Twitter Text Analysis with @WeAreRLadies

Katherine Simeon

 @kitkatbar429

R-Ladies Chicago

What is a RoCur?

Rotating **Cur**ation or **Ro**tating **Cur**ator

Rotating the spokesperson on a social media account

Fun Fact: the first RoCur was **@Sweden**

Other cool RoCurs: ***@IAmSciComm, @Neurotweeps***

@WeAreRLadies

Every week, a different R-Lady takes over our twitter account.

They discuss their experiences:

how they use R

tips & tricks, favorite resources

questions & confusions


We are R-Ladies
@WeAreRLadies
Following
I ended my day by trying to install `RcppParallel` and IT WOULD NOT WORK - until I installed directly from the github repo 😅
This happens to me often, but instead of staying zen, I always go 😩 and apply various chaotic strategies of making things work.🔨
11:25 AM - 1 Oct 2018
7 Likes
4
7


Rose Martin @rose_m_martin · 15h
Replying to @WeAreRLadies
I'm happy to know I'm not alone!
1
Thea Knowles @theaknowles · 15h
Replying to @WeAreRLadies
CAN RELATE to chaotic attempts at solutions as my baseline 🤣
1
Fiorella Flores @FioreFloresC · Oct 1
Replying to @WeAreRLadies
It reminds me when I had to install the RVAideMemoire package... Lots of package dependencies 🙇🙇🙇
Fortunately, everything is ok now 👌😌
1
Seth Dobson @sethdobson · Oct 1
Replying to @WeAreRLadies
Good to know I'm not the only one.
1

Since August 2018...

13K+ Followers

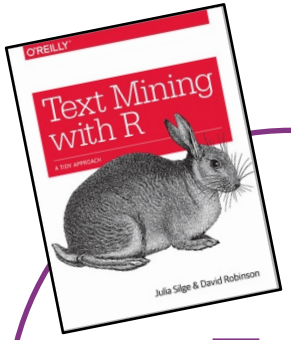
43 Awesome Curators

18 Countries represented

(**3 Chicago** R-Ladies!)

The Road Map

The Data



Text Analysis

Some Thoughts re: Data

The Data Set

Tweets from **August 2018** thru **June 2019**

Public RoCur Schedule (List of Curators)

Tweets

Info provided by Twitter

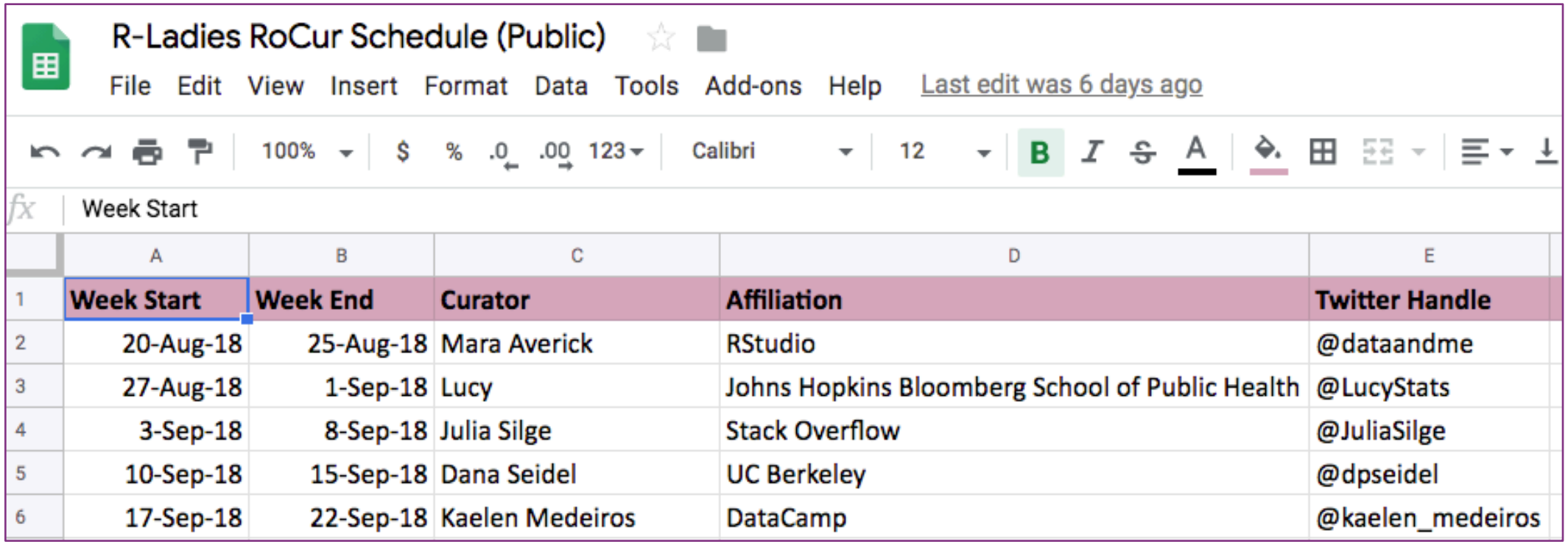
Tweet ID
Tweet URL
Tweet Text
Date & Time

Impressions
Engagements
- Engagement Rate
- Likes
- Retweets
- Expansions
- Clicks to link, media views
- Clicks to profile

RoCur Schedule

Manually-inputted public schedule

https://tinyurl.com/rladies-rocur-schedule



R-Ladies RoCur Schedule (Public)

File Edit View Insert Format Data Tools Add-ons Help Last edit was 6 days ago

100% $ % .0 .00 123 Calibri 12

fx Week Start

	A	B	C	D	E
1	**Week Start**	**Week End**	**Curator**	**Affiliation**	**Twitter Handle**
2	20-Aug-18	25-Aug-18	Mara Averick	RStudio	@dataandme
3	27-Aug-18	1-Sep-18	Lucy	Johns Hopkins Bloomberg School of Public Health	@LucyStats
4	3-Sep-18	8-Sep-18	Julia Silge	Stack Overflow	@JuliaSilge
5	10-Sep-18	15-Sep-18	Dana Seidel	UC Berkeley	@dpseidel
6	17-Sep-18	22-Sep-18	Kaelen Medeiros	DataCamp	@kaelen_medeiros

Prep Data for Combining

```
library(tidyverse)

# Two datasets
tweets <- read_csv("rocur_tweets.csv")
curators <- read_csv("curators.csv")

# Make sure dates are read as dates with lubridate
curators$Start <- dmy(curators$Start)
curators$End <- dmy(curators$End)
```

Placing Tweets to a Name

```
tweets_full <- tweets %>%
    mutate(id = "x") %>%
    left_join(curators %>% mutate(id = "x"), by = "id") %>%
    filter(Start <= date, End >= date)
```

```
# A tibble: 1,932 x 13
   Tweet.id Tweet.permalink     Tweet.text
      <dbl> <fct>               <fct>
 1  1.04e18 https://twitter.co… "👏, thanks for
 2  1.04e18 https://twitter.co… "@DataCamp @eamc
 3  1.04e18 https://twitter.co… "Would YOU like
 4  1.04e18 https://twitter.co… "🙌 @kellrstats
 5  1.04e18 https://twitter.co… "Already know so
 6  1.04e18 https://twitter.co… "✨ Want to buil
 7  1.04e18 https://twitter.co… "🎞 This present
 8  1.04e18 https://twitter.co… "☝ more spreads
```

```
# A tibble: 81,144 x 20
   Tweet.id Tweet.permalink Tweet.text
      <dbl> <fct>           <fct>
 1  1.04e18 https://twitte… "👏, thank…
 2  1.04e18 https://twitte… "👏, thank…
 3  1.04e18 https://twitte… "👏, thank…
 4  1.04e18 https://twitte… "👏, thank…
 5  1.04e18 https://twitte… "👏, thank…
 6  1.04e18 https://twitte… "👏, thank…
 7  1.04e18 https://twitte… "👏, thank…
 8  1.04e18 https://twitte… "👏, thank…
```

Placing Tweets to a Name

```
tweets_full <- tweets %>%
   mutate(id = "x") %>%
   left_join(curators %>% mutate(id = "x"), by = "id") %>%
   filter(Start <= date, End >= date)
```

Cleaning Data for Analysis

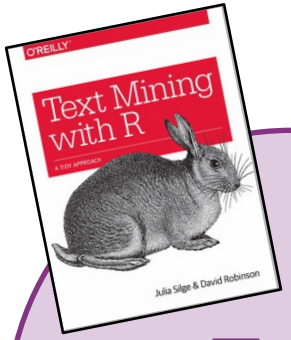
```
rocur_tweets <- tweets_full %>%
    select(Twitter,
           Tweet.text,
           date,
           time_only,
           engagement.rate,
           Student)
```

Our Data

	Twitter	Tweet.text	date	time_only	engagement.rate	Student
1	LucyStats	👋, thanks for a lovely week! It's been a delight to be …	2018-08-31	22:00:00	0.035996488	0
2	LucyStats	@DataCamp @eamcvey @daattali I didn't realize this …	2018-08-31	21:33:00	0.006329114	0
3	LucyStats	Would YOU like to be one of our #RLadies curators? W…	2018-08-31	21:30:00	0.027315023	0
4	LucyStats	🙌 @kellrstats has a delightful series of posts on the …	2018-08-31	21:00:00	0.011611275	0
5	LucyStats	Already know some #rstats Shiny basics? Try one of t…	2018-08-31	20:00:00	0.024627721	0
6	LucyStats	✨ Want to build interactive web apps using #rstats? …	2018-08-31	19:00:00	0.058688147	0
7	LucyStats	🎞 This presentation by @rctatman on data ethics is …	2018-08-31	18:02:00	0.028797696	0
8	LucyStats	☝ more spreadsheet reference! @kara_woo & @…	2018-08-31	17:00:00	0.027719298	0

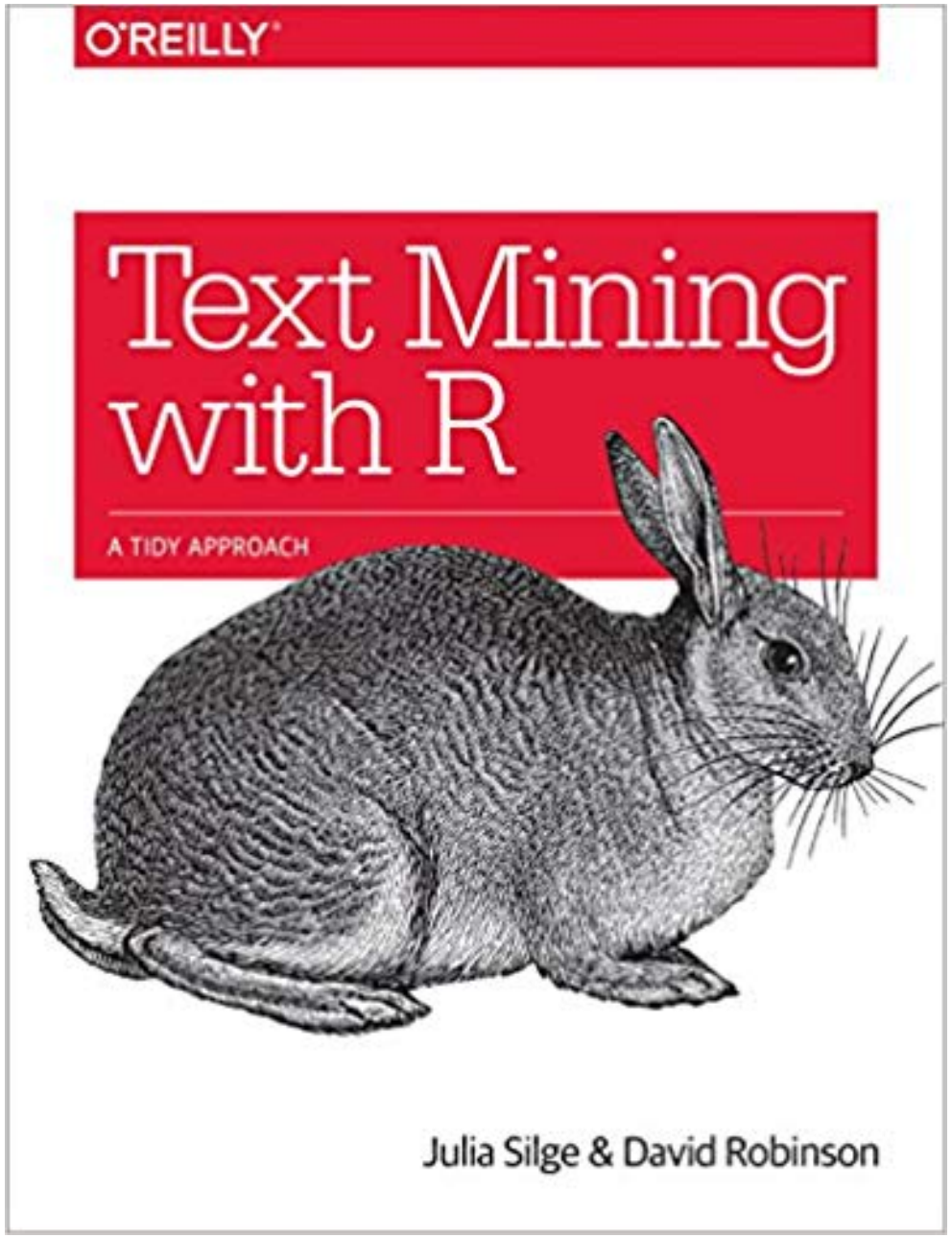
The Road Map

The Data



Text Analysis

Some Thoughts re: Data




```
library(tidytext)
```

← Some basics from this book

Publicly available at:
https://www.tidytextmining.com/

Text (language) is funky

There's a lot of variability!

There are **unreliable**, **inconsistent** cues

Dr. Smith prescribed the medicine.

Word meaning is **context-specific**

*The **rules** are confusing.*

*R **rules**!*

Tokenization

Splitting text into **tokens** → smaller meaningful units

```
rocur_tweets %>%
    unnest_tokens(word,Tweet.text)
```

- Filters punctuation
- Makes everything lowercase

Our Data

	Twitter	Tweet.text	date	time_only	engagement.rate	Student
1	LucyStats	👋, thanks for a lovely week! It's been a delight to be …	2018-08-31	22:00:00	0.035996488	0
2	LucyStats	@DataCamp @eamcvey @daattali I didn't realize this …	2018-08-31	21:33:00	0.006329114	0
3	LucyStats	Would YOU like to be one of our #RLadies curators? W…	2018-08-31	21:30:00	0.027315023	0
4	LucyStats	🙌 @kellrstats has a delightful series of posts on the …	2018-08-31	21:00:00	0.011611275	0
5	LucyStats	Already know some #rstats Shiny basics? Try one of t…	2018-08-31	20:00:00	0.024627721	0
6	LucyStats	✨ Want to build interactive web apps using #rstats? …	2018-08-31	19:00:00	0.058688147	0
7	LucyStats	🎞 This presentation by @rctatman on data ethics is …	2018-08-31	18:02:00	0.028797696	0
8	LucyStats	☝ more spreadsheet reference! @kara_woo & @…	2018-08-31	17:00:00	0.027719298	0

1,865 rows

Our Data – Tokenized!

	Twitter	date	time_only	engagement.rate	Student	word
1	LucyStats	2018-08-31	22:00:00	0.035996488	0	thanks
2	LucyStats	2018-08-31	22:00:00	0.035996488	0	for
3	LucyStats	2018-08-31	22:00:00	0.035996488	0	a
4	LucyStats	2018-08-31	22:00:00	0.035996488	0	lovely
5	LucyStats	2018-08-31	22:00:00	0.035996488	0	week
6	LucyStats	2018-08-31	22:00:00	0.035996488	0	it's
7	LucyStats	2018-08-31	22:00:00	0.035996488	0	been
8	LucyStats	2018-08-31	22:00:00	0.035996488	0	a

54,230 rows

Most Common Words

```
rocur_tweets %>%
    unnest_tokens(word,Tweet.text) %>%
    count(word, sort = TRUE)
```

*Many of these are **stop words***

```
# A tibble: 7,975 x 2
   word      n
   <chr> <int>
 1 to     1538
 2 the    1508
 3 t.co   1486
 4 https  1485
 5 i      1197
 6 a      1126
 7 and    1055
 8 of      797
 9 in      784
10 you     741
# … with 7,965 more rows
```

Get rid of stop words

```
words <- rocur_tweets %>%
    unnest_tokens(word,Tweet.text)

data("stop_words") # data set from tidytext

words <- words %>%
    anti_join(stop_words) %>%
    filter(!word %in% c('t.co', 'https'))
```

More informative set of words?

```
words %>%
    count(word, sort = TRUE)
```

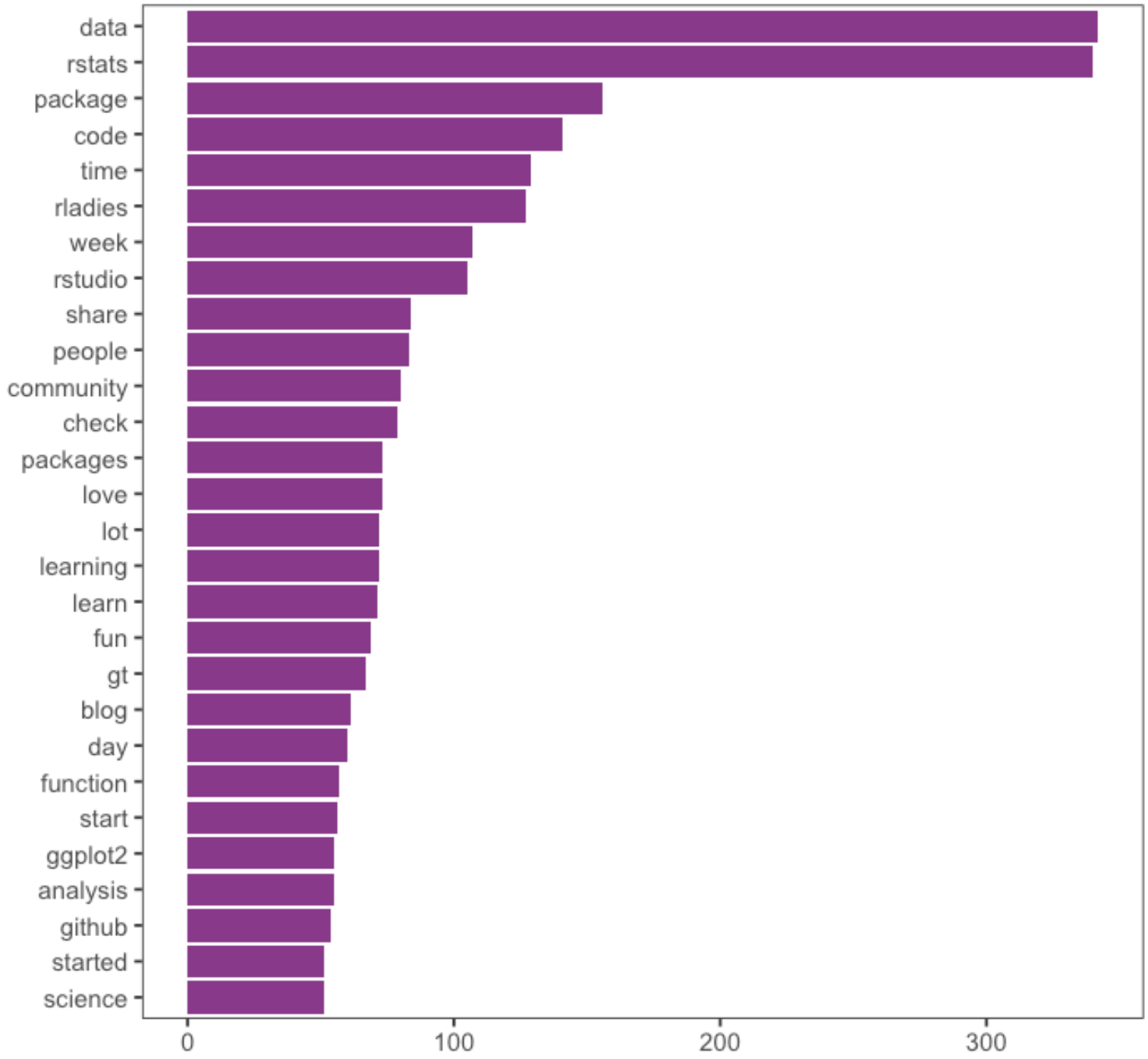
Before

```
# A tibble: 7,975 x 2
   word      n
   <chr> <int>
 1 to     1538
 2 the    1508
 3 t.co   1486
 4 https  1485
 5 i      1197
 6 a      1126
 7 and    1055
 8 of      797
 9 in      784
10 you     741
# … with 7,965 more rows
```

After

```
# A tibble: 7,394 x 2
   word        n
   <chr>   <int>
 1 data      342
 2 rstats    340
 3 package   156
 4 code      141
 5 time      129
 6 rladies   127
 7 week      107
 8 rstudio   105
 9 share      84
10 people     83
# … with 7,384 more rows
```

Word Frequency



Sentiment Analysis

Identify the emotional intent of text

→ Is it positive, negative, neutral?

One way to do this:

- Determine the sentiment of individual words using a sentiment lexicons
- Sum the sentiment of individual words in a given text.

get_sentiments()

Using the **Bing** sentiment lexicon

→ Binary positive/negative classification

→ `get_sentiments()` in tidytext

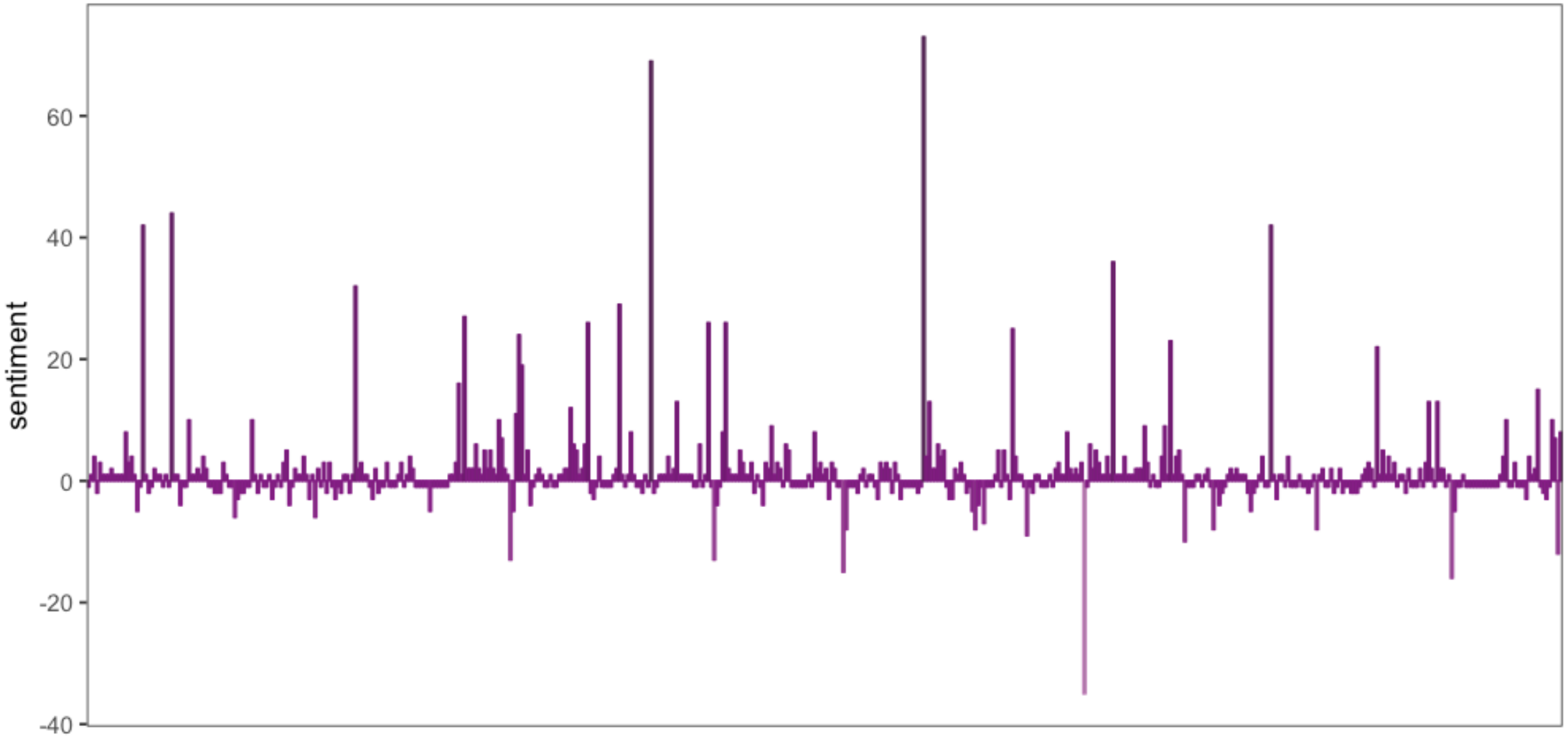
Sentiment of @WeAreRLadies

```
words %>%
    inner_join(get_sentiments("bing")) %>%
    count(word, sentiment, sort = TRUE) %>%
    spread(sentiment, n, fill = 0) %>%
    mutate(sentiment = positive - negative)
```

```
# A tibble: 514 x 3
   word    sentiment     n
   <chr>   <chr>     <int>
 1 love    positive     73
 2 fun     positive     69
 3 awesome positive     44
 4 amazing positive     42
```

```
# A tibble: 514 x 4
   word        negative positive sentiment
   <chr>          <dbl>    <dbl>     <dbl>
 1 aborts             1        0        -1
 2 abundance          0        1         1
 3 accessible         0        4         4
 4 accidental         2        0        -2
 5 accomplish         0        3         3
```

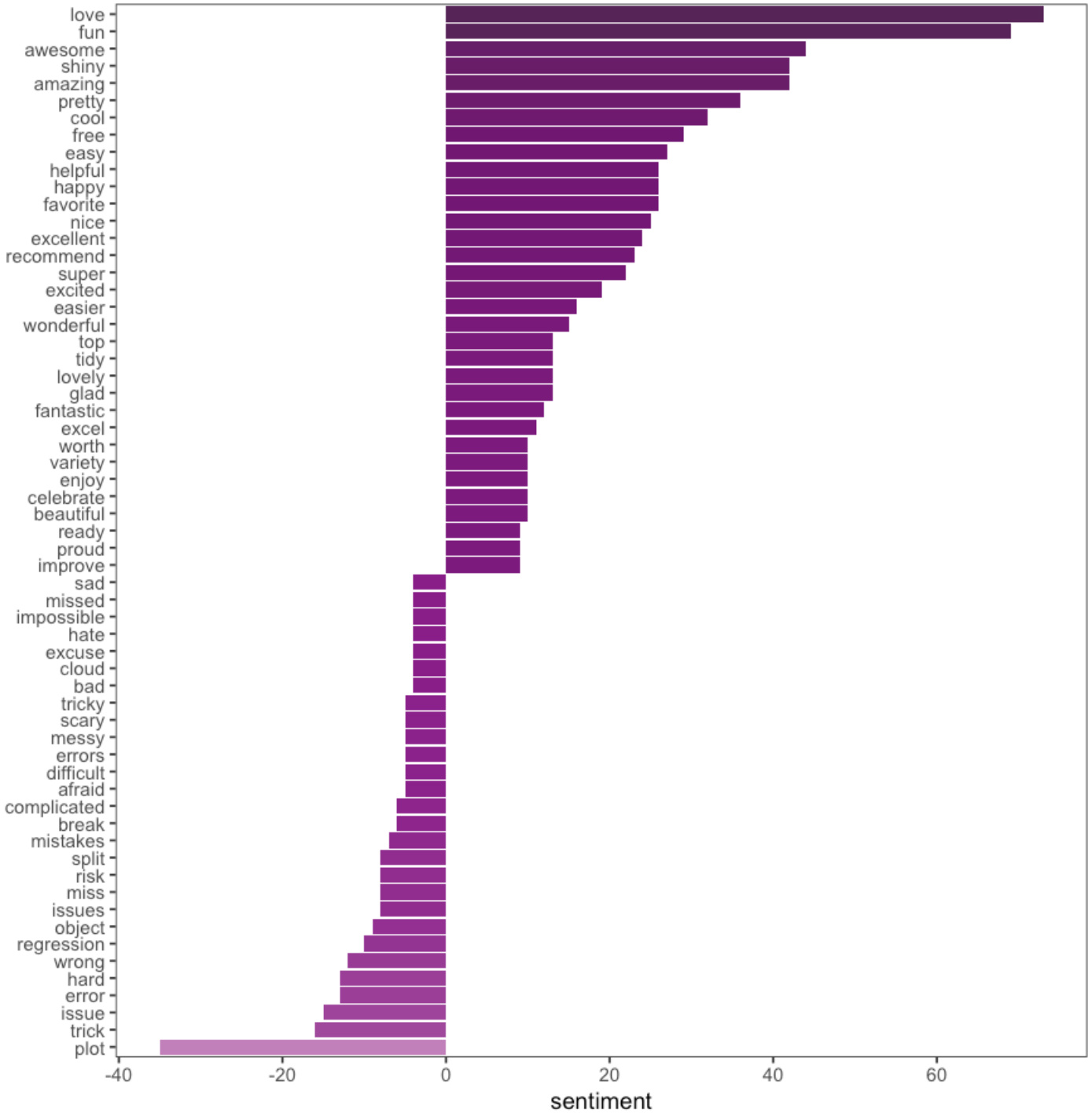
Sentiment of @WeAreRLadies

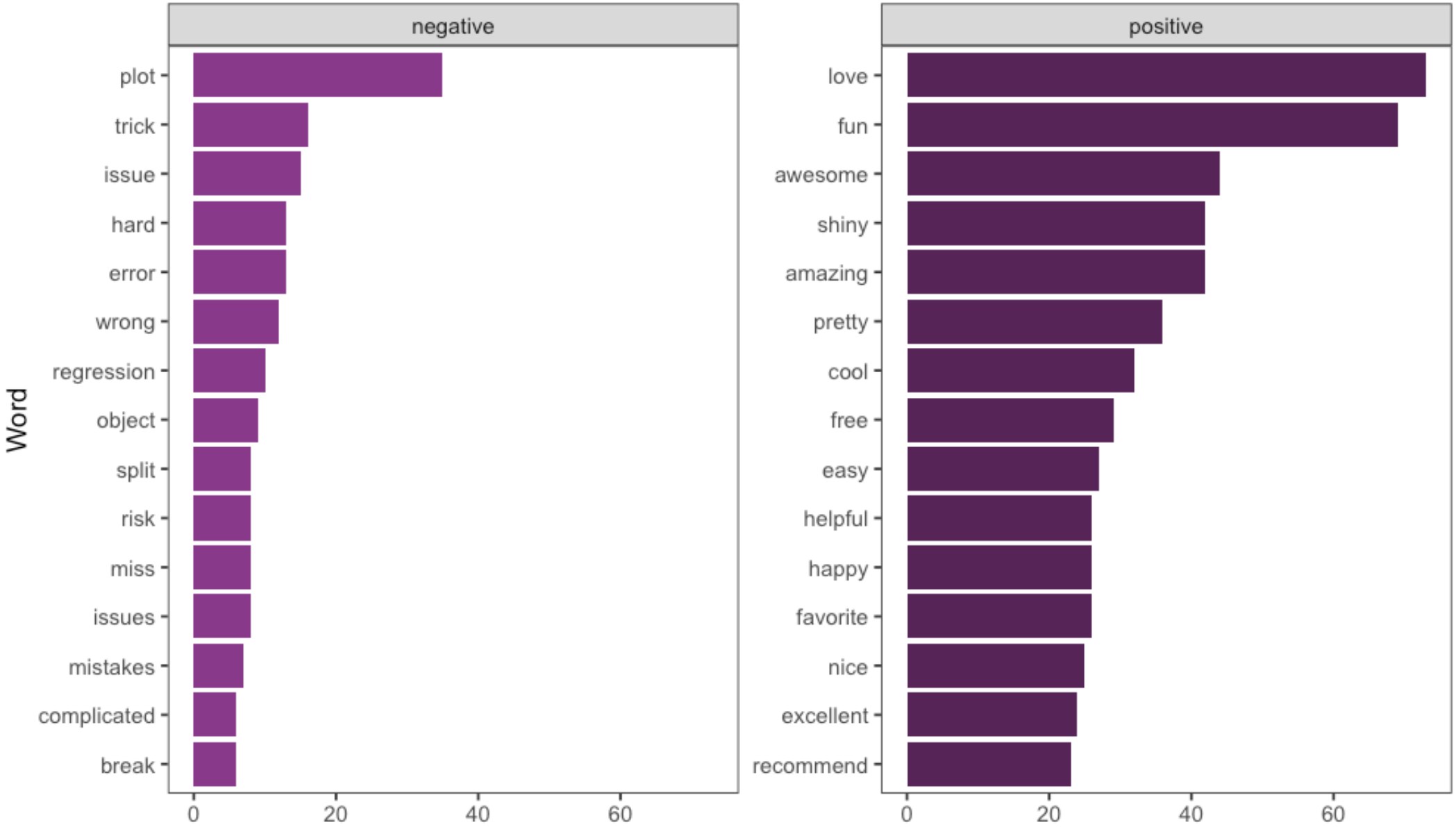


Look at the strong sentiments

```
words %>%
      inner_join(get_sentiments("bing")) %>%
      count(word, sentiment, sort = TRUE) %>%
      spread(sentiment, n, fill = 0) %>%
      mutate(sentiment = positive - negative) %>%
      filter(sentiment < -3 | sentiment > 8) %>%
      mutate(word = reorder(word, sentiment))
```

More Readable
(sort of)




negative
positive
Word
plot
trick
issue
hard
error
wrong
regression
object
split
risk
miss
issues
mistakes
complicated
break
love
fun
awesome
shiny
amazing
pretty
cool
free
easy
helpful
happy
favorite
nice
excellent
recommend
0
20
40
60

N-grams for context

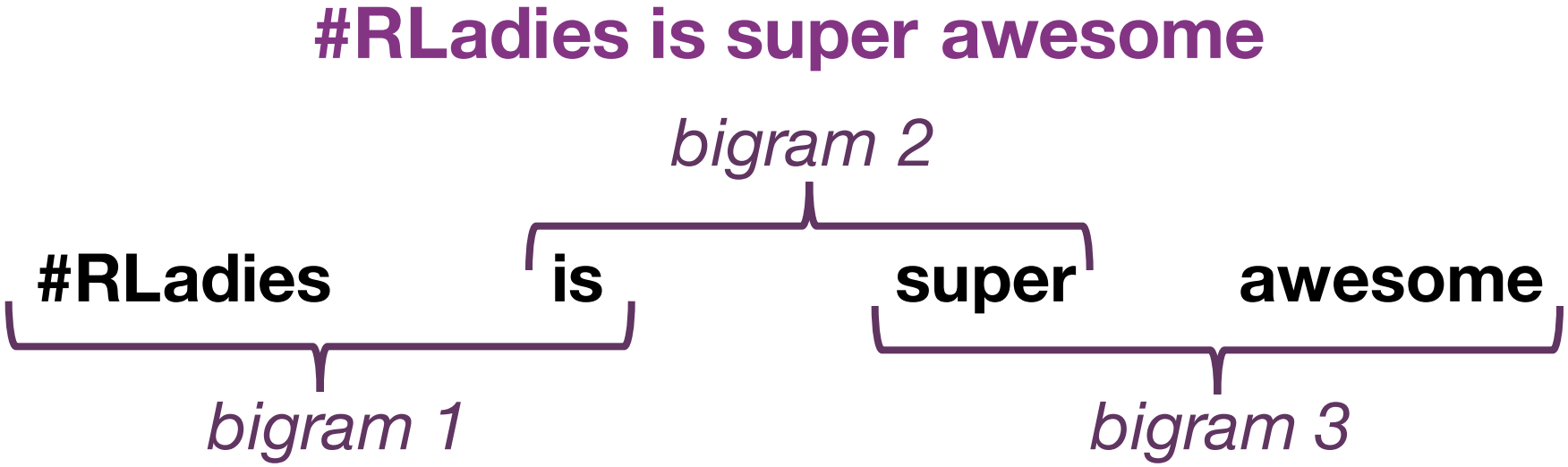
Tokenizing by n-grams, or word sequences.

#RLadies is super awesome

#RLadies	is	super	awesome
Word 1	*Word 2*	*Word 3*	*Word 4*

N-grams for context

Tokenizing by n-grams, or word sequences.



Obtaining bigrams from our data

```
rocur_tweets %>%
    unnest_tokens(bigram,Tweet.text,
                  token = "ngrams", n = 2)
```

```
# A tibble: 52,461 x 6
   Twitter date       time_only engagement.rate Student bigram
   <chr>   <date>     <time>                <dbl>   <dbl> <chr>
 1 @162880 2019-06-10 11:42                0.0267       0 hi everyone
 2 @162880 2019-06-10 11:42                0.0267       0 everyone i'm
 3 @162880 2019-06-10 11:42                0.0267       0 i'm margaret
```

Most common bigrams

```
rocur_tweets %>%
     unnest_tokens(bigram,Tweet.text,
                   token = "ngrams", n = 2) %>%
     count(bigram, sort = TRUE)
```

```
# A tibble: 33,337 x 2
   bigram          n
   <chr>       <int>
 1 https t.co   1485
 2 of the        130
 3 you can       126
 4 in the        115
 5 if you         93
 6 for the        85
 7 in r           84
 8 is a           79
 9 i have         72
10 a lot          70
# … with 33,327 more rows
```

Clean the data

Step 1 – Separate bigrams

```
bigrams <- rocur_tweets %>%
    unnest_tokens(bigram,Tweet.text,
                  token = “ngrams”, n = 2)

bigrams_separated <- bigrams %>%
    separate(bigram,
             c("word1", "word2"), sep = " ")
```

Clean the data

Step 2 – Filter out stop words

```
bigrams_filtered <- bigrams_separated %>%
     filter(!word1 %in% stop_words$word) %>%
     filter(!word2 %in% stop_words$word)
```

More meaningful bigrams?

```
bigrams_filtered %>%
    count(word1, word2, sort = TRUE)
```

Before

```
# A tibble: 33,337 x 2
   bigram          n
   <chr>       <int>
 1 https t.co   1485
 2 of the        130
 3 you can       126
 4 in the        115
 5 if you         93
 6 for the        85
 7 in r           84
 8 is a           79
 9 i have         72
10 a lot          70
# … with 33,327 more rows
```

After

```
# A tibble: 6,625 x 3
   word1      word2          n
   <chr>      <chr>      <int>
 1 data       science       35
 2 rstats     community     21
 3 blog       post          17
 4 data       table         17
 5 data       scientist     11
 6 rstats     packages      11
 7 dplyr      trick         10
 8 jennybryan statquant     10
 9 swmpkim    jennybryan    10
10 data       analysis       9
# … with 6,615 more rows
```

plot

`filter(word1 == "plot")`

```
# A tibble: 26 x 3
   word1 word2          n
   <chr> <chr>      <int>
 1 plot  https          4
 2 plot  amp            2
 3 plot  can            2
 4 plot  is             2
 5 plot  an             1
 6 plot  and            1
 7 plot  anonymize      1
 8 plot  below          1
 9 plot  confirm        1
10 plot  conveys        1
# … with 16 more rows
```

`filter(word2 == "plot")`

```
# A tibble: 24 x 3
   word1   word2     n
   <chr>   <chr> <int>
 1 your    plot      4
 2 a       plot      3
 3 sad     plot      3
 4 the     plot      3
 5 to      plot      3
 6 bar     plot      1
 7 box     plot      1
 8 can     plot      1
 9 could   plot      1
10 density plot      1
# … with 14 more rows
```

Other “negative” words

trick

```
# A tibble: 6 x 3
  word1    word2      n
  <chr>    <chr>  <int>
1 dplyr    trick     10
2 knitr    trick      2
3 last     trick      1
4 rprofile trick      1
5 that     trick      1
6 the      trick      1
```

object

```
# A tibble: 7 x 3
  word1   word2      n
  <chr>   <chr>  <int>
1 spatial object      3
2 3d      object      1
3 an      object      1
4 ggplot2 object      1
5 lt      object      1
6 the     object      1
7 with    object      1
```

regression

```
# A tibble: 9 x 3
  word1      word2          n
  <chr>      <chr>      <int>
1 regression models         2
2 regression analysis       1
3 regression as             1
4 regression data           1
5 regression image          1
6 regression ml             1
7 regression model          1
8 regression output         1
9 regression problems       1
```

tf-idf

Frequency of a term adjusted for how often it is used.

#rstats

✅

Good hashtag to search the entirety of Twitter

❓

Not helpful for searching @WeAreRLadies

Obtaining *tf-idf*

The product of…

term frequency (*tf*)

How often a given word occurs

inverse document frequency (*idf*)

Weighting:

⬇ words used a lot; ⬆ words used less frequently

tf-idf for @WeAreRLadies

```
# Get the word count for each curator
word_count_curator <- words %>%
    count(Twitter, word, sort = TRUE)

# Get term frequency
total_words_curator <- word_count_curator %>%
    group_by(Twitter) %>%
    summarize(total = sum(n))

curator_tf <- left_join(word_count_curator,
                        total_words)
```

tf-idf for @WeAreRLadies

```
# Get tf-idf
curator_tf_idf <- curator_tf %>%
    bind_tf_idf(word, Twitter, n)
```

```
> curator_tf
```

```
# A tibble: 14,686 x 7
   Twitter         word          n total
   <chr>           <chr>     <int> <int>
 1 @MaryELennon    rladies      59  1301
 2 @EmmaVitz       data         50  1102
 3 @littlemissdata rstats       42  1761
```

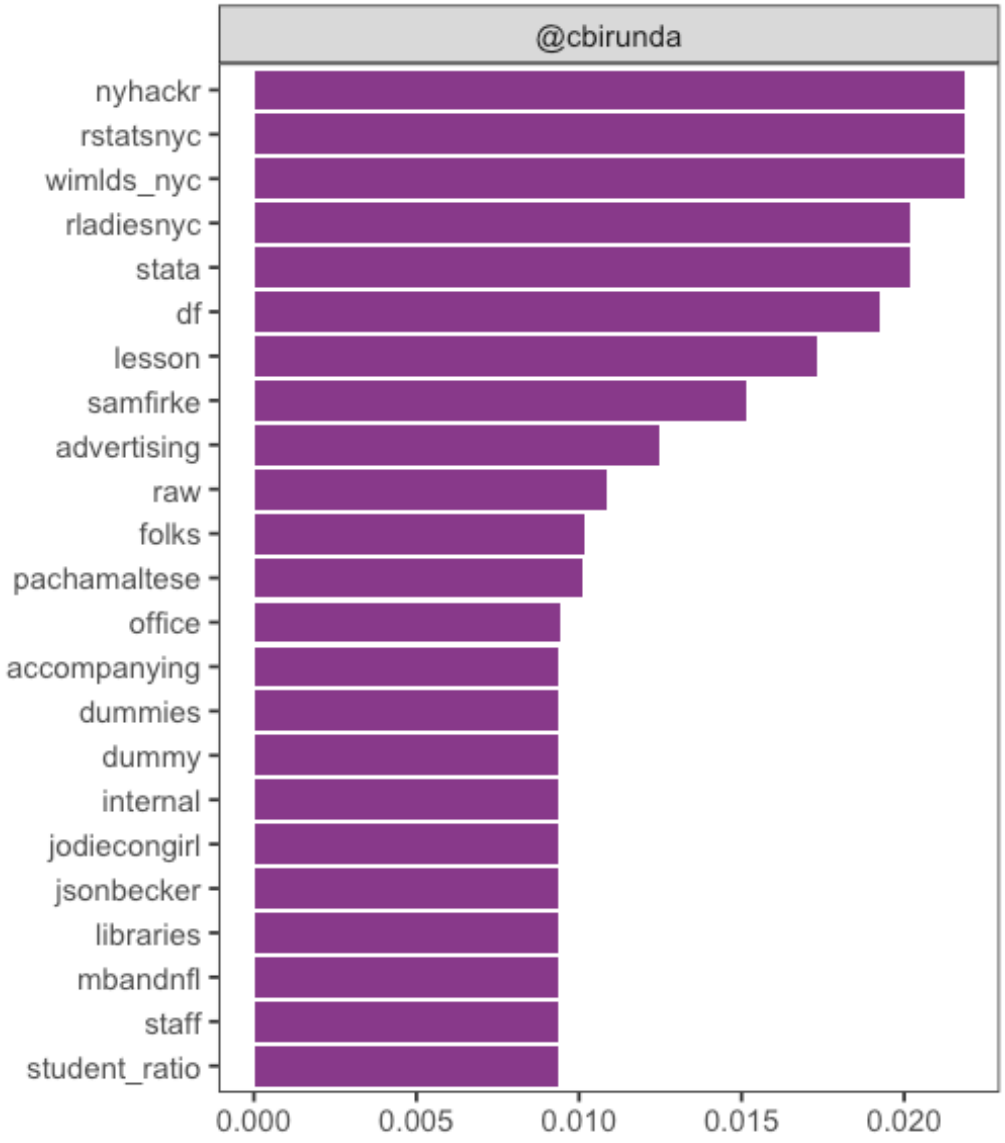
tf-idf for @WeAreRLadies

```
# Get tf-idf
curator_tf_idf <- curator_tf %>%
    bind_tf_idf(word, Twitter, n)
```

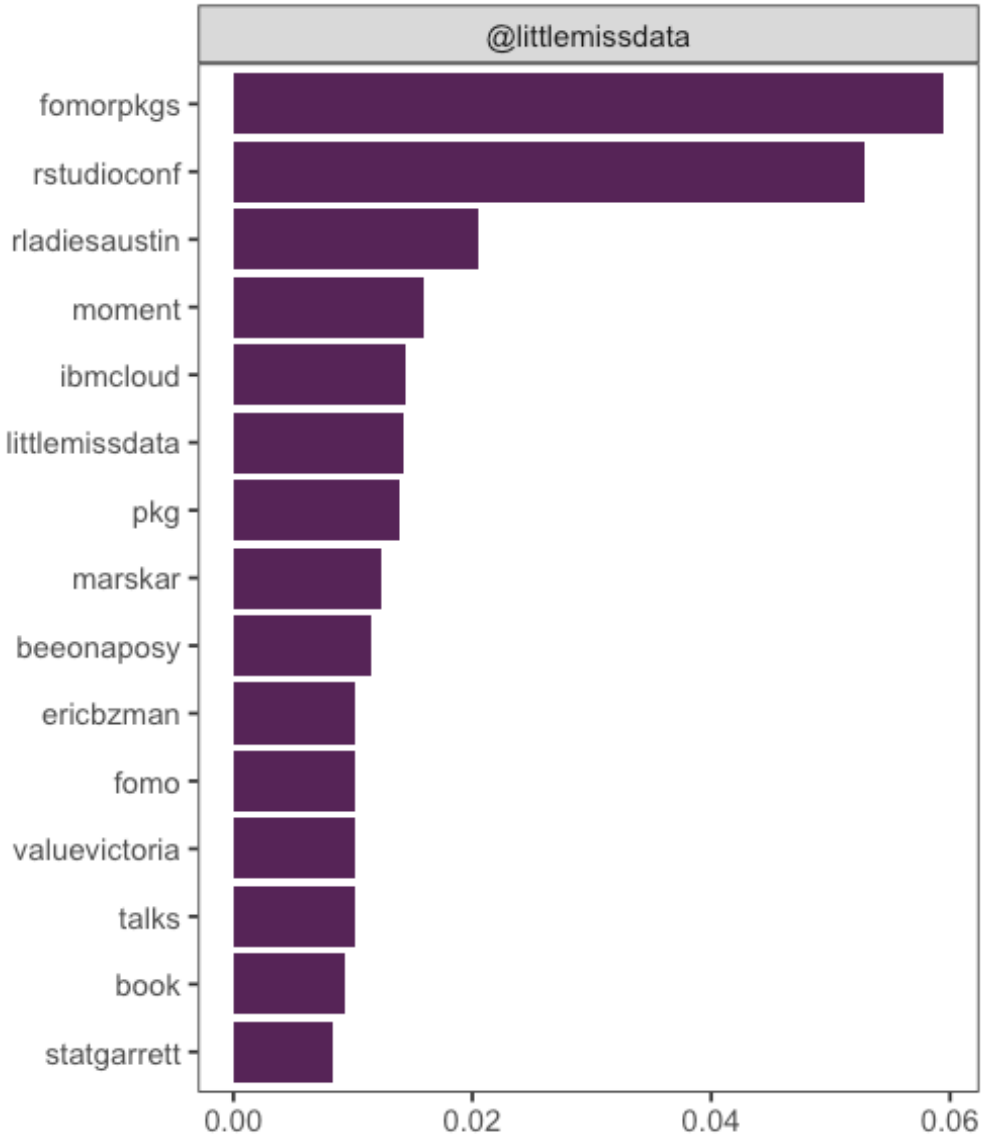
```
> curator_tf_idf
```

```
# A tibble: 14,686 x 7
   Twitter         word          n total     tf    idf   tf_idf
   <chr>           <chr>     <int> <int>  <dbl>  <dbl>    <dbl>
 1 @MaryELennon    rladies      59  1301 0.0453 0.392  0.0178
 2 @EmmaVitz       data         50  1102 0.0454 0.0274 0.00124
 3 @littlemissdata rstats       42  1761 0.0239 0.210  0.00500
```

NYC R Conference



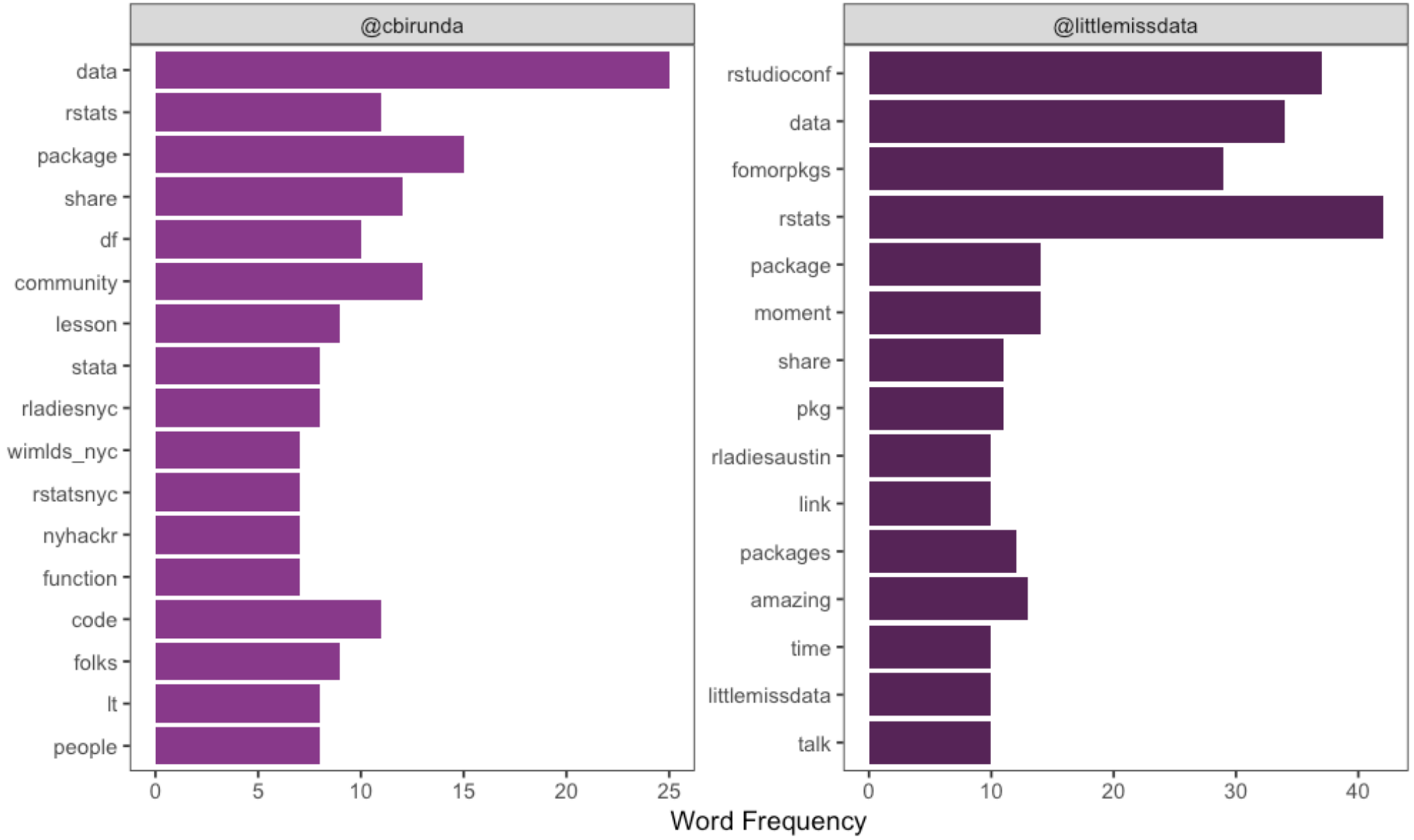
RStudio::conf



td-idf

NYC R Conference

RStudio::conf



Text (language) is funky

There's a lot of variability!

There are **unreliable**, **inconsistent** cues

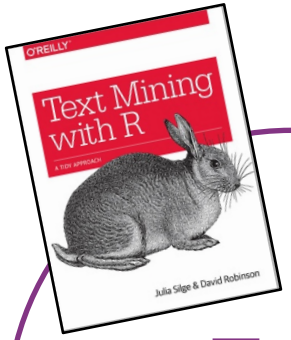
We can **tokenize**

Word meaning is **context-specific**

We can use **n-grams, tf-idf**

The Road Map

The Data



Text Analysis

Some Thoughts re: Data

Differences in...

Curators

- Students and Professionals
  →Data Scientists, Consultants, Academics, etc.

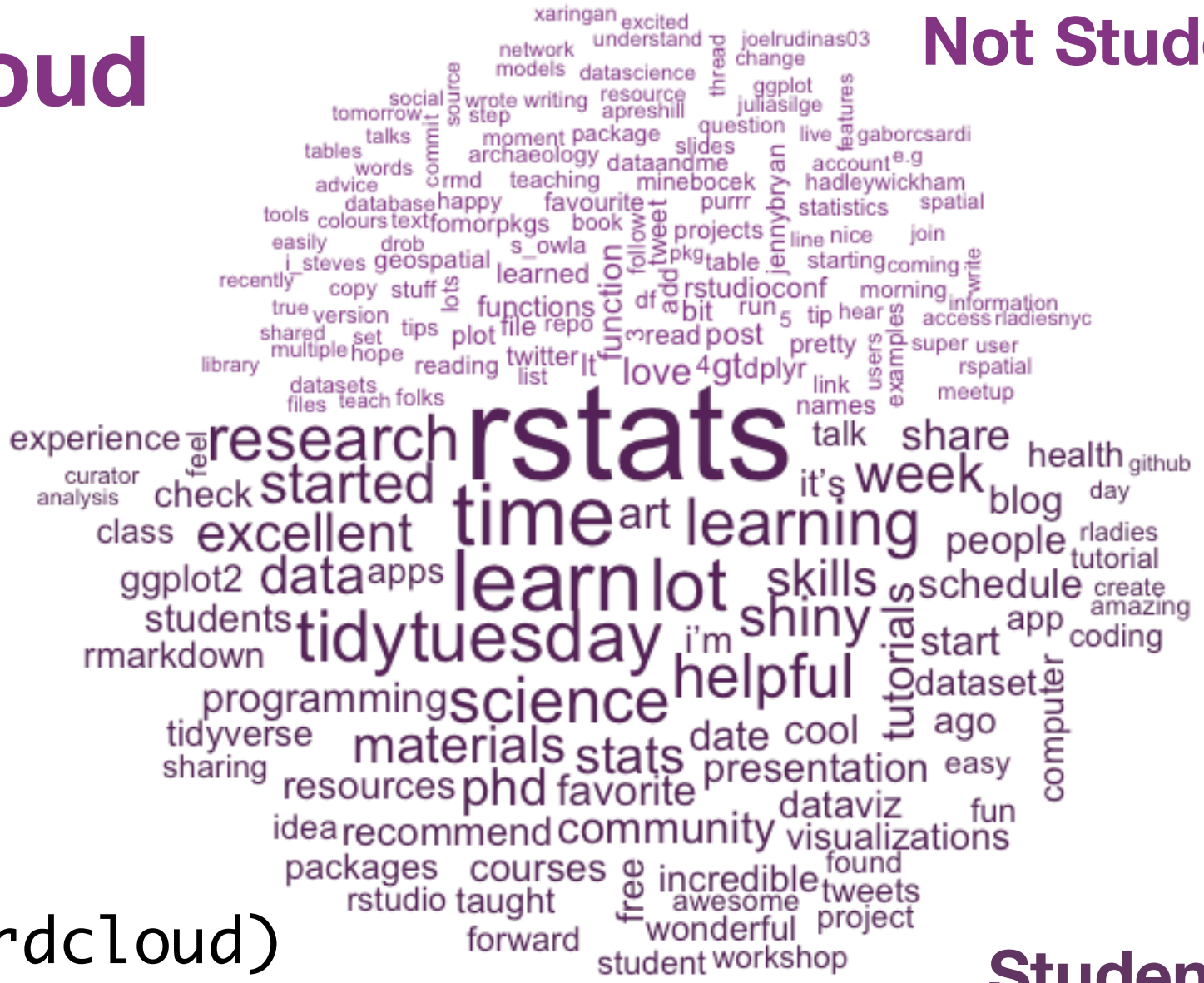
Experiences

- Multiple use cases
- Methods for learning
- Varying levels of proficiency in both R and Twitter

Intent of Tweet

- Sharing resources vs. asking a question

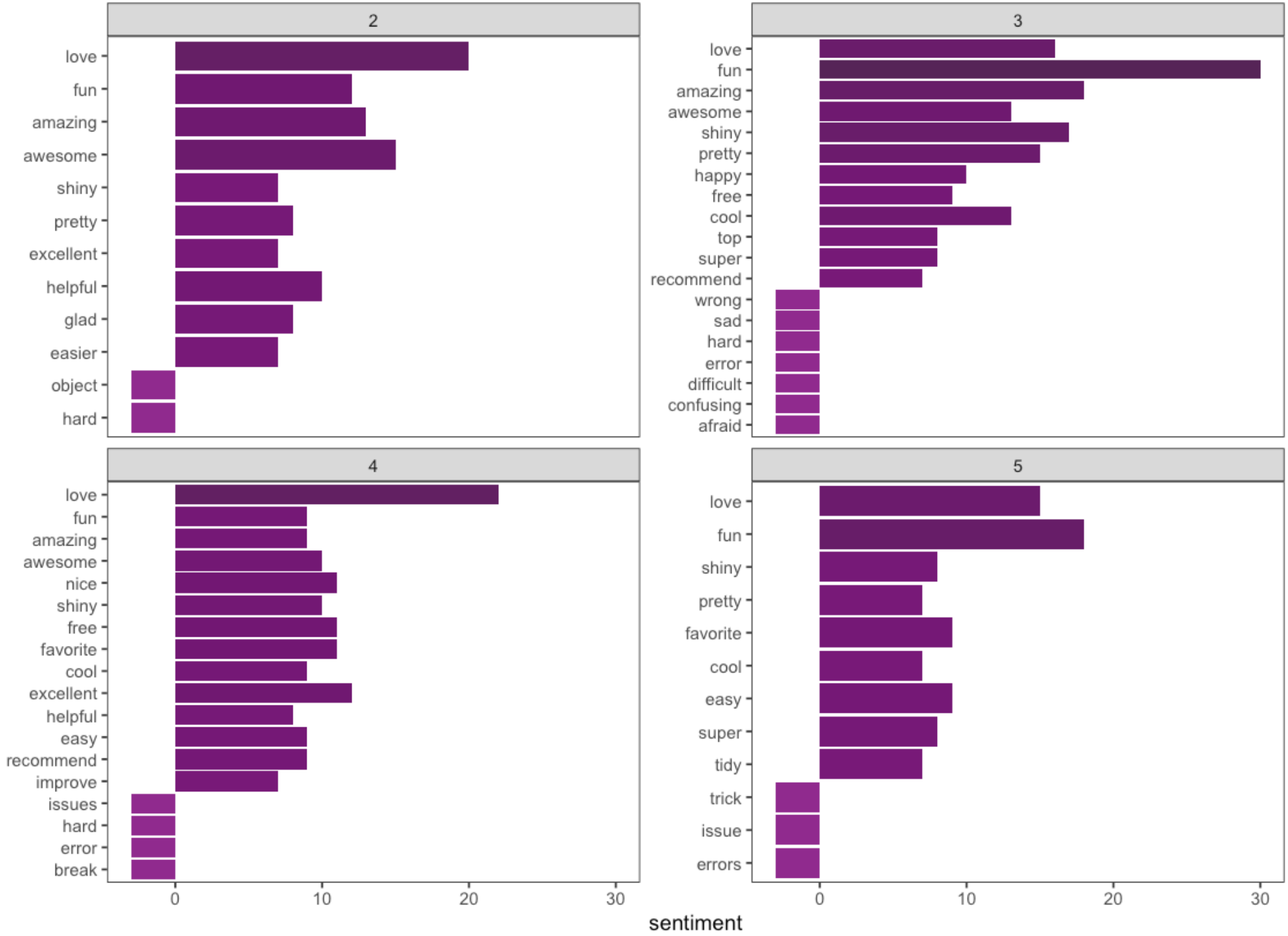
Word Cloud



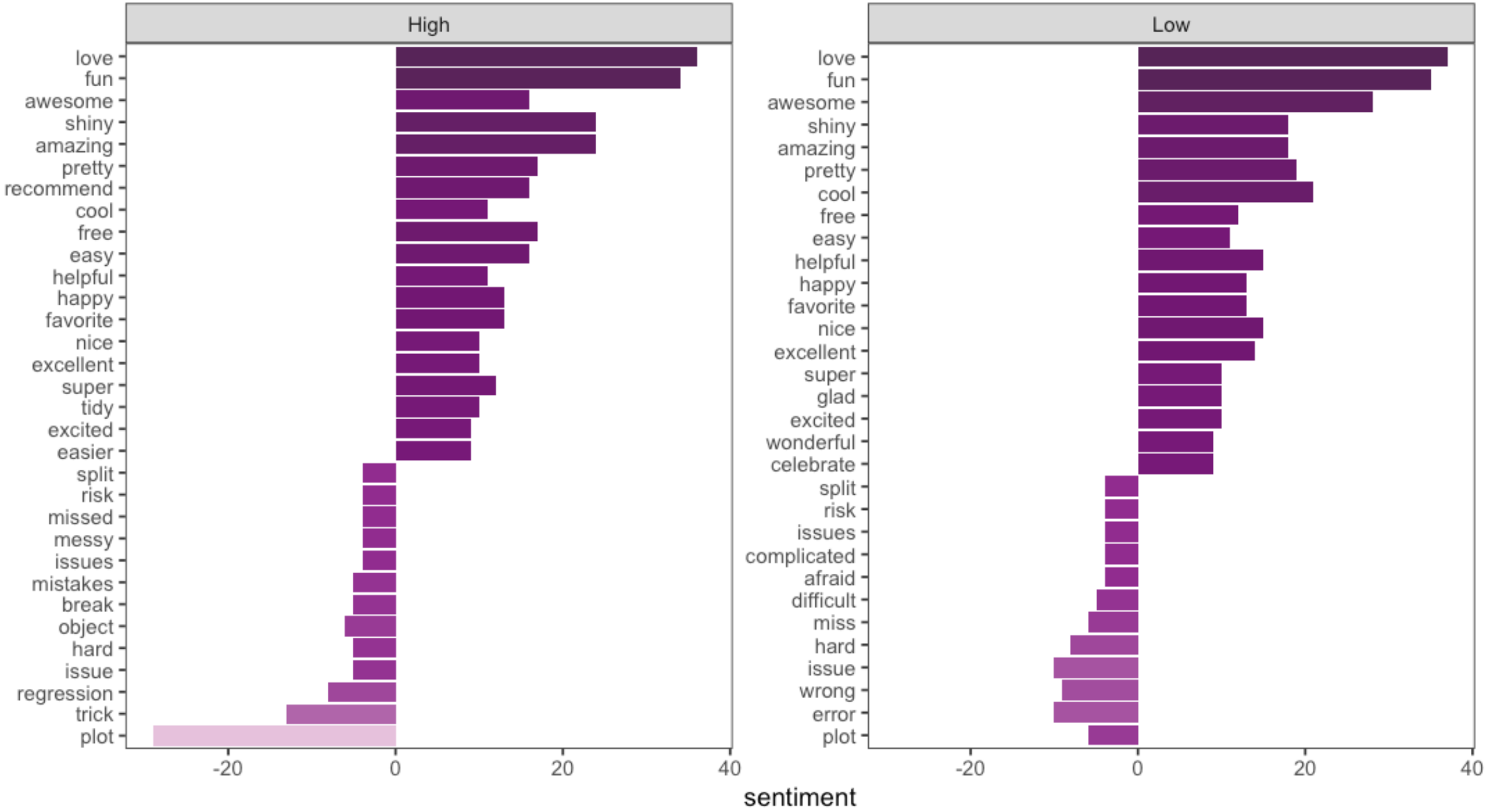
Not Students

Students

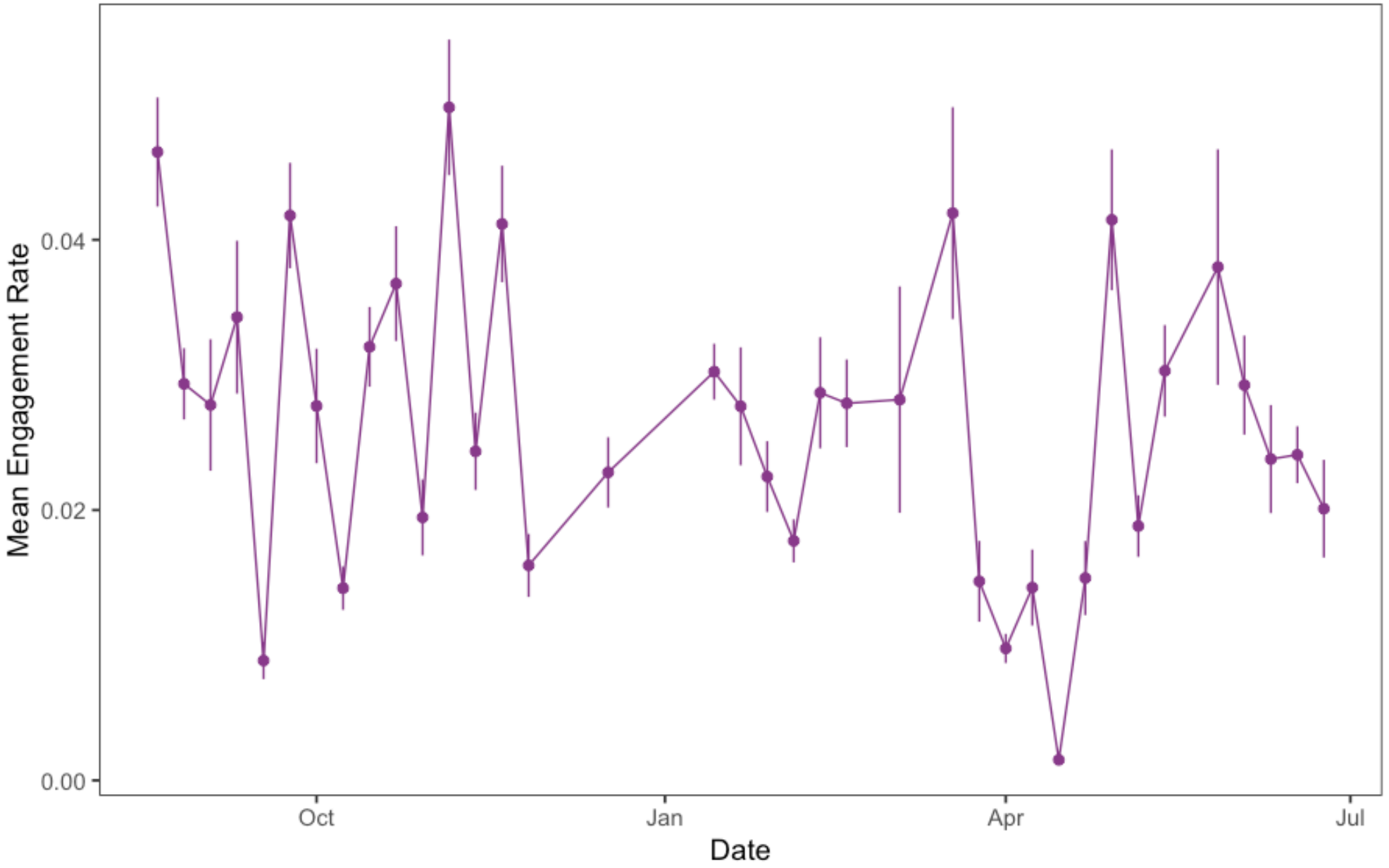
```
library(wordcloud)
```


2
love
fun
amazing
awesome
shiny
pretty
excellent
helpful
glad
easier
object
hard
3
love
fun
amazing
awesome
shiny
pretty
happy
free
cool
top
super
recommend
wrong
sad
hard
error
difficult
confusing
afraid
4
love
fun
amazing
awesome
nice
shiny
free
favorite
cool
excellent
helpful
easy
recommend
improve
issues
hard
error
break
5
love
fun
shiny
pretty
favorite
cool
easy
super
tidy
trick
issue
errors
0
10
20
30
sentiment

Sentiment: High vs. Low Engagement Rate



Engagement Rate Over Time



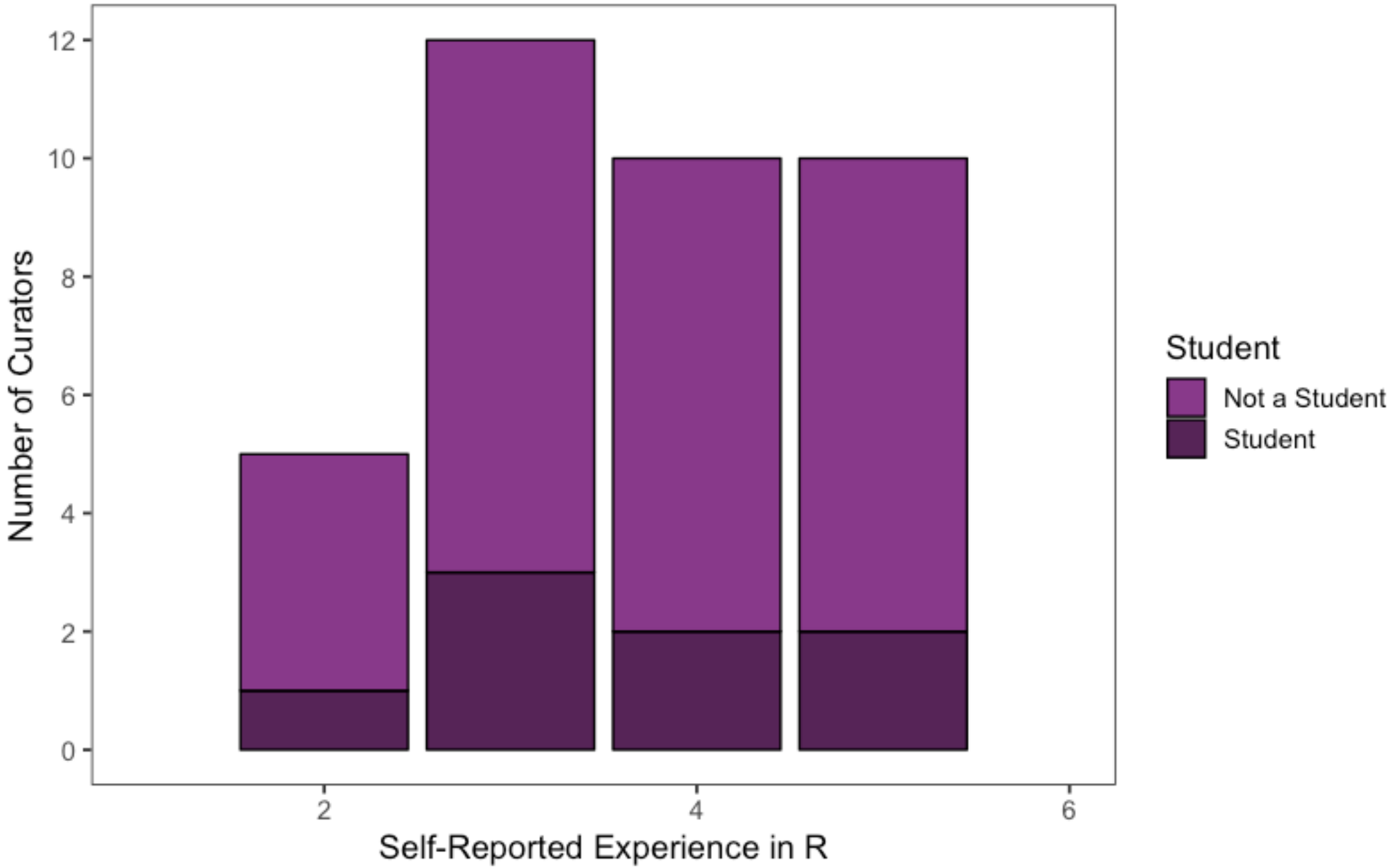
Call to Action

Consider curating for **@WeAreRLadies!**

We are all learning **together**

Everyone's perspective is **valuable**

https://github.com/rladies/starter-kit/tree/master/RoCur-Twitter


Number of Curators
12
10
8
6
4
2
0
2
4
6
Self-Reported Experience in R
Student
Not a Student
Student

Acknowledgments

Lucy D'Agostino McGowan (R-Ladies Nashville)

Janani Ravi (R-Ladies East Lansing)

Nujcharee Haswell (North Yorkshire, UK)

Sush Gopalan (R-Ladies Chicago)

R-Ladies Global

- Maëlle Salmon
- Gabriela de Queiroz

Thank you!

github.com/katherinesimeon