2017). *tidytext: Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools*. R package version 0.1.4.

Shakespear, W. (1597). *Romeo and Juliet : A tragic Love story*. unauthorized quarto.

Silge, J. (2017). *janeaustenr: Jane Austen's Complete Novels*. R package version 0.1.5.

Wickham, H. (2017). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.2.0.

Wickham, H., Francois, R., Henry, L., and Mller, K. (2017). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.4.