

Sense & Sensibility Wordcloud

Ron Richardson

November 1, 2017

Abstract

This article we construct a wordcloud using Jane Austen's book "Sense & Sensibility", via the R package, janeaustenr.

1 Introduction

Sense & Sensibility is a novel was written in 1811 by Jane Austen¹. This article will show how to construct a wordcloud with the most commonly used words in the book.

2 The Jane Austen Package

There is a relatively new package for R called janeaustenr. This package contains all the novels written by Jane Austen (Silge and Robinson, 2017). First, we need to install this package and load it in with the library command. Then, by calling the following function and store the result into a dataframe.

```
library(janeaustenr)
sns<-austen_books()
```

This dataframe has two columns, one for each line in the novel, and another with the title of novel the line of text is from. Let's first filter using dplyr so we have only just the lines from *Sense & Sensibility*

```
library(dplyr)
sns<-sns%>%
  filter(book == 'Sense & Sensibility')
print(sns, n=20)

## # A tibble: 12,624 x 2
##
##
```

text
<chr>

¹The novel was published anonymously.

```
## 1          SENSE AND SENSIBILITY
## 2
## 3          by Jane Austen
## 4
## 5          (1811)
## 6
## 7
## 8
## 9
## 10         CHAPTER 1
## 11
## 12
## 13 The family of Dashwood had long been settled in Sussex. Their estate
## 14 was large, and their residence was at Norland Park, in the centre of
## 15 their property, where, for many generations, they had lived in so
## 16 respectable a manner as to engage the general good opinion of their
## 17 surrounding acquaintance. The late owner of this estate was a single
## 18 man, who lived to a very advanced age, and who for many years of his
## 19 life, had a constant companion and housekeeper in his sister. But her
## 20 death, which happened ten years before his own, produced a great
## # ... with 1.26e+04 more rows, and 1 more variables: book <fctr>
```

Now we are ready for some data cleaning.

3 Data Cleaning

We would like to get the records that only contain the name and number of the chapter. To do this, we find any records that start with ‘Chapter’. We can use `dplyr` again, along with the R package called `stringr`.

```
library(stringr)
sns<-sns%>%
  filter(!str_detect(sns$text, "^CHAPTER"))
print(sns, n=20)

## # A tibble: 12,574 x 2
##                               text
##                               <chr>
## 1          SENSE AND SENSIBILITY
## 2
## 3          by Jane Austen
## 4
## 5          (1811)
## 6
## 7
```

```
## 8
## 9
## 10
## 11
## 12 The family of Dashwood had long been settled in Sussex. Their estate
## 13 was large, and their residence was at Norland Park, in the centre of
## 14 their property, where, for many generations, they had lived in so
## 15 respectable a manner as to engage the general good opinion of their
## 16 surrounding acquaintance. The late owner of this estate was a single
## 17 man, who lived to a very advanced age, and who for many years of his
## 18 life, had a constant companion and housekeeper in his sister. But her
## 19 death, which happened ten years before his own, produced a great
## 20 alteration in his home; for to supply her loss, he invited and received
## # ... with 1.255e+04 more rows, and 1 more variables: book <fctr>
```

We also want to remove the introduction text, as that is not relevant. We can find the positions of this by using the `head()` function to see what the line indexes are. Doing this, we see that the header ends at record 11, so we can redefine it to start at 12.

```
sns<-sns[12:12562,]
```

4 The Wordcloud

First off, we want break apart each line into individual words, by using a `tidytext` function to un-nest tokens()

```
library(tidytext)
snsWords<-sns%>%
  unnest_tokens(word, text)
```

Now we want to remove any unimportant words, called stop words. These words include ‘the’, ‘and’, ‘a’, and several others.

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

head(snsWords)

## # A tibble: 6 x 2
##           book      word
##           <fctr>   <chr>
## 1 Sense & Sensibility family
## 2 Sense & Sensibility dashwood
```

```
## 3 Sense & Sensibility    settled
## 4 Sense & Sensibility    sussex
## 5 Sense & Sensibility    estate
## 6 Sense & Sensibility    residence
```

The last step before building our wordcloud is to get a count of each word using dplyr.

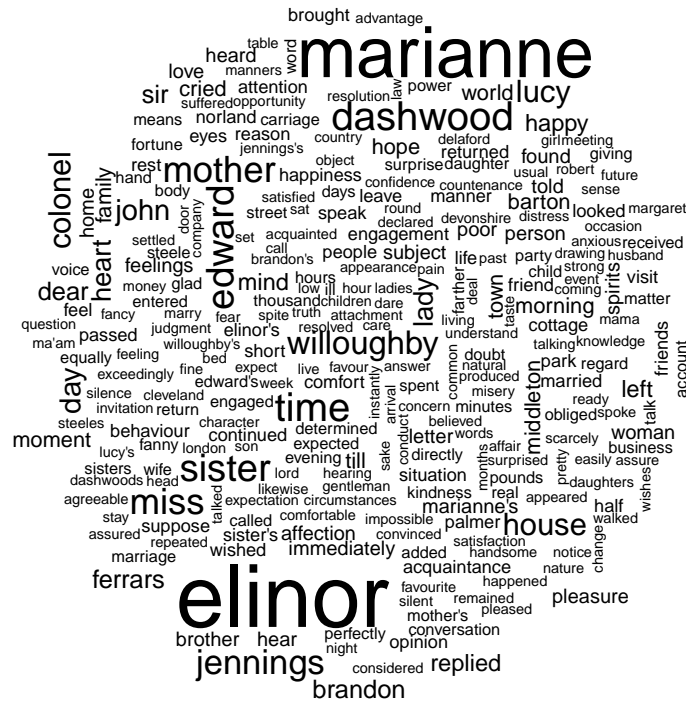
```
snsWordFreq<-snsWords%>%
  group_by(word)%>%
  summarize(count=n())

head(snsWordFreq)

## # A tibble: 6 x 2
##       word count
##   <chr> <int>
## 1      1      1
## 2     200      1
## 3   70001      1
## 4 abandoned      1
## 5 abatement      1
## 6 abbeyland      1
```

Finally, we can add the words and frequencies to the wordcloud function. Since there are so many unique words in the novel, we want to only include the words that occur more than 25 times.

```
library(wordcloud)
wordcloud(snsWordFreq$word, snsWordFreq$count, min.freq=25)
```



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.
- Silge, J. (2017). *janeaustenr: Jane Austen's Complete Novels*. R package version 0.1.5.
- Silge, J. and Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media.
- 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 0.7.2.
- Wickham, H. and Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.