Post 1: Sentiment Analysis & its Applications

Stat 133, Fall 2017

Tyler Larsen

October 31, 2017

> ## Motivation

Up until this point, we have learned a diverse set of skills in R that on their own may not seem to be too remarkable. However, when combined together, the realm of possibilities these different skills and functionalities afford is endless. When I was looking for a topic area in R to explore further, I wanted to find a way that the skills we have already learned can and are currently being used in industry. One such way that the skills we have learned so far are being applied in industry is through the use of a technique known as Sentiment Analysis.

> ## Introduction

In this blog post, I will first describe what Sentiment Analysis is, then where it is applied, and then I will go through an example of Sentiment Analysis using the Jane Austen novels.

> ## Background

Sentiment Analysis, also known as Opinion Mining, in a process in which the sentiment, or attitude, from a piece of text is extracted and then quantified in order to draw some sort of actionable conclusion. First a given text is broken down, usually by word or sentence, which are referred to as tokens, and then these tokens are assigned a value, in the simplest case these values could be -1 for negative opinion, 0 for a neutral opinion, and then 1 for a positive opinion. These values can be summed to find the sentiment at any given moment in time, but more interestingly they can also be graphed over a time interval to see how the sentiment changes over time.

There are many places where this analysis gets applied in the real world. For example, companies may use this analysis to gauge how much their customers like a new product. By analyzing customer reviews, a company can quickly find out whether people like, dislike, or are neutral towards a product, and then go and make changes accordingly. Without this form of analysis, companies would have to manually read through reviews to gauge customer response, an incredibly time consuming task, but with this analysis they can reach the same conclusions almost immediately.

This form of analysis is also commonly applied to social media networks to answer a host of different questions. In politics, Sentiment Analysis is often used to gauge the impact of campaign tactics. For example, after giving a speech in San Francisco an aspiring politician could then record and analyze the Tweets after his or her speech to see what people did and didn't like, and then tailor his or her message to better suit the audience the next time around.

> ## Example: Sentiment Analysis on the Jane Austen Novels

Sentiment Analysis is a three step process. First the data must be cleaned and formatted. In this example we'll be using the package 'janeaustenr,' and so our data has already been cleaned, but still it needs to be formatted. Then, we must apply the actual analysis. This includes choosing which form of Sentiment analysis we would like to use (I go more into the different options later), and then storing this data in a data-frame. Last comes the fun part, choosing how we would like to display our data.

In order to conduct Sentiment Analysis of our own, we first need to download the following packages:

```
library(tidytext)       # text mining and sentiment analysis functions
library(janeaustenr)    # collection of the Jane Austen novels
library(dplyr)          # data manipulation functions
library(tidyr)          # format data table for sentiment analysis
library(ggplot2)        # graphically displays our results
```

The following code creates a data frame "books", with a column of the text (separated by line), the book, and the corresponding line number.

```
books <- austen_books() %>%
  group_by(book) %>%
  mutate(linenumber = row_number()) %>%
  ungroup()

head(books, 20)
```

```
## # A tibble: 20 x 3
##                                                                       text
##                                                                      <chr>
##  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 2 more variables: book <fctr>, linenumber <int>
```

In order to to sentiment analysis, we must further break down the data-frame using the tidytext package, so that it is in a one-word-per-line format:

```
books_by_words <- unnest_tokens(books, word, text)

head(books_by_words)
```

```
## # A tibble: 6 x 3
##                   book linenumber        word
##                 <fctr>      <int>       <chr>
## 1 Sense & Sensibility          1       sense
## 2 Sense & Sensibility          1         and
## 3 Sense & Sensibility          1 sensibility
## 4 Sense & Sensibility          3          by
## 5 Sense & Sensibility          3        jane
## 6 Sense & Sensibility          3      austen
```

Next, we create a data-frame janeaustensentient to capture the sentiment in every word. There are three different arguments that get_sentiments() can take: “afinn,” “bing,” and “nrc.” These three arguments, known as lexicons, represent three different ways of evaluating the sentiment of each word, all with very similar output.

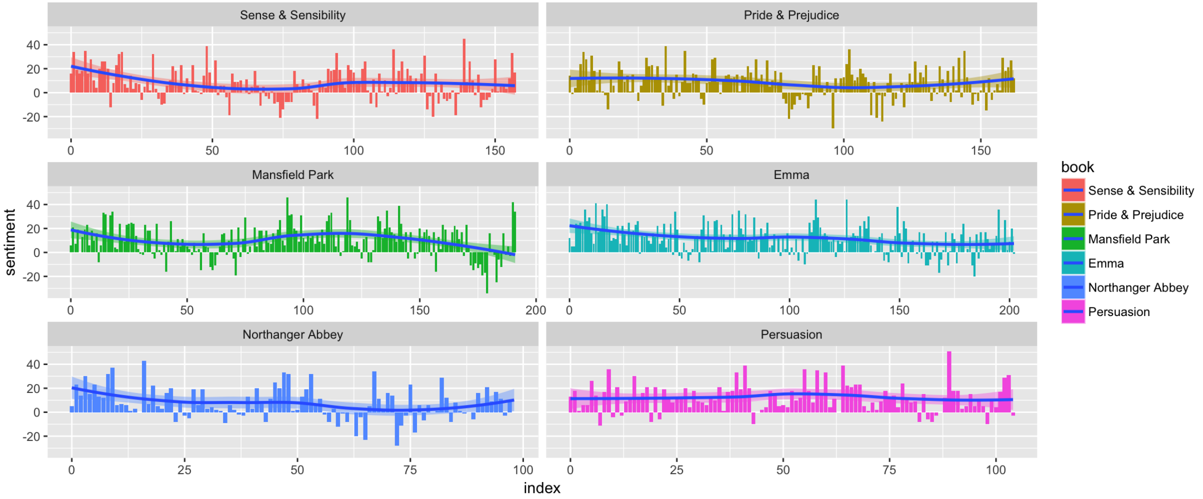
```
janeaustensentiment <- books_by_words %>%
  inner_join(get_sentiments("bing"), by = "word") %>%
  count(book, index = linenumber %/% 80, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)

head(janeaustensentiment, 20)
```

```
## # A tibble: 20 x 5
##                    book index negative positive sentiment
##                  <fctr> <dbl>    <dbl>    <dbl>     <dbl>
##  1 Sense & Sensibility     0       16       32        16
##  2 Sense & Sensibility     1       19       53        34
##  3 Sense & Sensibility     2       12       31        19
##  4 Sense & Sensibility     3       15       31        16
##  5 Sense & Sensibility     4       16       34        18
##  6 Sense & Sensibility     5       16       51        35
##  7 Sense & Sensibility     6       24       40        16
##  8 Sense & Sensibility     7       23       51        28
##  9 Sense & Sensibility     8       30       40        10
## 10 Sense & Sensibility     9       15       19         4
## 11 Sense & Sensibility    10       12       32        20
## 12 Sense & Sensibility    11       14       40        26
## 13 Sense & Sensibility    12       22       48        26
## 14 Sense & Sensibility    13       22       42        20
## 15 Sense & Sensibility    14       36       24       -12
## 16 Sense & Sensibility    15       23       31         8
## 17 Sense & Sensibility    16       15       38        23
## 18 Sense & Sensibility    17       15       47        32
## 19 Sense & Sensibility    18       19       53        34
## 20 Sense & Sensibility    19       31       38         7
```

Now that we have the sentiment for each novel, we can graph it for each novel. The vertical axis on the following graphs represents the sentiment. The more positive the value of the sentiment, the more positive the tone of the text, and the more negative the sentiment the more negative the tone of the text is. The horizontal axis represents the progression through the novel. I have also included a best fit line to track the relative changes in sentiment as the novel progresses.

```
ggplot(janeaustensentiment, aes(index, sentiment, fill = book)) +
  geom_bar(stat = "identity", show.legend = TRUE) +
  facet_wrap(~book, ncol = 2, scales = "free_x") +
  geom_smooth()
```



Discussion

In this example, the hardest part of the Data Analysis process, the data collection, has been taken care of for us. To take this one step further, I want to find a way to import live Twitter data, instead of using the preloaded Jane Austen books. From there, the analysis would be the exact same.

Take Home Message

Sentiment Analysis is a highly applicable technique for mining opinions out of pieces of text and then quantifying these opinions in order to make business decisions.

Sources

https://www.rdocumentation.org/packages/dplyr/versions/0.7.3/topics/join
http://uc-r.github.io/sentiment_analysis
https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf
http://stat545.com/bit001_dplyr-cheatsheet.html
http://tidytextmining.com/sentiment.html
https://www.r-bloggers.com/intro-to-text-analysis-with-r/
https://en.wikipedia.org/wiki/Sentiment_analysis