Untitled

Jia Wang (UNI: jw3315)

November 26, 2016

```
library(tidytext)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(RColorBrewer)
library(NLP)
library(janeaustenr)
library(stringr)
library(tidyverse)
```

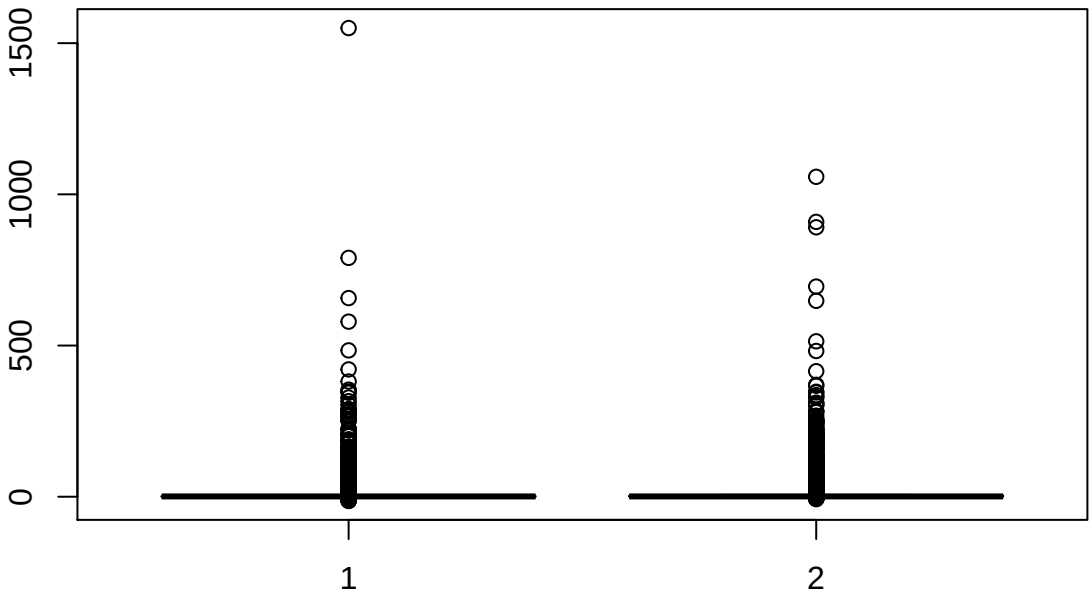
```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

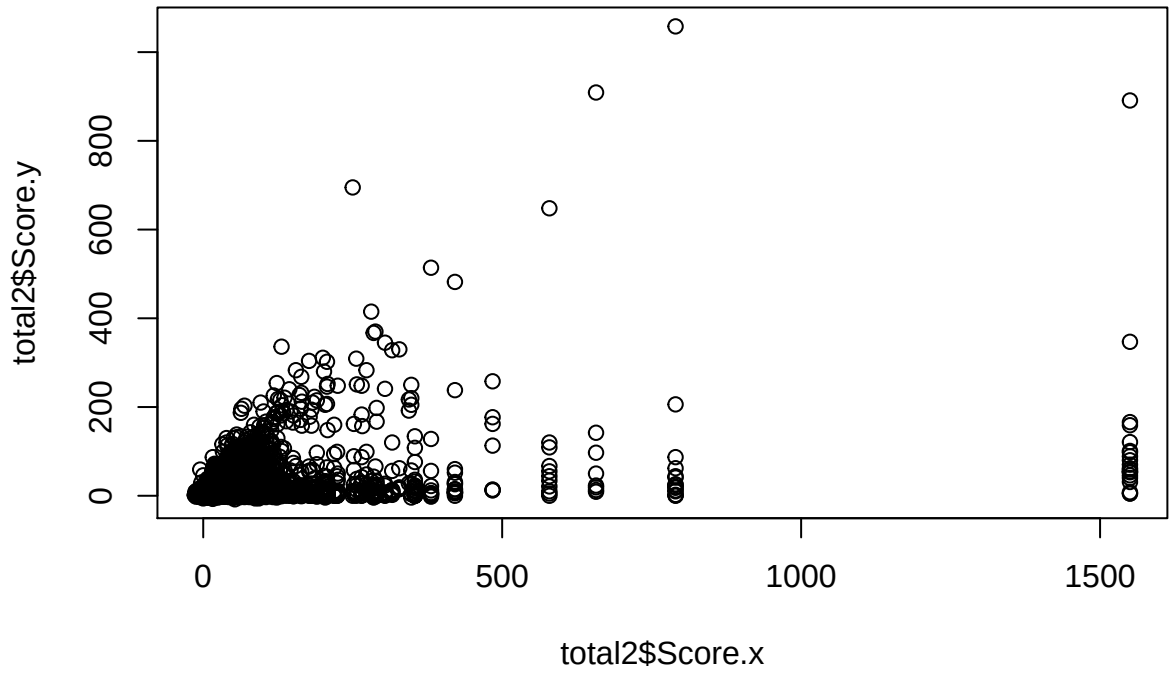
```
## annotate(): ggplot2, NLP
## filter():   dplyr, stats
## lag():      dplyr, stats
```

```
library(readr)
library(tm)
library(wordcloud)

#######
#  r  #
#######

r_questions_clean <- suppressMessages(read_csv("/var/folders/sg/2ybq97vj0nx4k1n4yxm321bw0000gn/T//Rtmpf
r_answers_clean <- suppressMessages(read_csv("/var/folders/sg/2ybq97vj0nx4k1n4yxm321bw0000gn/T//Rtmp6KH
r_Tags <- read.csv("~/Downloads/rquestions/Tags.csv")
```

```r
# merge questions and tags to find connections
names(r_Tags)[names(r_Tags)=="Id"]="ParentId"
names(r_questions_clean)[names(r_questions_clean)=="Id"]="ParentId"
total1<- merge(r_questions_clean, r_Tags,by="ParentId")
total2<-merge(r_questions_clean,r_answers_clean,by="ParentId")

# Score
boxplot(r_questions_clean$Score,r_answers_clean$Score)
```




```r
plot(total2$Score.x,total2$Score.y)
```



```r
mean(r_answers_clean$Score[r_answers_clean$IsAcceptedAnswer=="True"])
```

```
## [1] 3.741573
```

```r
mean(r_answers_clean$Score[r_answers_clean$IsAcceptedAnswer=="False"])
```

```
## [1] 2.115549
```

```r
median(r_answers_clean$Score[r_answers_clean$IsAcceptedAnswer=="True"])
```

```
## [1] 2
```

```r
median(r_answers_clean$Score[r_answers_clean$IsAcceptedAnswer=="False"])
```

```
## [1] 1
```

```r
# how long you need to wait after creating a question
time_wait.d<-difftime(total2$CreationDate.y,total2$CreationDate.x,units = "days")
mean(time_wait.d)
```

```
## Time difference of 47.96628 days
```

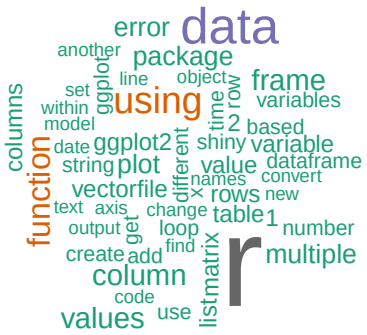
```r
median(time_wait.d)
```

```
## Time difference of 0.03233796 days
```

```r
# word frequency for question title
title_words <- r_questions_clean %>%
  select(ParentId, Score, CreationDate,Title) %>%
  unnest_tokens(word,Title)
freq_title_words<-title_words %>%
  count(word, sort = TRUE)
head(freq_title_words)
```

```
## # A tibble: 6 × 2
##       word     n
##      <chr> <int>
## 1        r 73552
## 2     data 23286
## 3    using 14315
## 4 function 10657
## 5   column  8772
## 6    frame  8619
```

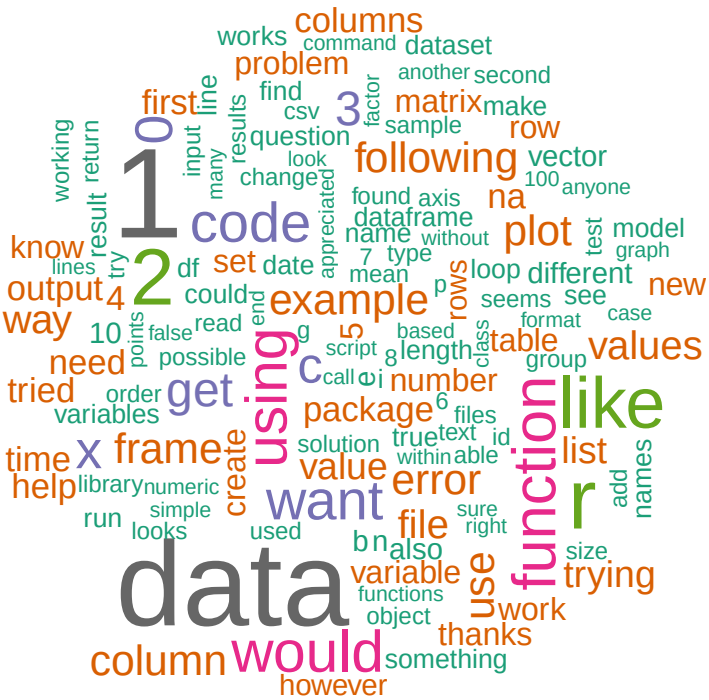
```r
set.seed(142)
dark2 <- brewer.pal(8, "Dark2")
wordcloud(freq_title_words$word,freq_title_words$n,min.freq = 2000,rot.per=0.2, colors=dark2)
```




```
# word frequency for question body
questionbody_words <- r_questions_clean %>%
  select(ParentId, Score, CreationDate,Body) %>%
  unnest_tokens(word,Body)
freq_questionbody_words<-questionbody_words %>%
  count(word, sort = TRUE)
head(freq_questionbody_words)
```

```
## # A tibble: 6 × 2
##       word      n
##      <chr>  <int>
## 1        1 195705
## 2     data 181039
## 3        r 119299
## 4        2 109700
## 5     like  99026
## 6 function  78152
```

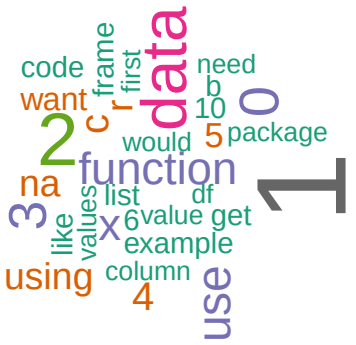
```
wordcloud(freq_questionbody_words$word,freq_questionbody_words$n,min.freq = 10000,rot.per=0.2, colors=da
```



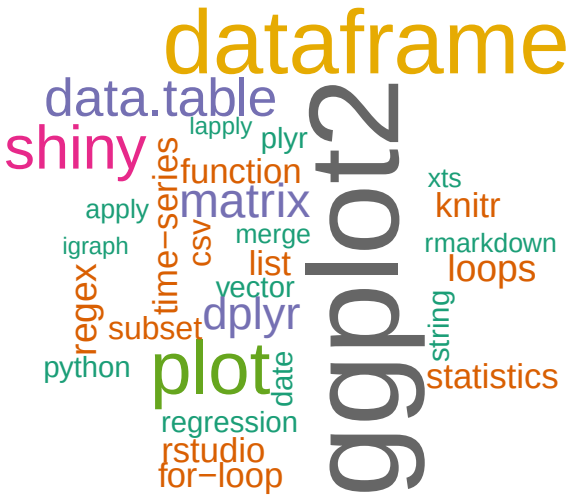
```
# word frequency for answer body
answerbody_words <- r_answers_clean %>%
  select(ParentId, Score, CreationDate,Body) %>%
  unnest_tokens(word,Body)
freq_answerbody_words<-answerbody_words %>%
  count(word, sort = TRUE)
head(freq_answerbody_words)
```

```
## # A tibble: 6 × 2
##    word      n
##   <chr>  <int>
## 1     1 313610
## 2     2 163089
## 3  data 133029
## 4     0 112440
## 5     3 102593
## 6   use  88825
```

```
wordcloud(freq_answerbody_words$word,freq_answerbody_words$n,min.freq = 30000,rot.per=0.2, colors=dark2)
```




```
# Tags
freq_r_Tags<-r_Tags %>%
  count(Tag, sort = TRUE)
wordcloud(freq_r_Tags$Tag,freq_r_Tags$n,min.freq = 1000,rot.per=0.2, colors=dark2)
```



```
############
#  python  #
############
python_questions_features <- suppressMessages(read_csv("/var/folders/sg/2ybq97vj0nx4k1n4yxm321bw0000gn/
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
python_answers_features <- suppressMessages(read_csv("/var/folders/sg/2ybq97vj0nx4k1n4yxm321bw0000gn/T/
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
python_Tags <- read.csv("~/Downloads/Tags.csv")
names(python_Tags)[names(python_Tags)=="Id"]="ParentId"
names(python_questions_features)[names(python_questions_features)=="Id"]="ParentId"
total.python.tag.question<-merge(python_Tags,python_questions_features,by="ParentId")
total.python<-merge(python_questions_features,python_answers_features,by="ParentId")

####tag for python wordcloud
freq_python_Tags<-python_Tags %>%
  count(Tag, sort = TRUE)
wordcloud(freq_python_Tags$Tag,freq_python_Tags$n,min.freq = 5000,rot.per=0.2, colors=dark2)
```


```
################
#  comparison  #
################

## compare time to wait for an answer
time_wait.python.d<-difftime(total.python$CreationDate.y,total.python$CreationDate.x,units = "days")
mean(time_wait.python.d)
```

```
## Time difference of 70.01714 days
```

```
median(time_wait.python.d)
```

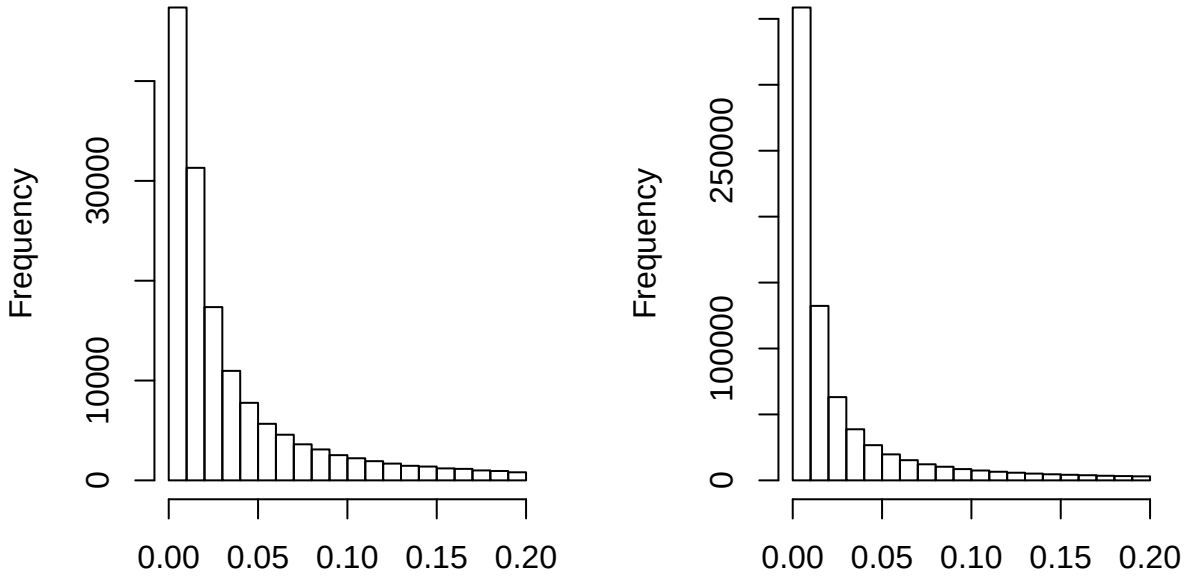
```
## Time difference of 0.02016204 days
```

```
wilcox.test(as.numeric(time_wait.d),as.numeric(time_wait.python.d))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  as.numeric(time_wait.d) and as.numeric(time_wait.python.d)
## W = 1.0871e+11, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
par(mfrow=c(1,2))
hist(as.numeric(time_wait.d[which(as.numeric(time_wait.d)<0.2 & as.numeric(time_wait.d)>0)]))
hist(as.numeric(time_wait.python.d[which(as.numeric(time_wait.python.d)<0.2 & as.numeric(time_wait.pytho
```

ait.d[which(as.numeric(time_wait.d)[which(as.numeric(time_wait.pytho



.d[which(as.numeric(time_wait.d) < 0.2 & as.hich(as.numeric(time_wait.python.d) < 0.2 &

```
###compare score of questions
wilcox.test(r_questions_clean$Score,python_questions_features$Score)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  r_questions_clean$Score and python_questions_features$Score
## W = 4.6141e+10, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
###compare score of answer
wilcox.test(r_answers_clean$Score,python_answers_features$Score)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  r_answers_clean$Score and python_answers_features$Score
## W = 1.0936e+11, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```