Sense and Sensibility Wordcloud

Justin Minsk

October 27, 2017

Abstract

In this atricle we will construct a wordcloud, using the tidytext R pakage. The text we will be using is Jane Austen's Sense and Sensibility.

1 Sense and Sensibility

Sense and Sensibilty is a novel by Jane Austen, published in 1811. We will create a wordcloud of the most used words in the book.

2 The Jane Austen Package

There is a relatively new package for R, janeaustenr, that gives access to all of her books. You need to install the package then bring in the library. Then use austenbooks to get the data and can input the data into a data frame.

```
library(janeaustenr)
sns <- austen_books()
#bring in all of jane austen's books
```

This data frame has two columns named books and text.

```
library(dplyr)
library(tidytext)
library(tm)
sns <- sns%>%
  filter(book == "Sense & Sensibility")
#filter to just get Sense and Sensibility text
head(sns)

## # A tibble: 6 x 2
##                    text               book
##                   <chr>             <fctr>
## 1 SENSE AND SENSIBILITY Sense & Sensibility
## 2                       Sense & Sensibility
```

```
## 3        by Jane Austen Sense & Sensibility
## 4                       Sense & Sensibility
## 5                (1811) Sense & Sensibility
## 6                       Sense & Sensibility
```

This data frame only contins the book Sense and Sensibility now. Now to clean the data.

3 Some Data-Cleaning

First we want to remove all of the 'Chapter' lines.

```
library(stringr)
sns$book <- as.character(sns$book)
#un factor the book column
sns <- sns%>%
  filter(!str_detect(sns$text, "^CHAPTER"))
#filter out all of the CHAPTER lines
```

Next we use indexes to remove the beging and the ending.

```
sns <- sns[12:12574, ]
#make sns start after the title

sns <- sns[1:12560, ]
#get rid of the end
```

Now we need to get the words by themselves and make sure that we have a count.

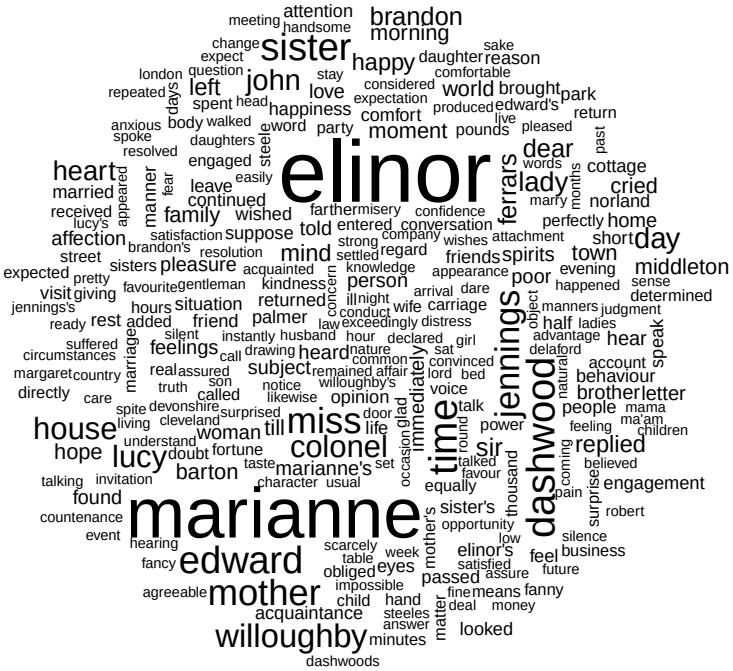
```
words_df <- sns%>%
  unnest_tokens(word, text)
#split the lines into words

words_df <- words_df%>%
  filter(!(word %in% stop_words$word))
#get rid of common words that are not unique (the, a, etc.)

words_free <- words_df%>%
  group_by(word)%>%
  summarise(count = n())
#make a count of the word
```

Then we make a wordcloud

4 Wordcloud

First we need the library then we cna make the wordcloud.

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




```
#make a word cloud
```