HW 11

James Chen, Hoik Jang, Steven Oliver, and Adrian Perez

November 15, 2018

Combining old plots

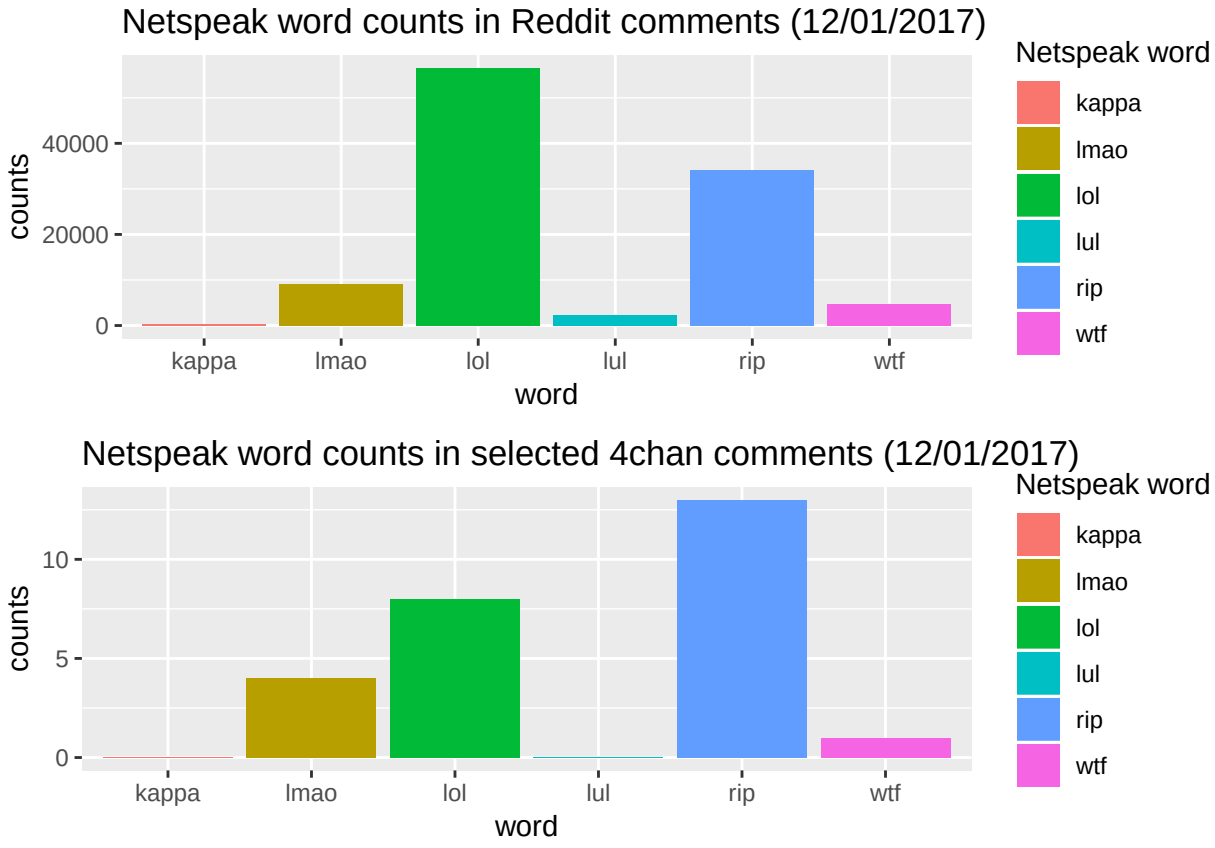
In the last two weeks, we investigated the relative frequency of certain netspeak words in Reddit and in 4chan. This week, we take the natural step of showing them side by side.

```
##
## 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
```

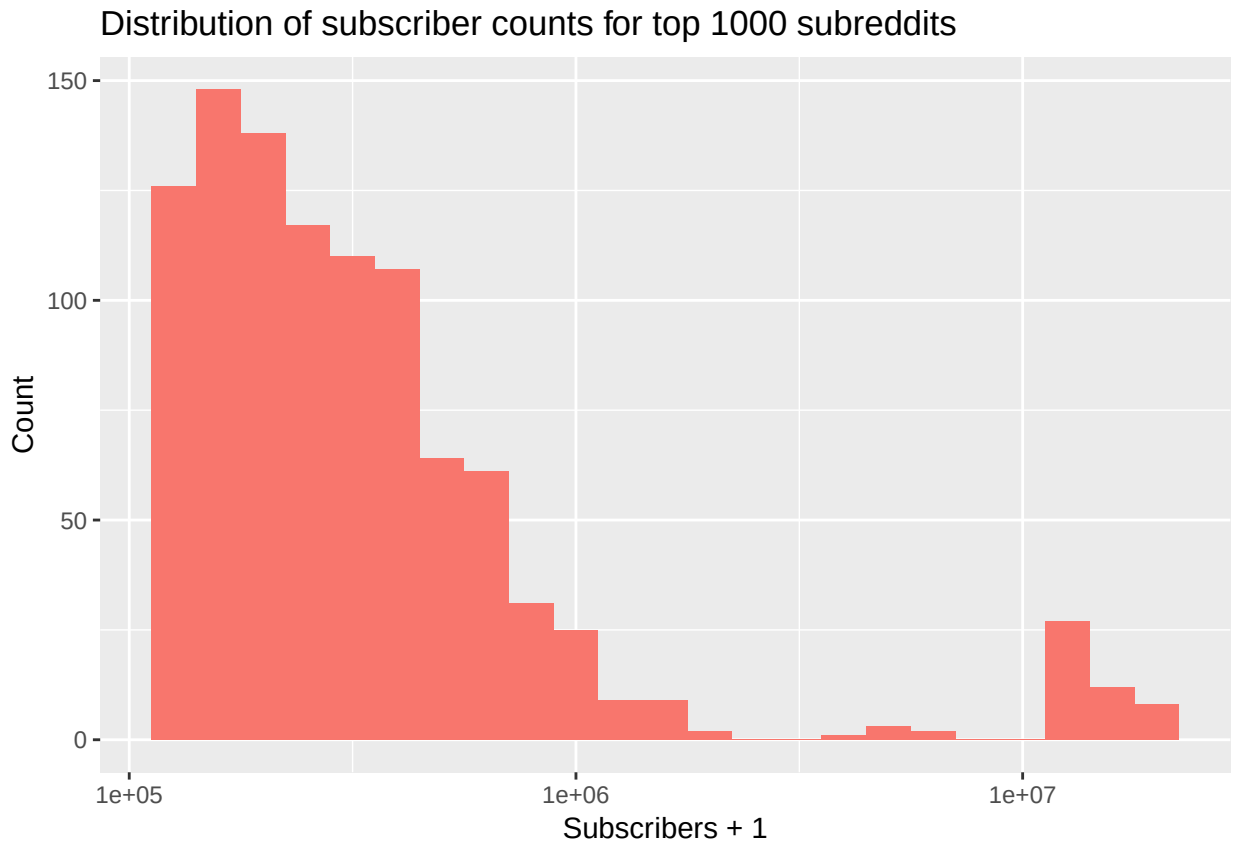
Again, we are limited because of the small sample size, as we weren't able to get the scraping of 4chan done on the desktop we'd set up as a server over the weekend. However, it is interesting to see that lmao, lol, and rip are the most prominent examples of netspeak on both Reddit and 4chan, while the other three are much more obscure.



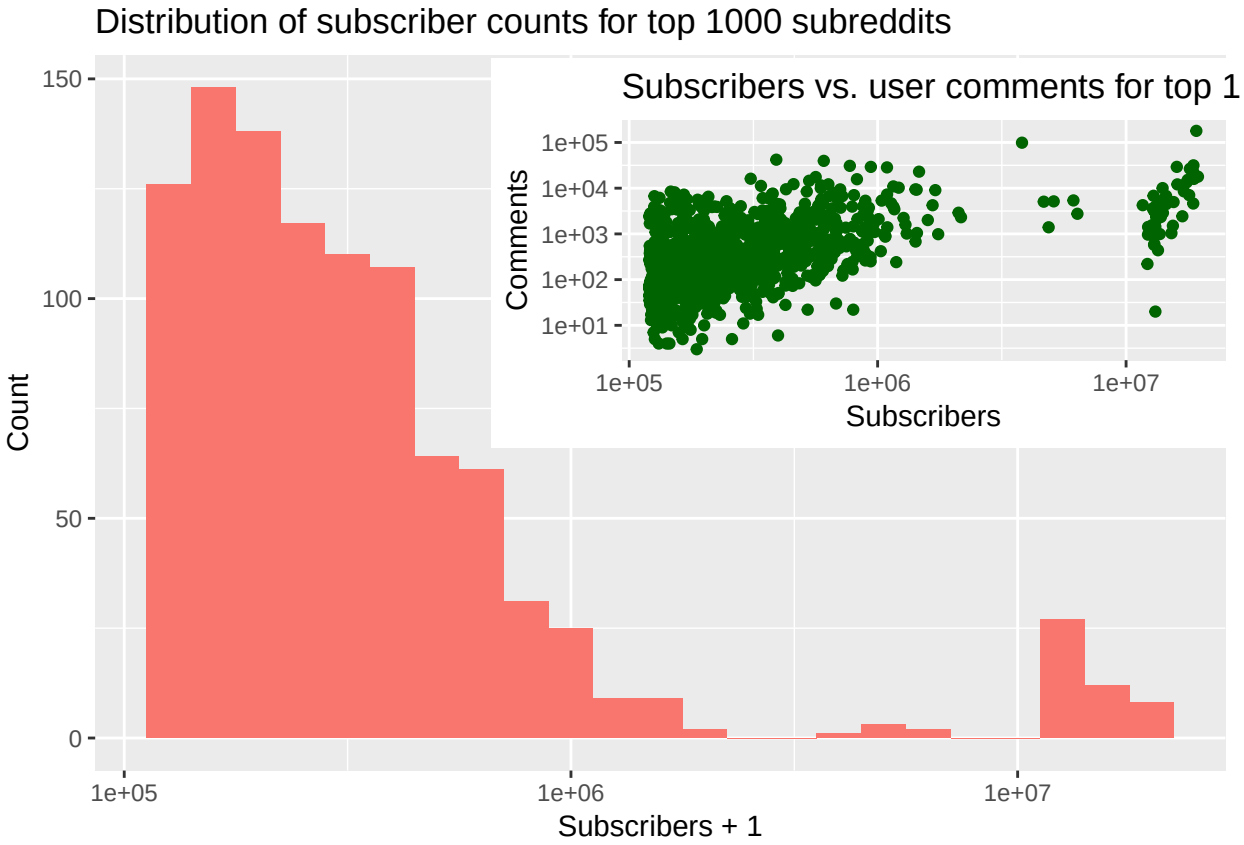
Grid for inset plot

Recall that from homework 8, we showed the following histogram plot of the distribution of subscriber counts for the top 1000 subreddits.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   120377   170467   262010  1045351   451114 19467413
```

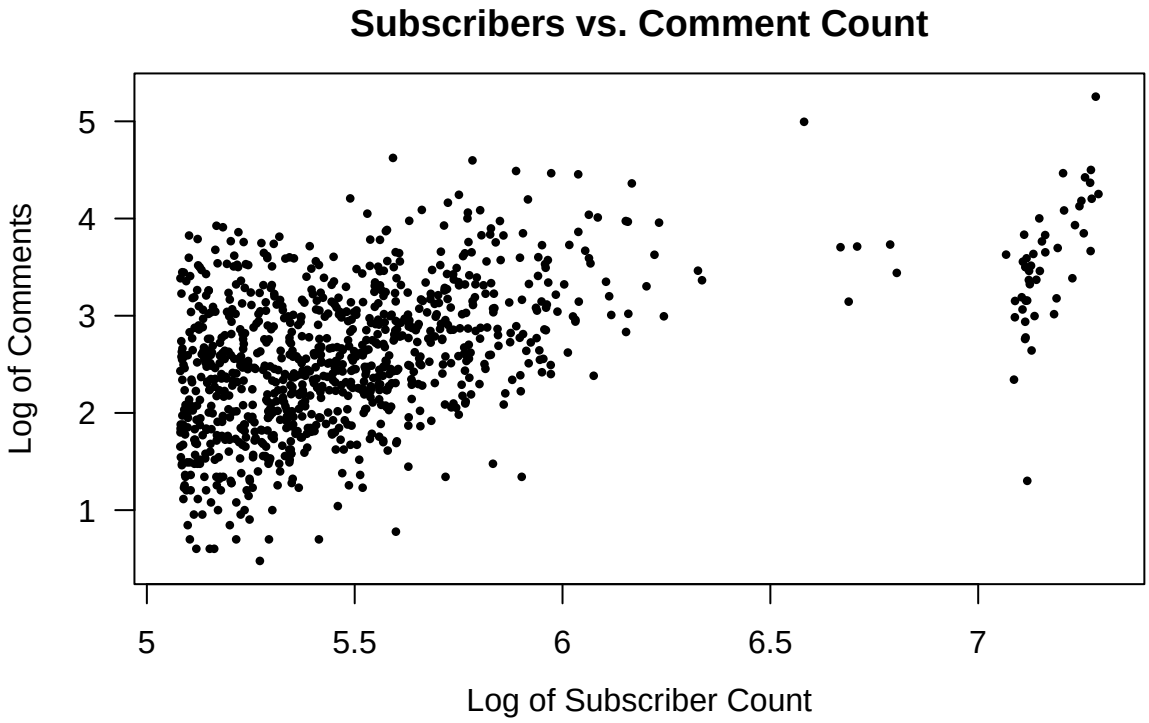


To update this plot, we decided to inset a scatterplot of the number of comments made on 12/01/2017 vs. the number of subscribers. We plotted this data with log scales because of the skewness of the data. The most interesting feature we observe here is that, as we previously discussed, there seems to be two groupings of subreddits, the regular ones and the top 100 subreddits that are the defaults.



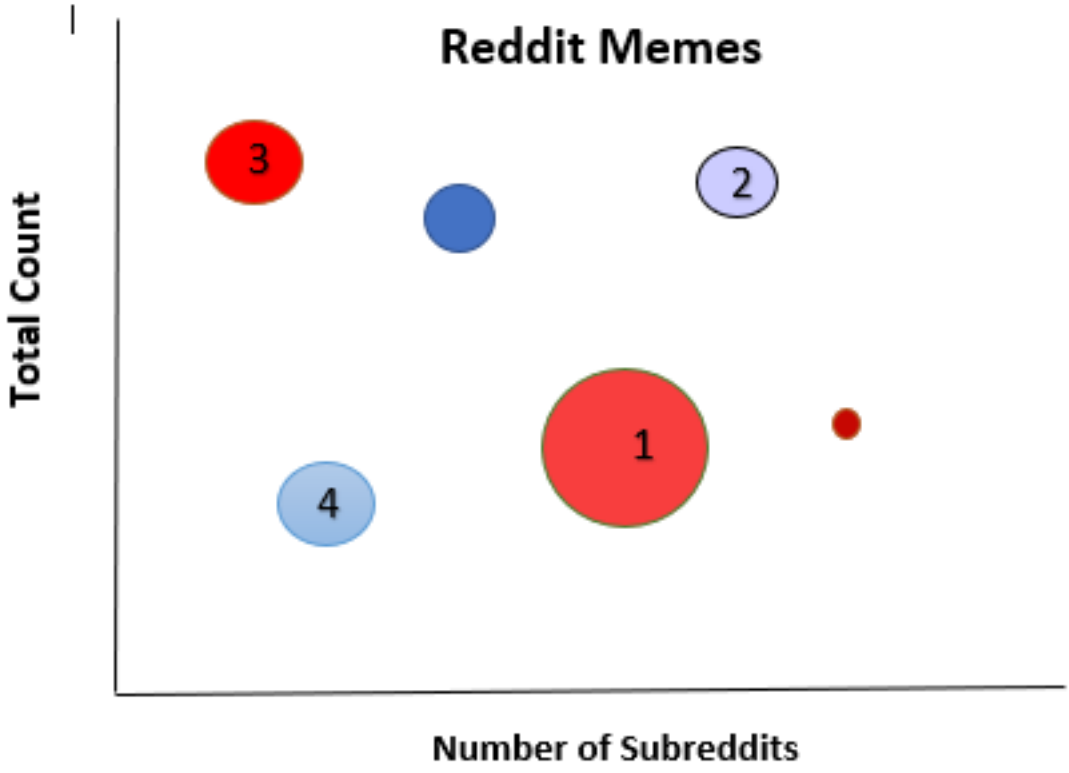
Grid for replicating plot

Here is a base Grid version of the new scatterplot we designed for the inset.



Killer plot proposal

Below you can see the basic mock-up of our killer plot. The main innovative feature is the color/size scaling to denote the length of the text of a meme, on top of the information presented by a regular scatterplot. We have an alternative in mind if this is too simple, but it seems to us to be a relatively simple and effective plot.



Total count of copypasta in y axis and total number of subreddits it is found in is in the x axis. The color of the circle will be the amount of text it has (gradient from blue to red). The size of the color will be the amount of upvotes it gets (standardized). Each copypasta will have a number and a corresponding key to the full text of the copypasta.