Wordcloud of Top 75 Words in Charlotte Bronte’s Jane Eyre

Using the text of Jane Eyre from Project Gutenberg, we will plot the top 75 words in the novel. [http://www.gutenberg.org/cache/epub/1260/pg1260.txt]

Global parameters

Set working directory by pointing to the location on your computer where you have stored the files. Below, we have chosen to Save the folder “RAnalysis” on the Desktop on a Mac. It contains all the other R scripts, texts, notebooks, and results. If you have branched the github, simply note where you have save the folder. If you are on a PC, you will need to use an absolute path such as “C:Users:XXX.” ***

```r
setwd("~/Desktop/R/Text_Analysis/RNotebooks/")
```

Include necessary packages for notebook ***

```r
library(knitr)
library(markdown)
library(rmarkdown)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```r
library(tm)
```

```
## Loading required package: NLP
```

```r
library(RColorBrewer)
```

Load data (plaintext file), telling the computer that it is formatted by line breaks
**

```r
text_raw<-scan("~/Desktop/R/Text_Analysis/data/bronte/janeEyre.txt", what="character", sep="\n")
```

Prepare text data The tm library works with a “corpus,” which can represent a single text, many texts, and even different types of inputs, such as CSV files. In this case, we only have one text in our corpus, but will follow tm’s naming conventions for the variables.

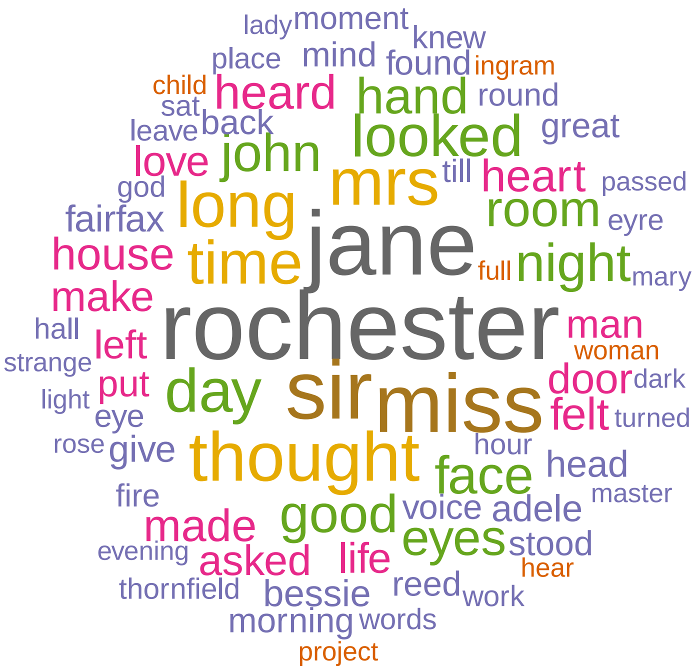
```r
#Create a corpus
corpus <- Corpus(VectorSource(text_raw))
```

One of the great things about using tm is that it has many text cleaning functions built into it. Here we have removed extra white space, transformed all the text to lower case, removed stopwords, and removed punctuation. It is possible to simply use an "english" (or other language) stopword list, but I have found the SMART list to be more complete. For example, "said" is on the SMART stopword list but not on the standard "english" list. ***

```r
corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removeWords, stopwords("SMART"))
corpus <- tm_map(corpus, removePunctuation)
```

Plot the wordcloud! There are a number of arguments you can customize: "random.order" is false so that words are plotted in order of decreasing frequency; "scale"" indicates the size of the words; "rot.per" lets you customize the proportion of words that are rotated 90 degrees; "max.words" controls how many words show up in the wordcloud; and we've used the library "RColorBrewer" to give us access to some predefined palettes. Note that if you change the palette, you need to tell the argument how many colors are in the new palette. ***

```r
wordcloud(corpus,random.order=FALSE,scale=c(3,.1),rot.per=0,
          max.words=75,colors=brewer.pal(8, "Dark2"))
```



Voila!