Homework 9

James Chen

October 31, 2018

```
##
## 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
```

Introduction

This week, we were able to load monthly JSON data through the stream_in function from jsonlite directly into our SQL database to avoid the memory issue. However, for the text mining part of our project for the group homework, we decided to work with a single day of data so we could easily manipulate it in R. For now, we're going to keep working on the daily data to prototype our methods, since they're still large enough to be fairly representative of typical Reddit activity. Adapting our analysis to the monthly data sets should only require minor tweaking to process the data in manageable chunks through a loop.

Netspeak frequency

We began our analysis of the reddit comments by using regular expression and stringr to count the number of times popular "netspeak" was used in a comment. We started with the simple case of netspeak lingo due to the ease of implementation with our data and to practice extracting strings before we tackle more complicated information. The end goal of our project is to implement similar code to search for popular memes and copypasta and to explore their propagation throughout reddit over time.

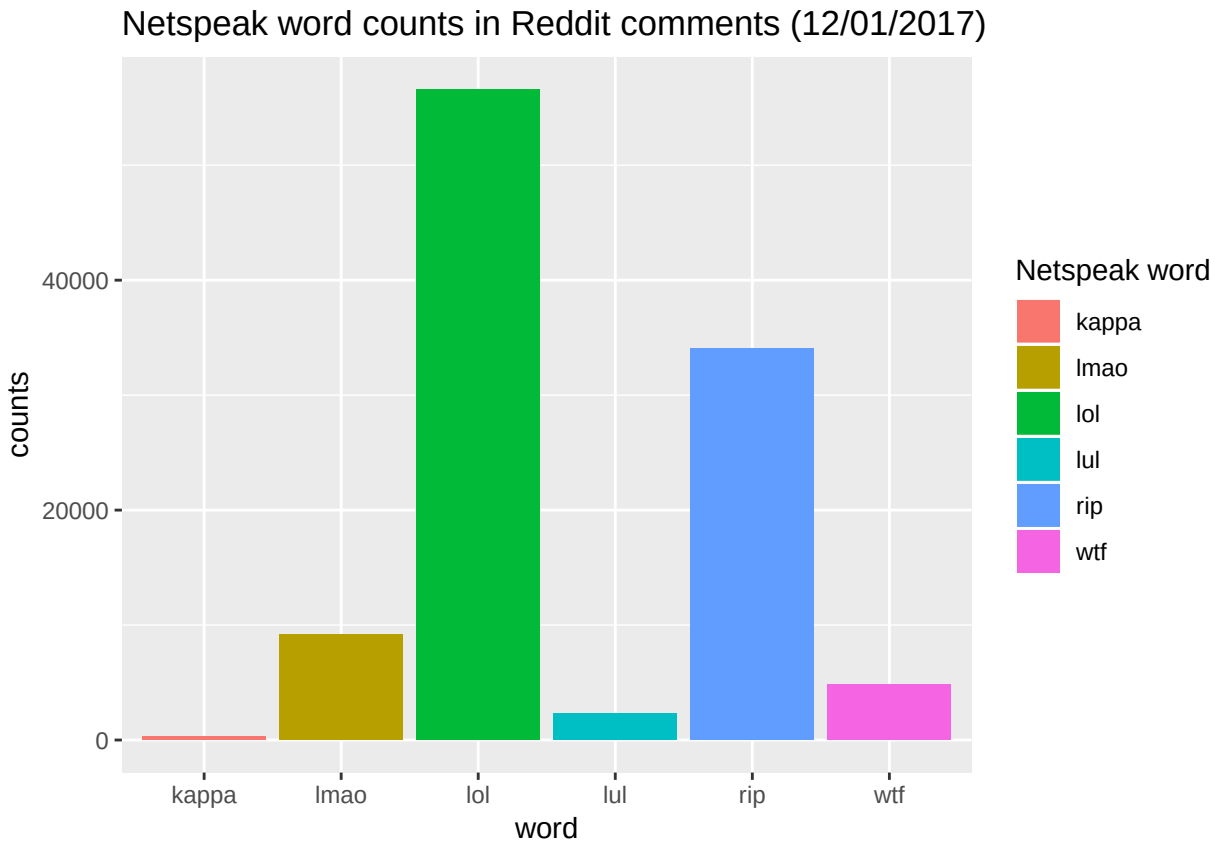
Note that these were found out of a total of about 3 million comments.

The table below shows "lol" was the most frequently used phrase in our list of comments, but we can see that the new "lul" phrase with identical meaning is also seeing some use on reddit. We can also compare that to the popular phrase "lmao" which also conveys the same information. An interesting direction to take this knowledge for our project will be to explore how the frequency of the use of these phrases has changed over time as a way of visualizing the rise in popularity of new netspeak such as "lul". Another interesting observation in the table is the surprising amount of the use of "RIP". We also looked at "wtf" and "kappa" (new lingo that means sarcasm") and will expand our analysis to several other popular phrases for our project.

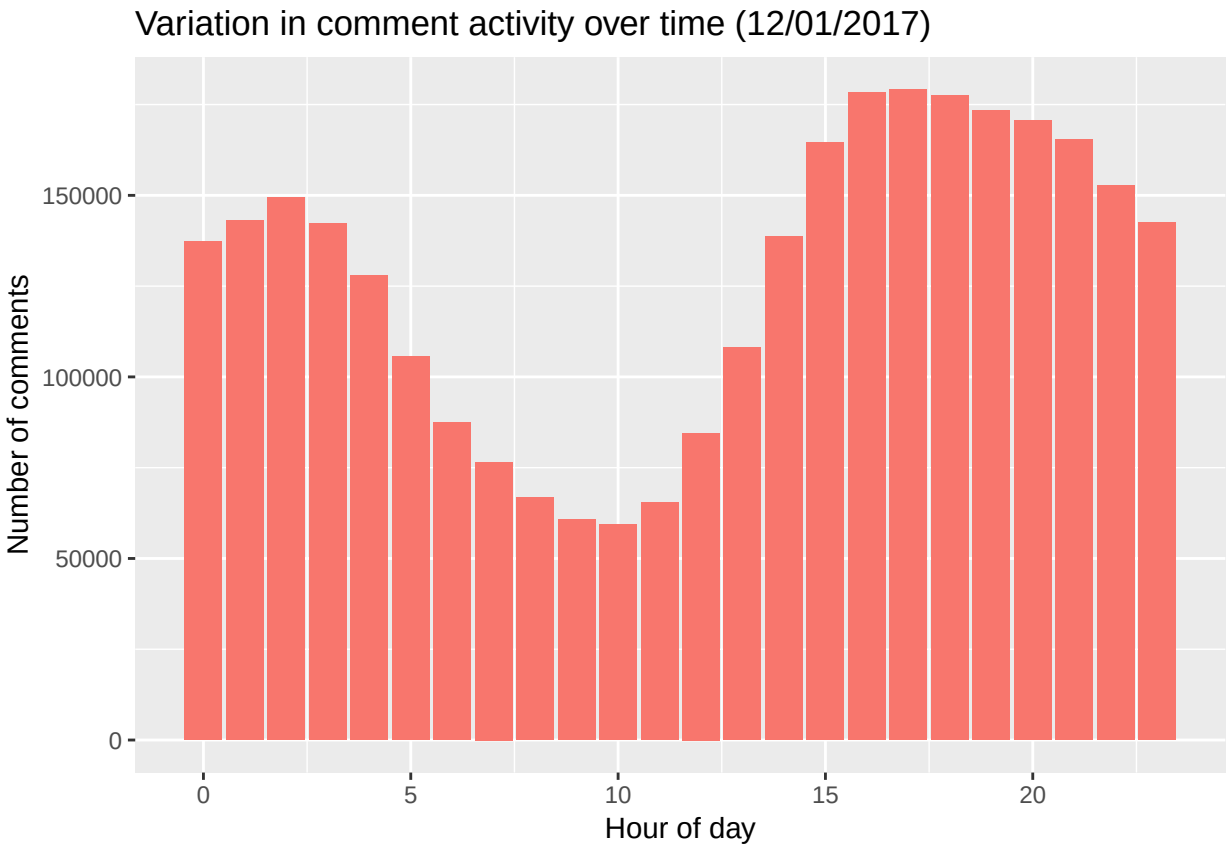
```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.0.0     v readr   1.1.1
## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------------------ tidy
## x dplyr::filter()  masks stats::filter()
## x purrr::flatten() masks jsonlite::flatten()
## x dplyr::lag()     masks stats::lag()
```



In addition to our netspeak analysis, we also wanted to visualize user activity on reddit on an hourly basis. We see a peak at 3pm (UTC), a second (smaller) peak at 2am (UTC). Since these match up to 11 am EST and 10 pm EST, this corresponds to peak usage as people are waking in the morning and as night time progresses. Assuming the previous is true (which makes sense for average use activity) this data also shows that there is a bias in the comments from the US, since seeing relatively equal bars would indicate equal usage worldwide at respective local peak times. An interesting direction to take this data analysis further as we work on our project will be to correlate the length of the comment (and/or amount of comments) to the age of the user account on reddit to see if there is a relationship between how much a user writes as their time on reddit increases.


Variation in comment activity over time (12/01/2017)
Number of comments
150000
100000
50000
0
0
5
10
15
20
Hour of day