Activity in the d3 github repository

Martin Gascon

11/20/2015

Procedure

a) Fork the mbostock/d3 repository into my github account (terminal)

- git clone https://github.com/martingascon/d3

b) retrieve the commit activity and drop it into a log file

- git log –pretty=format:"%h,%an,%ad,%s" > log.txt

c) Load log file, format the date column

```
# read log.txt
log = read.csv("./d3/log.txt", sep=",",head=F,)
colnames(log) <-c("commit","author","date","subject")

# remove those with empty spaces in subject
log <-log[-which(log$subject == ""),]

# Convert the date column into a date type
log$date <- as.Date(log$date , "%a %b %d %H:%M:%S %Y %z")
```

Responses

1. What week in the last year had the greatest number of commits? 1.1 There are three possible answers depending on the ISO coding

- %U - Week Nb, starting with the first Sunday as the first day of the first week (00..53)
- %V - Week number of year according to ISO-8601 (01..53)
- %W - Week number of the year, starting with the first Monday as the first day of the first week (00..53)

```
Week_nb1 <- format(log$date, format="%y/%U")
sort(table(Week_nb1),decreasing=T)[1:3]
```

```
## Week_nb1
## 15/05 15/42 15/06
##    19    19    14
```

```
Week_nb2 <- format(log$date, format="%y/%V")
sort(table(Week_nb2),decreasing=T)[1:3]
```

```
## Week_nb2
## 15/06 15/43 14/49
##    25    23     8
```

```
Week_nb3 <- format(log$date, format="%y/%W")
sort(table(Week_nb3),decreasing=T)[1:3]
```

```
## Week_nb3
## 15/05 15/42 14/48
##    25    23     8
```

- *%U -> Weeks 5 and 42 in 2015 are ones with more commits (19)*
- *%V -> Week 6 in 2015 is one with more commits (25)*
- *%W -> Week 5 in 2015 is one with more commits (25)*

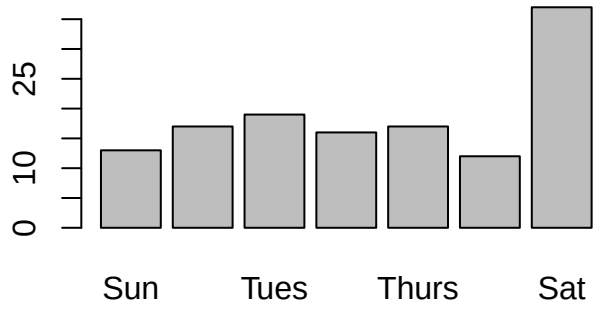
```
sort(table(log$date),decreasing=T)[1:3]
```

2. Over the last year, what day of the week had the most commits?

```
##
## 2015-02-07 2015-10-21 2014-12-06
##         16          8          7
```

- *Feb 7th, 2015 is the day with more commits (16)*

```
library(lubridate)
barplot(table(wday(log$date, label=TRUE)))
```



3. Graph the number of commits per day of the week.

- *Saturday is the day with more commits (37)*

---

4. Show us something else interesting about the d3 repository

- *Let's plot the most used words in the subject of every commit*

```
# tm: text mining library for R
library(tm)
```

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

```
# paste every subject together
subject_vector <- paste(log$subject, collapse=" ")

# set up the source and create a corpus
subject_source <- VectorSource(subject_vector)
corpus <- Corpus(subject_source)

# Now, we transform to lower, remove punctuation, whitespace, stopwords
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, removeWords, stopwords("english"))

# create the document-term matrix.
dtm <- DocumentTermMatrix(corpus)
dtm2 <- as.matrix(dtm)

#most frequently used words

frequency <- colSums(dtm2)
frequency <- sort(frequency, decreasing=TRUE)
head(frequency)
```

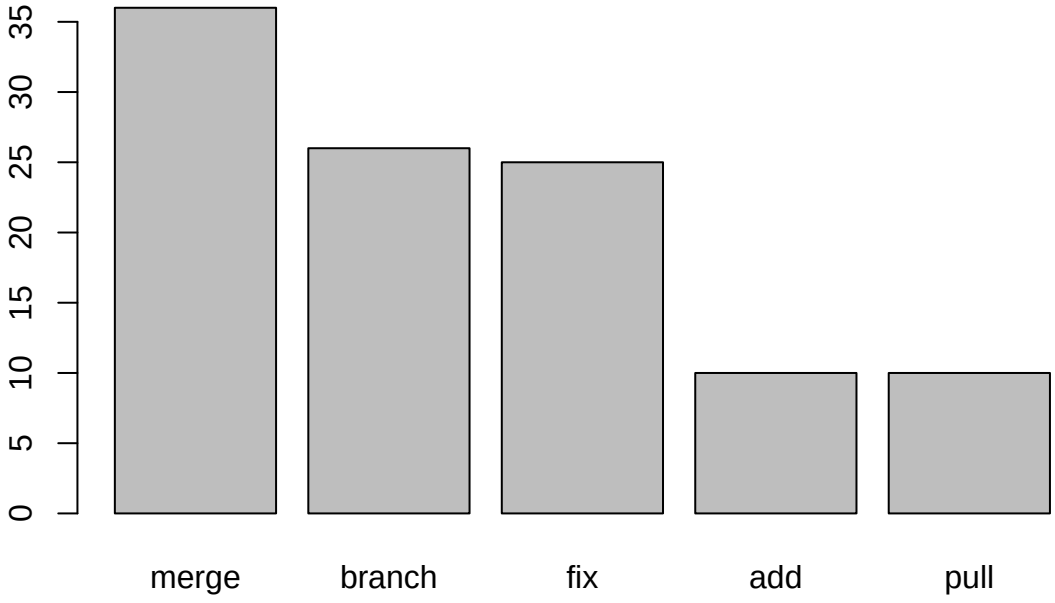
```
##   merge  branch     fix     add    pull request
##      36      26      25      10      10      10
```

```
barplot(frequency[1:5], label=TRUE)
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "label" is not a
## graphical parameter

## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "label" is not a graphical parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "label" is not a graphical parameter
```



- *Merge, branch and fix are the three most popular words in the subject of every commit.*

- *Another way to plot the 15 most frequently used words.*

```
library(wordcloud)
```

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

```
words <- names(frequency)
wordcloud(words[1:15], frequency[1:15])
```