

Text Analysis of Correlaid

Longhao Chen/Qianhui Rong/Wenjia Xie/Andrew Zhang

11/3/2018

Seperate Analysis on Each Article ## P-Value Article We want to analyze the passage from <https://correlaid.org/blog/posts/understand-p-values>.

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

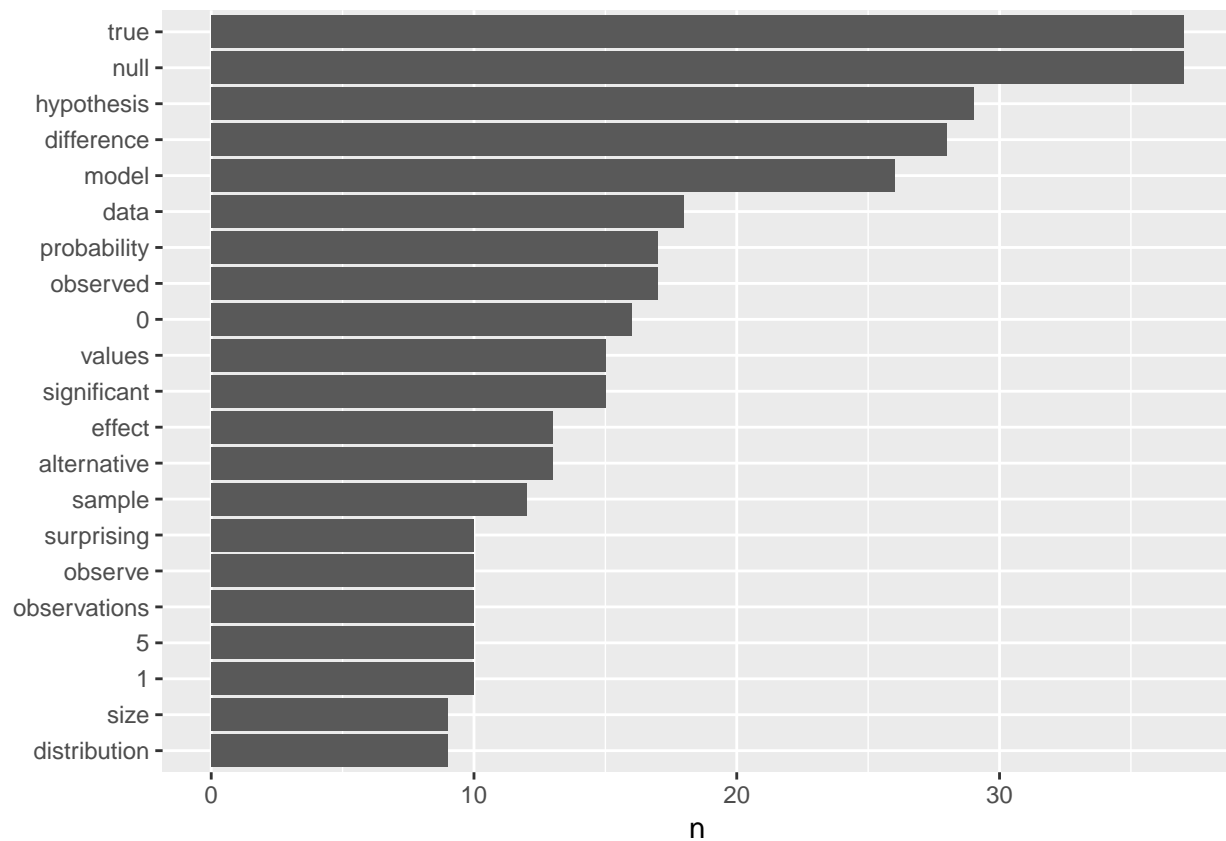
## The following objects are masked from 'package:stats':
##
##   filter, lag

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

## Joining, by = "word"

## # A tibble: 282 x 2
##   word          n
##   <chr>      <int>
## 1 null         37
## 2 true         37
## 3 hypothesis   29
## 4 difference   28
## 5 model        26
## 6 data         18
## 7 observed     17
## 8 probability  17
## 9 0            16
## 10 significant 15
## # ... with 272 more rows

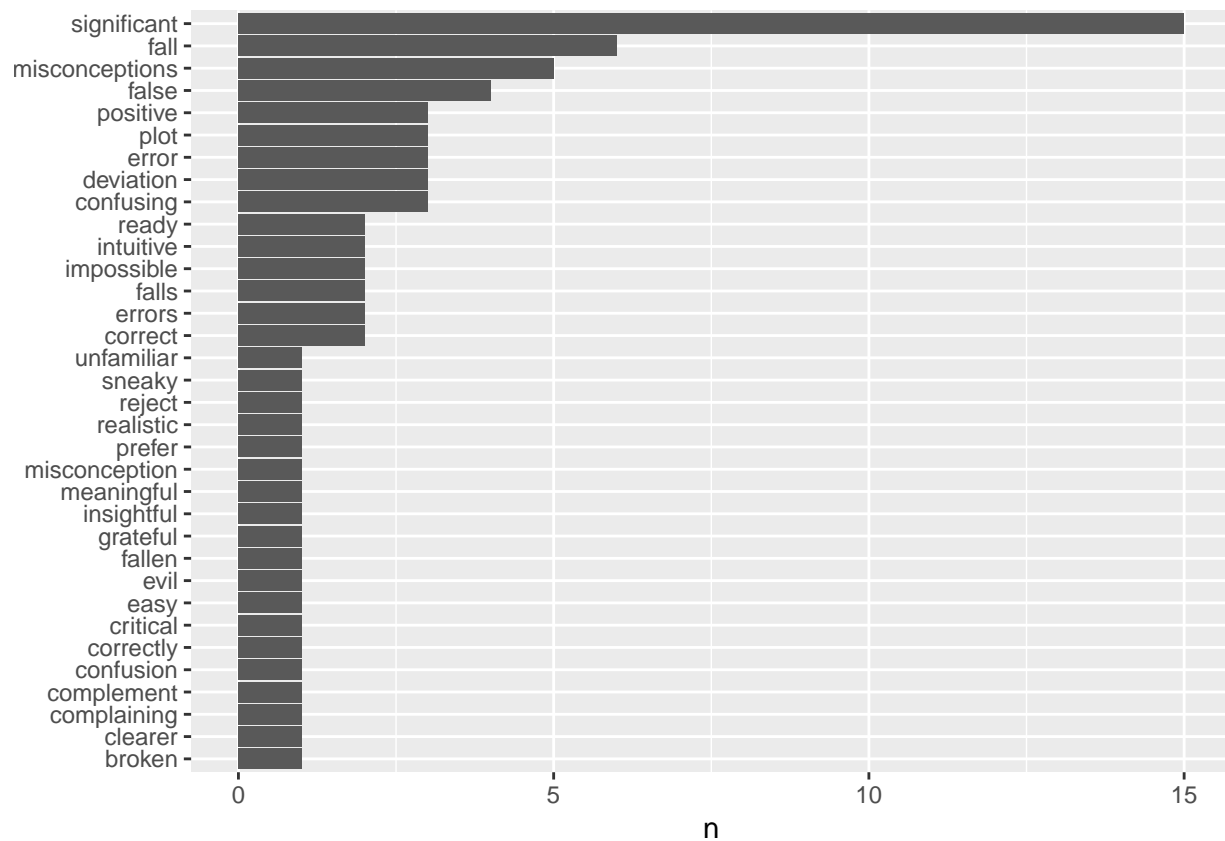
## Selecting by n
```



After eliminating the stop words in the article, we order the words appeared in the passage by frequency and we made a ggplot to show the 20 most frequent words appear in the article.

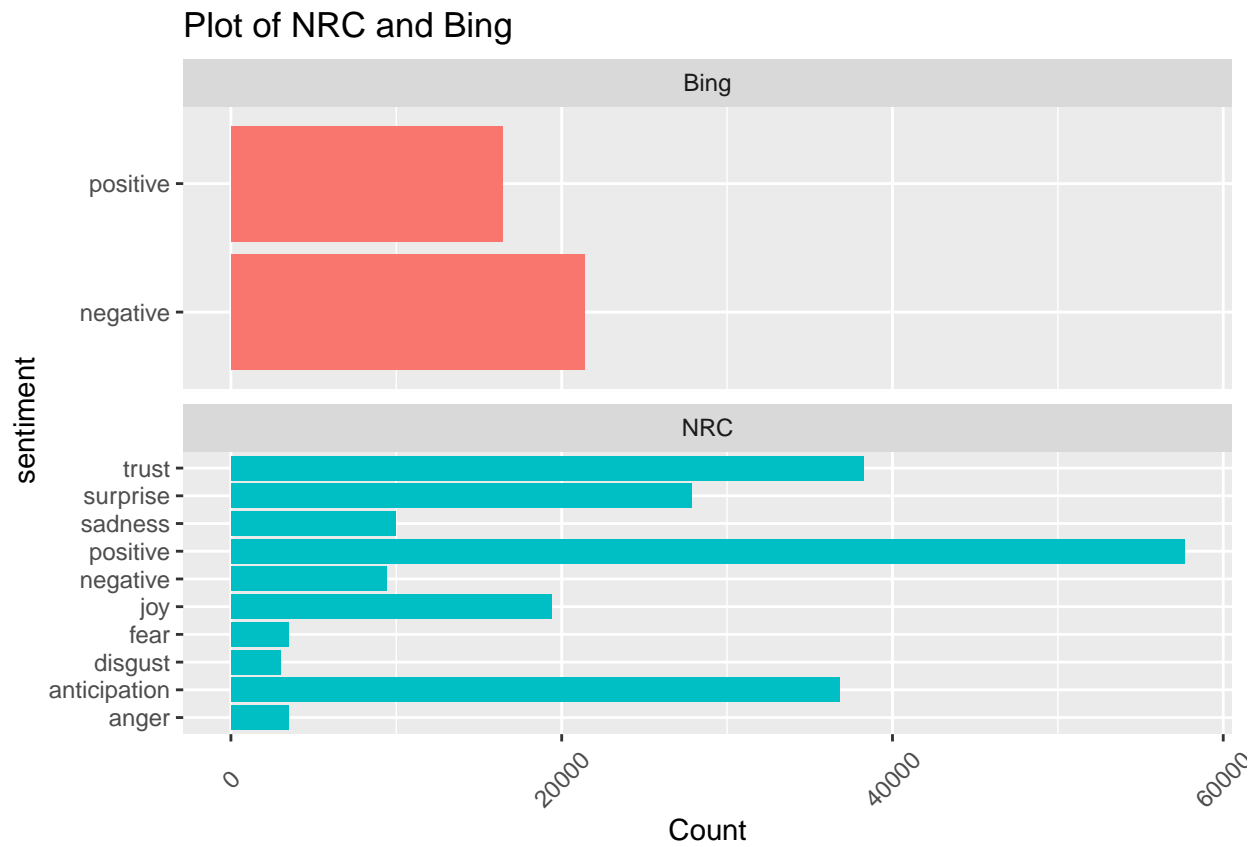
```
## Joining, by = "word"
```

```
## Selecting by n
```

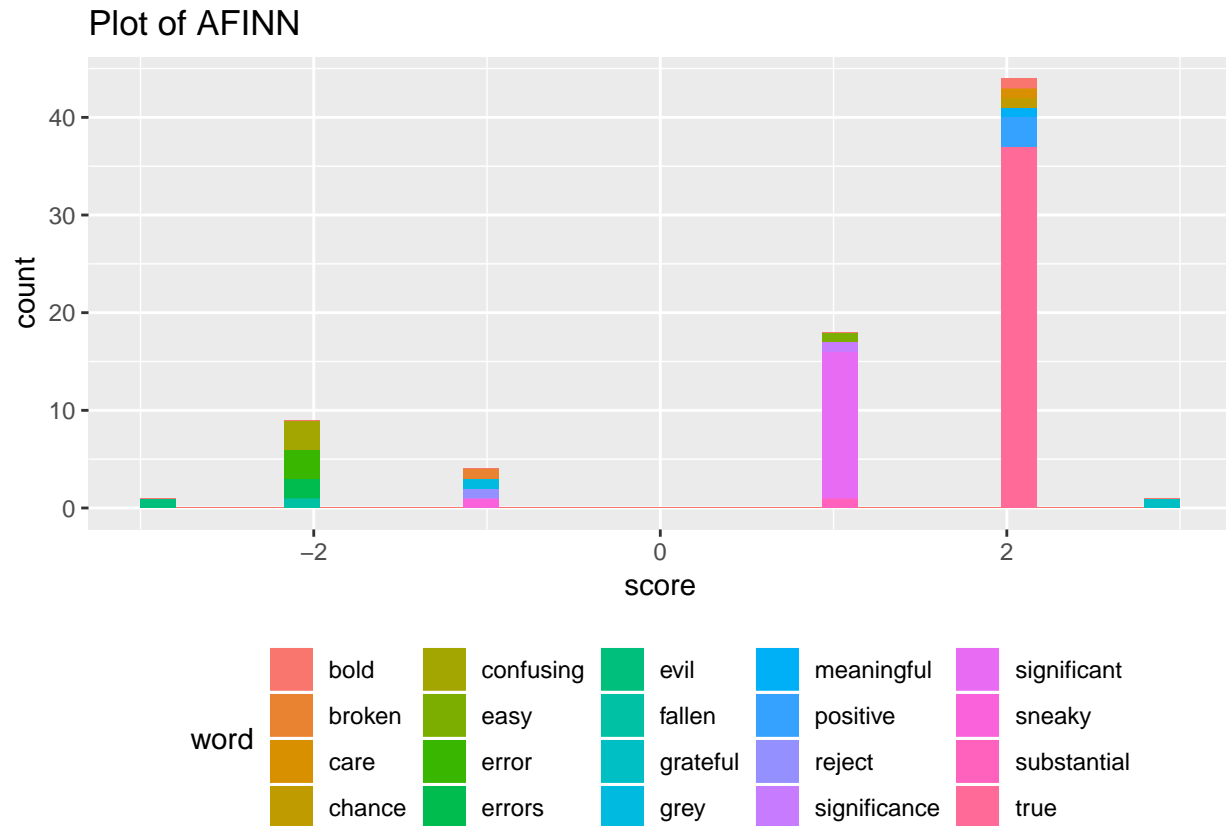


In order to get an odea of the passage's sentiment on P-Value, we applied Bing sentiment package, and made a ggplot of the top 20 sentimental words in the article. The first one "significant" is about 3 times more frequent than the second word in order. That should be due to the term "statistically significant". Then we want to compare the results from the other two packages of sentimenal words: AFINN and NRC.

```
## Joining, by = "word"
## Joining, by = "word"
```

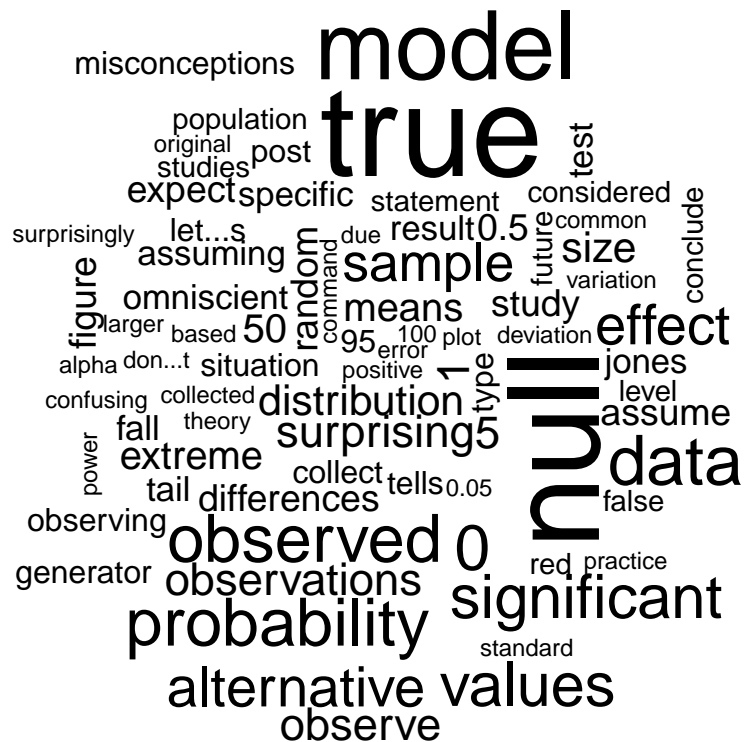


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We want to also see the wordcloud.

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

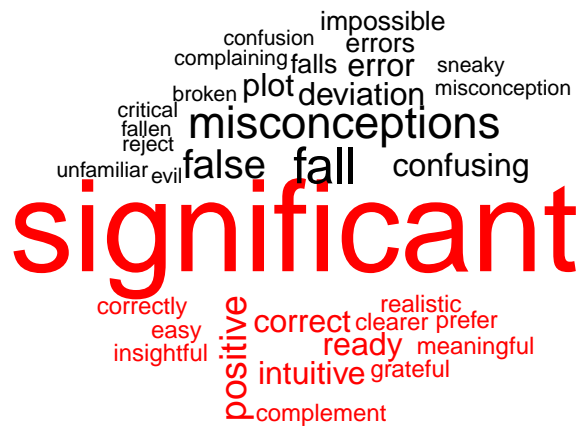


```
##
```

```
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##      smiths

## Joining, by = "word"
```



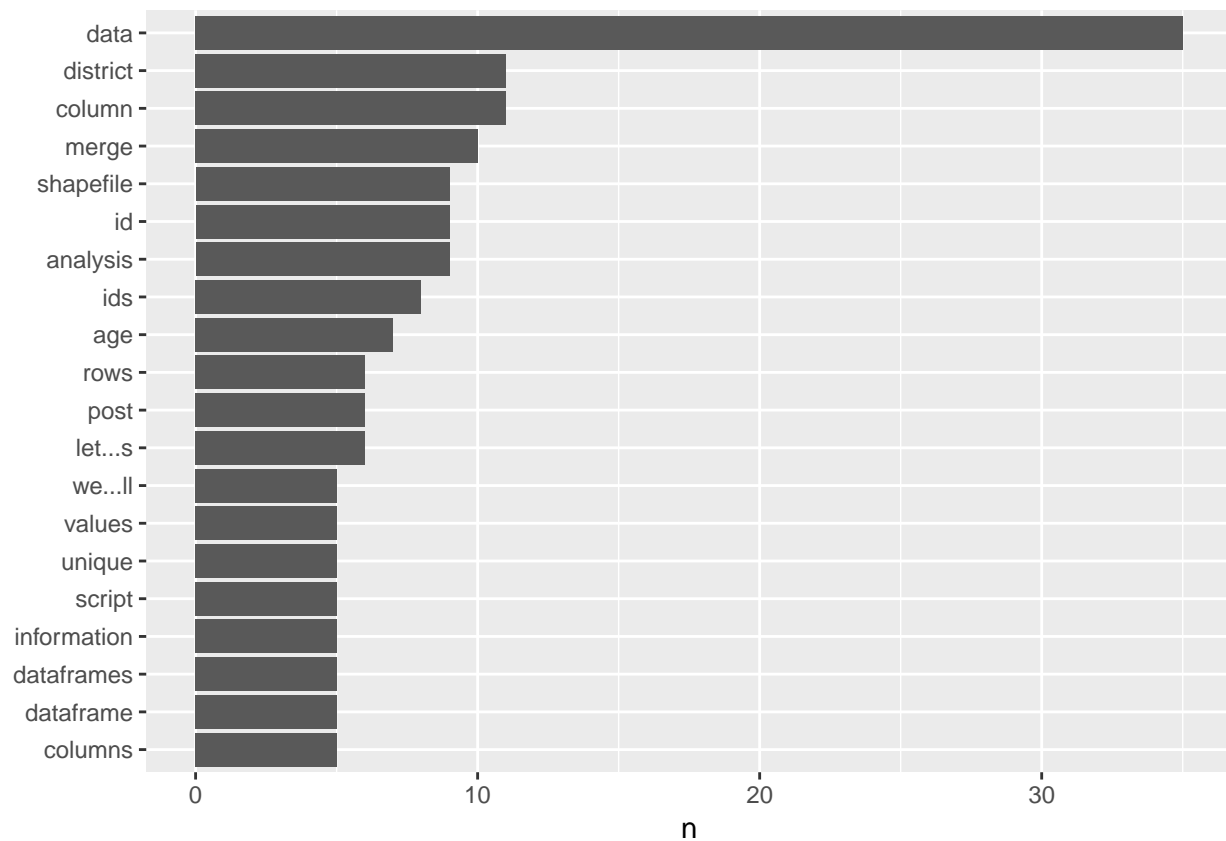
From the Data to the Story Article

We want to analyze the passage from <https://correlaid.org/blog/posts/journocode-workflow>.

```
## Joining, by = "word"

## # A tibble: 285 x 2
##   word      n
##   <chr>    <int>
## 1 data      35
## 2 column    11
## 3 district  11
## 4 merge     10
## 5 analysis   9
## 6 id         9
## 7 shapefile  9
## 8 ids        8
## 9 age        7
## 10 let's     6
## # ... with 275 more rows

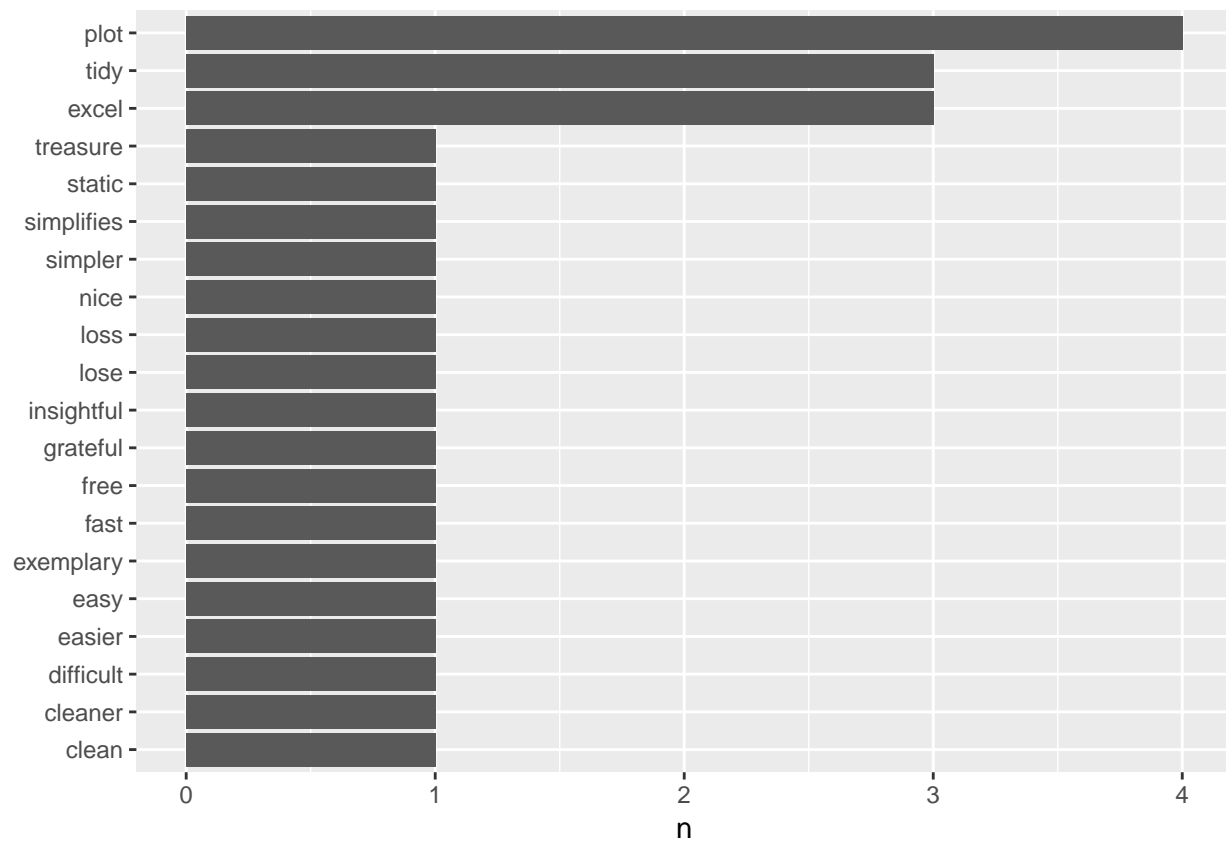
## Selecting by n
```



After eliminating the stop words in the article, we order the words appeared in the passage by frequency and we made a ggplot to show the 20 most frequent words appear in the article.

```
## Joining, by = "word"
```

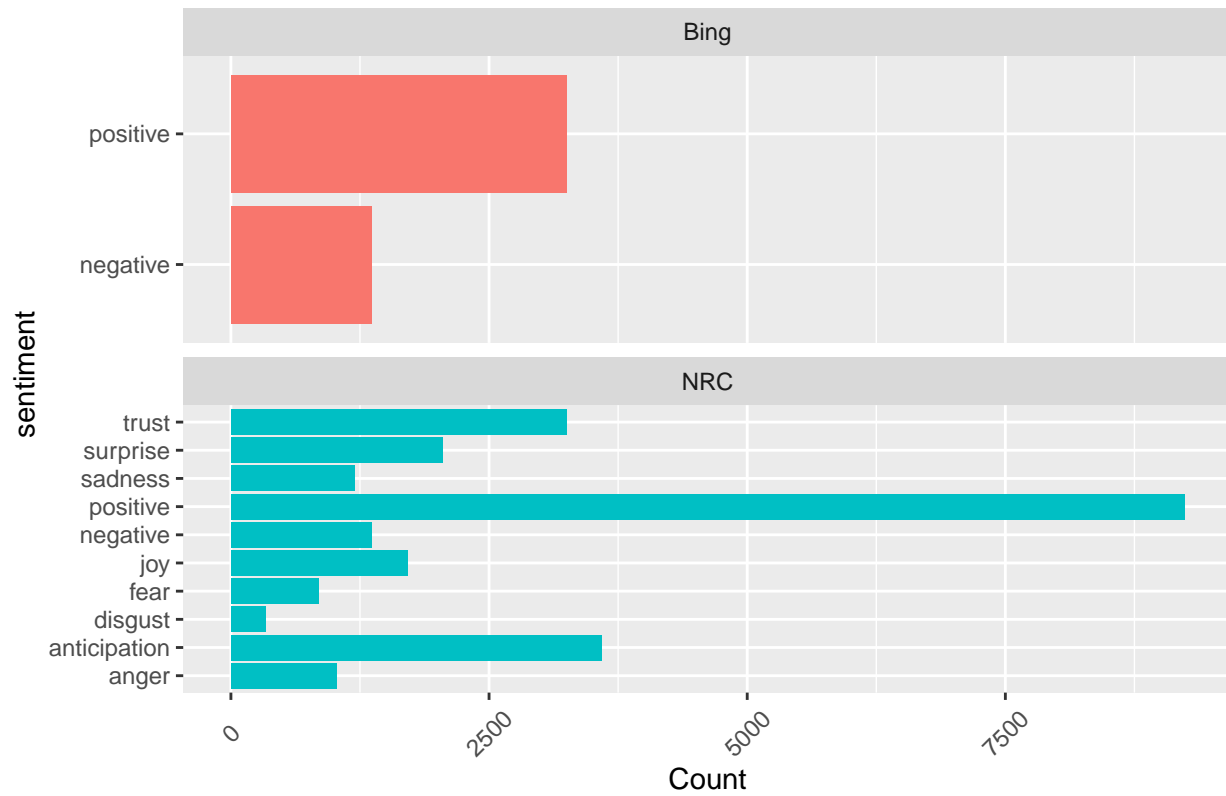
```
## Selecting by n
```



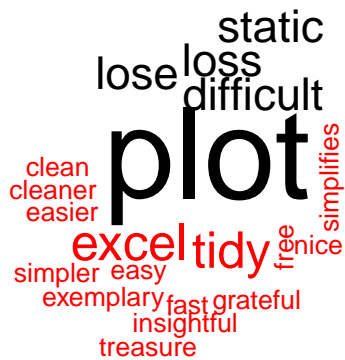
In order to get an odea of the passage’s sentiment on this article, we applied Bing sentiment package, and made a ggplot of the top 20 sentimental words in the article. The first three are “plot”, “excel” and “tidy”, which is reasonable because this is a tutorial of R. Then we want to compare the results from the other two packages of sentimenal words: AFINN and NRC.

```
## Joining, by = "word"
## Joining, by = "word"
```


Plot of NRC and Bing

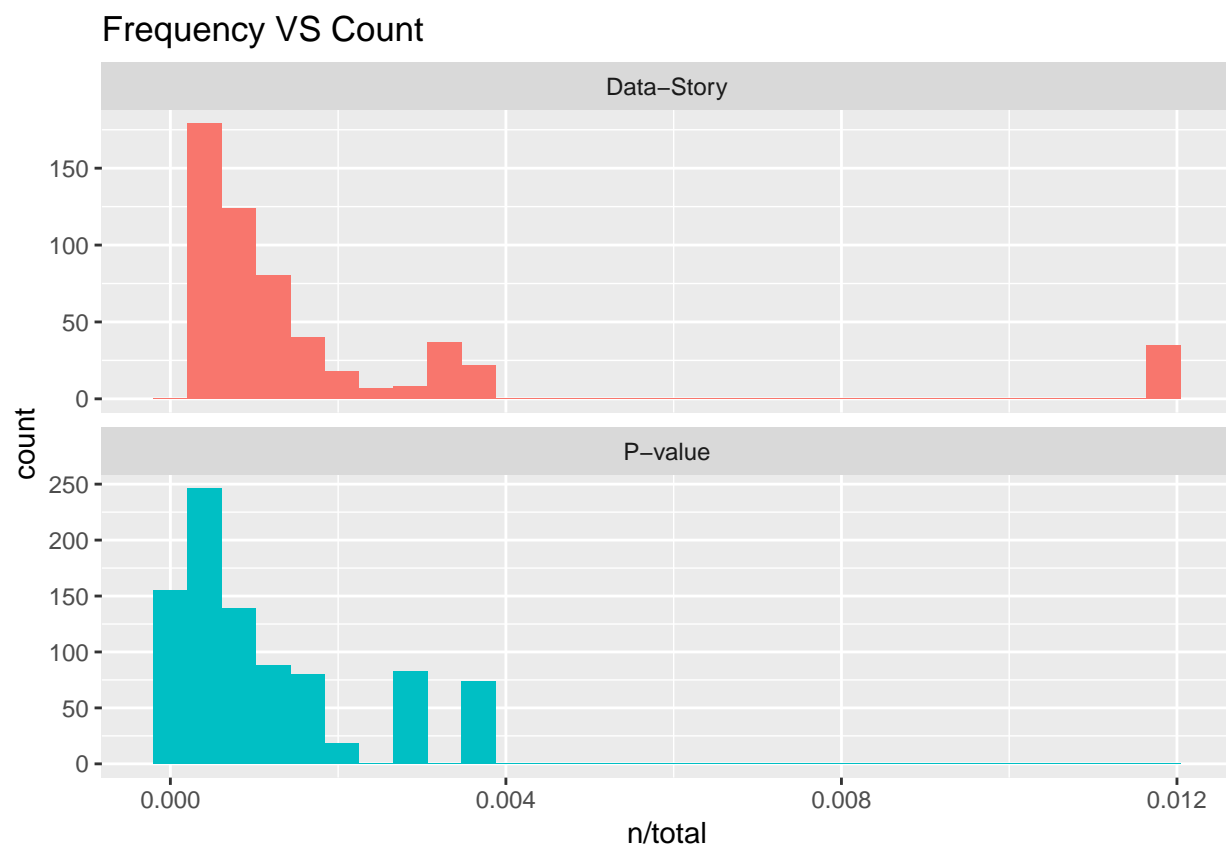


`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



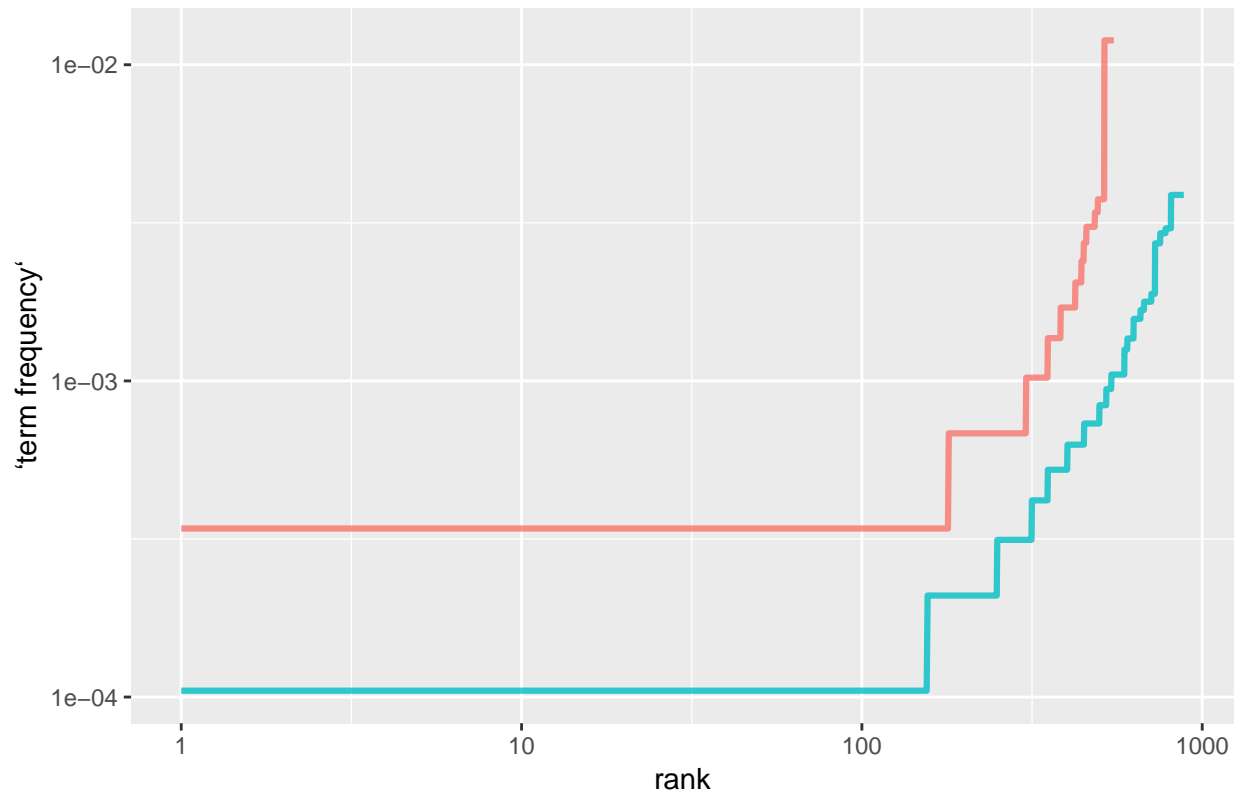
Combined Analysis on Two Articles To find important words for the context by decreasing the weight for commonly used words, we apply `bind_tf_idf` function for these two article.

``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



We can see that the tails are not so long and these two article exhibit similar distribution. Their peaks are at sim-

Rand VS Term Frequency



ilar points.

The result is totally opposite to the Zipf's Law, which states that a word appears is inversely proportional to its rank.

Then we apply `bind_tf_idf` function to find the important words for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that not used very much.

N-grams and Correlations We want to check the words as bigrams from now on.

Then we apply `bind_tf_idf` function to find the important bigrams.

##	bigram	article	n	tf	idf
## 1	null hypothesis	P-Value	439	0.003891051	-2.3025851
## 2	null hypothesis	P-Value	439	0.003891051	-2.3025851
## 3	null hypothesis	P-Value	439	0.003891051	-2.3025851
## 4	text books	P-Value	439	0.003891051	0.6931472
## 5	power analysis	P-Value	439	0.003891051	0.6931472
## 6	analysis software	P-Value	439	0.003891051	0.6931472
## 7	horizontal axis	P-Value	439	0.003891051	0.6931472
## 8	calculated based	P-Value	439	0.003891051	0.6931472
## 9	normal distribution	P-Value	439	0.003891051	0.6931472
## 10	sample size	P-Value	439	0.003891051	-1.0986123
## 11	null hypothesis	P-Value	439	0.003891051	-2.3025851
## 12	null model	P-Value	439	0.003891051	-2.0149030
## 13	null hypothesis	P-Value	439	0.003891051	-2.3025851
## 14	post i've	P-Value	439	0.003891051	0.6931472
## 15	i've recently	P-Value	439	0.003891051	0.6931472
## 16	recently realized	P-Value	439	0.003891051	0.6931472
## 17	lot clearer	P-Value	439	0.003891051	0.6931472
## 18	null model	P-Value	439	0.003891051	-2.0149030
## 19	model assuming	P-Value	439	0.003891051	0.0000000

## 20	standard deviation	P-Value 439 0.003891051 -0.4054651
## 21	test comparing	P-Value 439 0.003891051 0.6931472
## 22	sd 1	P-Value 439 0.003891051 0.0000000
## 23	effect size	P-Value 439 0.003891051 0.0000000
## 24	null model	P-Value 439 0.003891051 -2.0149030
## 25	true standard	P-Value 439 0.003891051 0.6931472
## 26	standard deviation	P-Value 439 0.003891051 -0.4054651
## 27	sample size	P-Value 439 0.003891051 -1.0986123
## 28	null model	P-Value 439 0.003891051 -2.0149030
## 29	raw scores	P-Value 439 0.003891051 0.6931472
## 30	null model	P-Value 439 0.003891051 -2.0149030
## 31	sample size	P-Value 439 0.003891051 -1.0986123
## 32	size increases	P-Value 439 0.003891051 0.6931472
## 33	collect 5000	P-Value 439 0.003891051 0.6931472
## 34	50 observations	P-Value 439 0.003891051 -0.9162907
## 35	null model	P-Value 439 0.003891051 -2.0149030
## 36	null model	P-Value 439 0.003891051 -2.0149030
## 37	0 due	P-Value 439 0.003891051 0.6931472
## 38	larger sample	P-Value 439 0.003891051 0.6931472
## 39	sample size	P-Value 439 0.003891051 -1.0986123
## 40	sample closer	P-Value 439 0.003891051 0.6931472
## 41	0 compared	P-Value 439 0.003891051 0.6931472
## 42	null model	P-Value 439 0.003891051 -2.0149030
## 43	50 observations	P-Value 439 0.003891051 -0.9162907
## 44	colored red	P-Value 439 0.003891051 0.6931472
## 45	represent 2.5	P-Value 439 0.003891051 0.6931472
## 46	left tail	P-Value 439 0.003891051 0.6931472
## 47	0 representing	P-Value 439 0.003891051 0.6931472
## 48	alpha level	P-Value 439 0.003891051 0.0000000
## 49	vertical axis	P-Value 439 0.003891051 0.6931472
## 50	curves let's	P-Value 439 0.003891051 0.6931472
## 51	let's assume	P-Value 439 0.003891051 -0.4054651
## 52	figure visualizing	P-Value 439 0.003891051 0.6931472
## 53	null model	P-Value 439 0.003891051 -2.0149030
## 54	observation falls	P-Value 439 0.003891051 0.6931472
## 55	tailed test	P-Value 439 0.003891051 0.6931472
## 56	null model	P-Value 439 0.003891051 -2.0149030
## 57	collected 5000	P-Value 439 0.003891051 0.6931472
## 58	5000 observations	P-Value 439 0.003891051 0.6931472
## 59	collected 50	P-Value 439 0.003891051 0.6931472
## 60	50 observations	P-Value 439 0.003891051 -0.9162907
## 61	address common	P-Value 439 0.003891051 0.6931472
## 62	common misconceptions	P-Value 439 0.003891051 -0.4054651
## 63	alternative hypothesis	P-Value 439 0.003891051 -0.9162907
## 64	alternative model	P-Value 439 0.003891051 -1.2527630
## 65	let's assume	P-Value 439 0.003891051 -0.4054651
## 66	knowing entity	P-Value 439 0.003891051 0.0000000
## 67	paul meehl	P-Value 439 0.003891051 0.6931472
## 68	knowing entity	P-Value 439 0.003891051 0.0000000
## 69	entity omniscient	P-Value 439 0.003891051 0.6931472
## 70	omniscient jones	P-Value 439 0.003891051 -1.0986123
## 71	50 observations	P-Value 439 0.003891051 -0.9162907
## 72	observations omniscient	P-Value 439 0.003891051 0.0000000
## 73	omniscient jones	P-Value 439 0.003891051 -1.0986123

## 74	expected data	P-Value 439 0.003891051 0.6931472
## 75	data pattern	P-Value 439 0.003891051 0.6931472
## 76	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 77	grey line	P-Value 439 0.003891051 0.6931472
## 78	alternative model	P-Value 439 0.003891051 -1.2527630
## 79	model assuming	P-Value 439 0.003891051 0.0000000
## 80	0.5 exists	P-Value 439 0.003891051 0.6931472
## 81	black line	P-Value 439 0.003891051 0.6931472
## 82	omniscient jones	P-Value 439 0.003891051 -1.0986123
## 83	true difference	P-Value 439 0.003891051 0.6931472
## 84	larger let's	P-Value 439 0.003891051 0.6931472
## 85	let's assume	P-Value 439 0.003891051 -0.4054651
## 86	50 observations	P-Value 439 0.003891051 -0.9162907
## 87	observations omniscient	P-Value 439 0.003891051 0.0000000
## 88	omniscient jones	P-Value 439 0.003891051 -1.0986123
## 89	jones tells	P-Value 439 0.003891051 -0.4054651
## 90	null model	P-Value 439 0.003891051 -2.0149030
## 91	alternative model	P-Value 439 0.003891051 -1.2527630
## 92	finally ready	P-Value 439 0.003891051 0.6931472
## 93	common misconceptions	P-Value 439 0.003891051 -0.4054651
## 94	values interpreted	P-Value 439 0.003891051 0.6931472
## 95	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 96	true let's	P-Value 439 0.003891051 0.6931472
## 97	significant result	P-Value 439 0.003891051 -1.0986123
## 98	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 99	omniscient jones	P-Value 439 0.003891051 -1.0986123
## 100	jones tells	P-Value 439 0.003891051 -0.4054651
## 101	alpha level	P-Value 439 0.003891051 0.0000000
## 102	alternative model	P-Value 439 0.003891051 -1.2527630
## 103	null model	P-Value 439 0.003891051 -2.0149030
## 104	extremely surprising	P-Value 439 0.003891051 0.6931472
## 105	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 106	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 107	hypothesis true	P-Value 439 0.003891051 0.6931472
## 108	alternative hypothesis	P-Value 439 0.003891051 -0.9162907
## 109	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 110	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 111	false imagine	P-Value 439 0.003891051 0.6931472
## 112	command rnorm	P-Value 439 0.003891051 0.6931472
## 113	0 sd	P-Value 439 0.003891051 0.6931472
## 114	sd 1	P-Value 439 0.003891051 0.0000000
## 115	1 rcopy	P-Value 439 0.003891051 0.6931472
## 116	command generates	P-Value 439 0.003891051 0.6931472
## 117	generates 50	P-Value 439 0.003891051 0.6931472
## 118	50 random	P-Value 439 0.003891051 0.6931472
## 119	random observations	P-Value 439 0.003891051 0.6931472
## 120	standard deviation	P-Value 439 0.003891051 -0.4054651
## 121	test tells	P-Value 439 0.003891051 0.6931472
## 122	surprisingly extreme	P-Value 439 0.003891051 0.6931472
## 123	extreme assuming	P-Value 439 0.003891051 0.6931472
## 124	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 125	bold move	P-Value 439 0.003891051 0.6931472
## 126	observing surprising	P-Value 439 0.003891051 0.6931472
## 127	surprising data	P-Value 439 0.003891051 0.6931472

## 128	data assuming	P-Value 439	0.003891051	0.6931472
## 129	null hypothesis	P-Value 439	0.003891051	-2.3025851
## 130	conclude based	P-Value 439	0.003891051	0.6931472
## 131	extreme outcome	P-Value 439	0.003891051	0.6931472
## 132	considered surprising	P-Value 439	0.003891051	-0.6931472
## 133	null hypothesis	P-Value 439	0.003891051	-2.3025851
## 134	alternative hypothesis	P-Value 439	0.003891051	-0.9162907
## 135	evil hackers	P-Value 439	0.003891051	0.6931472
## 136	hackers taking	P-Value 439	0.003891051	0.6931472
## 137	chance note	P-Value 439	0.003891051	0.6931472
## 138	null hypothesis	P-Value 439	0.003891051	-2.3025851
## 139	random variation	P-Value 439	0.003891051	0.6931472
## 140	observe extreme	P-Value 439	0.003891051	0.6931472
## 141	extreme data	P-Value 439	0.003891051	0.0000000
## 142	basically 100	P-Value 439	0.003891051	0.6931472
## 143	null hypothesis	P-Value 439	0.003891051	-2.3025851
## 144	extreme data	P-Value 439	0.003891051	0.0000000
## 145	95 remember	P-Value 439	0.003891051	0.6931472
## 146	hypothesis 3	P-Value 439	0.003891051	0.6931472
## 147	null model	P-Value 439	0.003891051	-2.0149030
## 148	sample size	P-Value 439	0.003891051	-1.0986123
## 149	considered surprising	P-Value 439	0.003891051	-0.6931472
## 150	sample size	P-Value 439	0.003891051	-1.0986123
## 151	considered surprising	P-Value 439	0.003891051	-0.6931472
## 152	surprising due	P-Value 439	0.003891051	0.6931472
## 153	substantial level	P-Value 439	0.003891051	0.6931472
## 154	data note	P-Value 439	0.003891051	0.6931472
## 155	null hypothesis	P-Value 439	0.003891051	-2.3025851
## 156	observed data	P-Value 439	0.003891051	0.6931472
## 157	considered surprising	P-Value 439	0.003891051	-0.6931472
## 158	verbal label	P-Value 439	0.003891051	0.6931472
## 159	label significant	P-Value 439	0.003891051	0.6931472
## 160	significant effect	P-Value 439	0.003891051	0.6931472
## 161	surprising effect	P-Value 439	0.003891051	0.6931472
## 162	null model	P-Value 439	0.003891051	-2.0149030
## 163	automatically true	P-Value 439	0.003891051	0.6931472
## 164	interpret effect	P-Value 439	0.003891051	0.6931472
## 165	effect sizes	P-Value 439	0.003891051	0.6931472
## 166	hypothesis tests	P-Value 439	0.003891051	0.6931472
## 167	hypothesis test	P-Value 439	0.003891051	0.6931472
## 168	equivalence test	P-Value 439	0.003891051	0.6931472
## 169	observed difference	P-Value 439	0.003891051	0.0000000
## 170	observed difference	P-Value 439	0.003891051	0.0000000
## 171	surprisingly closer	P-Value 439	0.003891051	0.6931472
## 172	significant finding	P-Value 439	0.003891051	0.6931472
## 173	type 1	P-Value 439	0.003891051	-0.6931472
## 174	1 error	P-Value 439	0.003891051	-0.4054651
## 175	false positive	P-Value 439	0.003891051	-0.4054651
## 176	5 assume	P-Value 439	0.003891051	0.6931472
## 177	collect 20	P-Value 439	0.003891051	0.6931472
## 178	20 observations	P-Value 439	0.003891051	0.6931472
## 179	omniscient jones	P-Value 439	0.003891051	-1.0986123
## 180	jones tells	P-Value 439	0.003891051	-0.4054651
## 181	null hypothesis	P-Value 439	0.003891051	-2.3025851

## 182	significant result	P-Value 439 0.003891051 -1.0986123
## 183	false positive	P-Value 439 0.003891051 -0.4054651
## 184	significant results	P-Value 439 0.003891051 0.6931472
## 185	type 1	P-Value 439 0.003891051 -0.6931472
## 186	1 errors	P-Value 439 0.003891051 0.6931472
## 187	type 1	P-Value 439 0.003891051 -0.6931472
## 188	1 error	P-Value 439 0.003891051 -0.4054651
## 189	error rate	P-Value 439 0.003891051 0.6931472
## 190	rate controls	P-Value 439 0.003891051 0.6931472
## 191	red tail	P-Value 439 0.003891051 0.6931472
## 192	type 1	P-Value 439 0.003891051 -0.6931472
## 193	1 error	P-Value 439 0.003891051 -0.4054651
## 194	significant result	P-Value 439 0.003891051 -1.0986123
## 195	5 probability	P-Value 439 0.003891051 0.6931472
## 196	false positive	P-Value 439 0.003891051 -0.4054651
## 197	collect data	P-Value 439 0.003891051 0.6931472
## 198	run 5	P-Value 439 0.003891051 0.6931472
## 199	significant result	P-Value 439 0.003891051 -1.0986123
## 200	replicate based	P-Value 439 0.003891051 0.6931472
## 201	future studies	P-Value 439 0.003891051 -0.4054651
## 202	additional assumptions	P-Value 439 0.003891051 0.6931472
## 203	assumptions e.g	P-Value 439 0.003891051 0.6931472
## 204	alternative effect	P-Value 439 0.003891051 0.6931472
## 205	effect size	P-Value 439 0.003891051 0.0000000
## 206	original study	P-Value 439 0.003891051 0.0000000
## 207	future studies	P-Value 439 0.003891051 -0.4054651
## 208	specific situation	P-Value 439 0.003891051 -0.9162907
## 209	future studies	P-Value 439 0.003891051 -0.4054651
## 210	specific situation	P-Value 439 0.003891051 -0.9162907
## 211	null model	P-Value 439 0.003891051 -2.0149030
## 212	alternative model	P-Value 439 0.003891051 -1.2527630
## 213	150 observations	P-Value 439 0.003891051 0.6931472
## 214	difference falls	P-Value 439 0.003891051 0.6931472
## 215	significance level	P-Value 439 0.003891051 0.6931472
## 216	specific situation	P-Value 439 0.003891051 -0.9162907
## 217	95 probable	P-Value 439 0.003891051 0.6931472
## 218	significant result	P-Value 439 0.003891051 -1.0986123
## 219	replication study	P-Value 439 0.003891051 0.6931472
## 220	study assuming	P-Value 439 0.003891051 0.6931472
## 221	true effect	P-Value 439 0.003891051 0.6931472
## 222	alternative model	P-Value 439 0.003891051 -1.2527630
## 223	alternative model	P-Value 439 0.003891051 -1.2527630
## 224	true 95	P-Value 439 0.003891051 0.6931472
## 225	95 1	P-Value 439 0.003891051 0.6931472
## 226	observed means	P-Value 439 0.003891051 0.0000000
## 227	original study	P-Value 439 0.003891051 0.0000000
## 228	statistical power	P-Value 439 0.003891051 0.6931472
## 229	observed means	P-Value 439 0.003891051 0.0000000
## 230	type 2	P-Value 439 0.003891051 0.6931472
## 231	2 errors	P-Value 439 0.003891051 0.6931472
## 232	specific situation	P-Value 439 0.003891051 -0.9162907
## 233	alternative hypothesis	P-Value 439 0.003891051 -0.9162907
## 234	null hypothesis	P-Value 439 0.003891051 -2.3025851
## 235	specific situation	P-Value 439 0.003891051 -0.9162907

## 236	alternative hypothesis	P-Value	439	0.003891051	-0.9162907
## 237	specific size	P-Value	439	0.003891051	0.6931472
## 238	future study	P-Value	439	0.003891051	0.6931472
## 239	significant result	P-Value	439	0.003891051	-1.0986123
## 240	result conclusion	P-Value	439	0.003891051	0.6931472
## 241	conclusion probabilities	P-Value	439	0.003891051	0.6931472
## 242	intuitive grammar	P-Value	439	0.003891051	0.6931472
## 243	practice grammar	P-Value	439	0.003891051	0.6931472
## 244	don't practice	P-Value	439	0.003891051	0.6931472
## 245	researchers don't	P-Value	439	0.003891051	0.6931472
## 246	don't understand	P-Value	439	0.003891051	0.6931472
## 247	explain common	P-Value	439	0.003891051	0.6931472
## 248	common misconceptions	P-Value	439	0.003891051	-0.4054651
## 249	misconceptions multiple	P-Value	439	0.003891051	0.6931472
## 250	multiple times	P-Value	439	0.003891051	0.6931472
## 251	post originally	P-Value	439	0.003891051	0.0000000
## 252	december 5th	P-Value	439	0.003891051	0.6931472
## 253	5th 2017	P-Value	439	0.003891051	0.6931472
## 254	cross post	P-Value	439	0.003891051	0.0000000
## 255	insightful post	P-Value	439	0.003891051	0.0000000
## 256	original blog	P-Value	439	0.003891051	0.0000000
## 257	blog post	P-Value	439	0.003891051	0.0000000
## 258	standard steps	Data-Story	439	0.005494505	0.6931472
## 259	data driven	Data-Story	439	0.005494505	0.6931472
## 260	driven project	Data-Story	439	0.005494505	0.6931472
## 261	exemplary data	Data-Story	439	0.005494505	0.6931472
## 262	data journalism	Data-Story	439	0.005494505	0.6931472
## 263	workflow we'll	Data-Story	439	0.005494505	0.6931472
## 264	bbsr germany	Data-Story	439	0.005494505	0.6931472
## 265	commented code	Data-Story	439	0.005494505	0.6931472
## 266	github page	Data-Story	439	0.005494505	0.0000000
## 267	makes collaboration	Data-Story	439	0.005494505	0.6931472
## 268	collaboration easier	Data-Story	439	0.005494505	0.6931472
## 269	automatically set	Data-Story	439	0.005494505	0.6931472
## 270	double click	Data-Story	439	0.005494505	0.6931472
## 271	rproj file	Data-Story	439	0.005494505	0.6931472
## 272	rstudio window	Data-Story	439	0.005494505	0.6931472
## 273	load packages	Data-Story	439	0.005494505	0.6931472
## 274	data preprocessing	Data-Story	439	0.005494505	0.6931472
## 275	analysis analysis	Data-Story	439	0.005494505	0.6931472
## 276	don't necessarily	Data-Story	439	0.005494505	0.6931472
## 277	analysis cleaner	Data-Story	439	0.005494505	0.6931472
## 278	analysis script	Data-Story	439	0.005494505	0.0000000
## 279	preprocessed data	Data-Story	439	0.005494505	0.6931472
## 280	data saved	Data-Story	439	0.005494505	0.6931472
## 281	data source	Data-Story	439	0.005494505	0.6931472
## 282	analysis script	Data-Story	439	0.005494505	0.0000000
## 283	script head	Data-Story	439	0.005494505	0.6931472
## 284	talk let's	Data-Story	439	0.005494505	0.6931472
## 285	let's start	Data-Story	439	0.005494505	0.6931472
## 286	excel worksheet	Data-Story	439	0.005494505	0.6931472
## 287	data sheets	Data-Story	439	0.005494505	0.6931472
## 288	germany's 402	Data-Story	439	0.005494505	0.6931472
## 289	402 city	Data-Story	439	0.005494505	0.6931472

```

## 290         city districts Data-Story 439 0.005494505 0.6931472
## 291         unique district Data-Story 439 0.005494505 -0.4054651
## 292         district id Data-Story 439 0.005494505 -0.9162907
## 293         city district Data-Story 439 0.005494505 0.6931472
## 294         excel sheet Data-Story 439 0.005494505 0.6931472
## 295         average age Data-Story 439 0.005494505 0.0000000
## 296         district's male Data-Story 439 0.005494505 0.6931472
## 297         male population Data-Story 439 0.005494505 0.6931472
## 298         female population Data-Story 439 0.005494505 0.0000000
## 299 population preprocessing Data-Story 439 0.005494505 0.6931472
## 300         female population Data-Story 439 0.005494505 0.0000000
## 301         district let's Data-Story 439 0.005494505 0.6931472
## 302         dataframes aren't Data-Story 439 0.005494505 0.6931472
## 303         aren't sorted Data-Story 439 0.005494505 0.6931472
## 304         column names Data-Story 439 0.005494505 0.6931472
## 305         names aren't Data-Story 439 0.005494505 0.6931472
## 306         merge we'll Data-Story 439 0.005494505 0.6931472
## 307         german districts Data-Story 439 0.005494505 0.6931472
## 308         districts what's Data-Story 439 0.005494505 0.6931472
## 309         duplicated rows Data-Story 439 0.005494505 0.0000000
## 310         duplicated rows Data-Story 439 0.005494505 0.0000000
## 311         let's merge Data-Story 439 0.005494505 0.6931472
## 312         function cbind Data-Story 439 0.005494505 0.6931472
## 313         simply add Data-Story 439 0.005494505 0.6931472
## 314         age column Data-Story 439 0.005494505 0.0000000
## 315         merge merge Data-Story 439 0.005494505 0.6931472
## 316         merge joins Data-Story 439 0.005494505 0.6931472
## 317         joins dataframes Data-Story 439 0.005494505 0.6931472
## 318         unique values Data-Story 439 0.005494505 0.6931472
## 319         unique district Data-Story 439 0.005494505 -0.4054651
## 320         district id Data-Story 439 0.005494505 -0.9162907
## 321         age column Data-Story 439 0.005494505 0.0000000
## 322         fourth column Data-Story 439 0.005494505 0.6931472
## 323         indexing age_female Data-Story 439 0.005494505 0.6931472
## 324         matching column Data-Story 439 0.005494505 0.6931472
## 325         parameters by.x Data-Story 439 0.005494505 0.6931472
## 326         matching columns Data-Story 439 0.005494505 0.6931472
## 327         p.s merging Data-Story 439 0.005494505 0.6931472
## 328         data frames Data-Story 439 0.005494505 0.6931472
## 329         unmatchable rows Data-Story 439 0.005494505 0.6931472
## 330         rows type Data-Story 439 0.005494505 0.6931472
## 331         type merge Data-Story 439 0.005494505 0.6931472
## 332 columns average_age_males Data-Story 439 0.005494505 0.6931472
## 333         matching values Data-Story 439 0.005494505 0.6931472
## 334         parameter key Data-Story 439 0.005494505 0.6931472
## 335         attribute names Data-Story 439 0.005494505 0.6931472
## 336         columns indexes Data-Story 439 0.005494505 0.6931472
## 337         1 3 Data-Story 439 0.005494505 0.6931472
## 338         1,2,3 analysis Data-Story 439 0.005494505 0.6931472
## 339         nice package Data-Story 439 0.005494505 0.6931472
## 340         package that's Data-Story 439 0.005494505 0.6931472
## 341         tidyr it's Data-Story 439 0.005494505 0.6931472
## 342         hadley wickham Data-Story 439 0.005494505 0.6931472
## 343         tidy data Data-Story 439 0.005494505 0.0000000

```

```

## 344         data format Data-Story 439 0.005494505 0.6931472
## 345         format let's Data-Story 439 0.005494505 0.6931472
## 346         dplyr makes Data-Story 439 0.005494505 0.6931472
## 347         easily answer Data-Story 439 0.005494505 0.6931472
## 348         answer questions Data-Story 439 0.005494505 0.6931472
## 349         unique ids Data-Story 439 0.005494505 0.6931472
## 350         ids represented Data-Story 439 0.005494505 0.6931472
## 351         district id Data-Story 439 0.005494505 -0.9162907
## 352         base function Data-Story 439 0.005494505 0.6931472
## 353         function substr Data-Story 439 0.005494505 0.6931472
## 354         type substr Data-Story 439 0.005494505 0.6931472
## 355         console finally Data-Story 439 0.005494505 0.6931472
## 356         finally arrange Data-Story 439 0.005494505 0.6931472
## 357     district_mean visualize Data-Story 439 0.005494505 0.6931472
## 358         visualize we've Data-Story 439 0.005494505 0.6931472
## 359         simply filtering Data-Story 439 0.005494505 0.6931472
## 360     filtering summarizing Data-Story 439 0.005494505 0.6931472
## 361         simple visualization Data-Story 439 0.005494505 0.6931472
## 362         visualization helps Data-Story 439 0.005494505 0.6931472
## 363         finding patterns Data-Story 439 0.005494505 0.6931472
## 364         previous post Data-Story 439 0.005494505 0.0000000
## 365         time we'll Data-Story 439 0.005494505 0.6931472
## 366         static choropleth Data-Story 439 0.005494505 0.6931472
## 367         choropleth map Data-Story 439 0.005494505 0.6931472
## 368         esri shapefile Data-Story 439 0.005494505 0.0000000
## 369         germany's city Data-Story 439 0.005494505 0.6931472
## 370         county districts Data-Story 439 0.005494505 0.6931472
## 371         esri shapefile Data-Story 439 0.005494505 0.0000000
## 372         multiple files Data-Story 439 0.005494505 0.6931472
## 373         shp file Data-Story 439 0.005494505 0.6931472
## 374         rgdal's readogr Data-Story 439 0.005494505 0.6931472
## 375         readogr krs_shape Data-Story 439 0.005494505 0.6931472
## 376         krs_shape basically Data-Story 439 0.005494505 0.6931472
## 377         basically consists Data-Story 439 0.005494505 0.6931472
## 378         krs_shape data Data-Story 439 0.005494505 0.0000000
## 379     geographic information Data-Story 439 0.005494505 0.6931472
## 380         krs_shape polygons Data-Story 439 0.005494505 0.6931472
## 381     column krscontaining Data-Story 439 0.005494505 0.6931472
## 382         unique district Data-Story 439 0.005494505 -0.4054651
## 383         district ids Data-Story 439 0.005494505 0.0000000
## 384         age dataframe Data-Story 439 0.005494505 0.6931472
## 385         dataframes plot Data-Story 439 0.005494505 0.6931472
## 386         age values Data-Story 439 0.005494505 0.6931472
## 387         tidy data Data-Story 439 0.005494505 0.0000000
## 388         we'll loaded Data-Story 439 0.005494505 0.6931472
## 389         package broom Data-Story 439 0.005494505 0.6931472
## 390         simple broom's Data-Story 439 0.005494505 0.6931472
## 391         broom's tidy Data-Story 439 0.005494505 0.6931472
## 392         tidy function Data-Story 439 0.005494505 0.6931472
## 393     function simplifies Data-Story 439 0.005494505 0.6931472
## 394         district id Data-Story 439 0.005494505 -0.9162907
## 395         district id Data-Story 439 0.005494505 -0.9162907
## 396         krs_shape data Data-Story 439 0.005494505 0.0000000
## 397         id 0 Data-Story 439 0.005494505 0.6931472

```

```

## 398          würzburg id Data-Story 439 0.005494505 0.6931472
## 399              id 1 Data-Story 439 0.005494505 0.6931472
## 400      shapefiles district Data-Story 439 0.005494505 0.6931472
## 401          district ids Data-Story 439 0.005494505 0.0000000
## 402      tidied shapefile Data-Story 439 0.005494505 0.0000000
## 403          shapefile ids Data-Story 439 0.005494505 0.6931472
## 404      tidied shapefile Data-Story 439 0.005494505 0.0000000
## 405          shapefile rows Data-Story 439 0.005494505 0.6931472
## 406              set all.x Data-Story 439 0.005494505 0.6931472
## 407              id column Data-Story 439 0.005494505 0.6931472
## 408          final plotting Data-Story 439 0.005494505 0.6931472
## 409          plotting data Data-Story 439 0.005494505 0.6931472
## 410      district's shapefile Data-Story 439 0.005494505 0.6931472
## 411      additional information Data-Story 439 0.005494505 0.6931472
## 412      district's average Data-Story 439 0.005494505 0.6931472
## 413          average age Data-Story 439 0.005494505 0.0000000
## 414          age let's Data-Story 439 0.005494505 0.6931472
## 415          let's plot Data-Story 439 0.005494505 0.6931472
## 416      ggplot basically Data-Story 439 0.005494505 0.6931472
## 417          previous post Data-Story 439 0.005494505 0.0000000
## 418              post i'll Data-Story 439 0.005494505 0.6931472
## 419          data isn't Data-Story 439 0.005494505 0.6931472
## 420          data story Data-Story 439 0.005494505 0.6931472
## 421          story treasure Data-Story 439 0.005494505 0.6931472
## 422      eastern germany Data-Story 439 0.005494505 0.6931472
## 423          arranged lists Data-Story 439 0.005494505 0.6931472
## 424              dig deeper Data-Story 439 0.005494505 0.6931472
## 425          data analysis Data-Story 439 0.005494505 0.6931472
## 426          methods we've Data-Story 439 0.005494505 0.6931472
## 427              we've shown Data-Story 439 0.005494505 0.6931472
## 428          entire code Data-Story 439 0.005494505 0.6931472
## 429          github page Data-Story 439 0.005494505 0.0000000
## 430      questions suggestions Data-Story 439 0.005494505 0.6931472
## 431          feedback feel Data-Story 439 0.005494505 0.6931472
## 432              feel free Data-Story 439 0.005494505 0.6931472
## 433          comment we'll Data-Story 439 0.005494505 0.6931472
## 434              answer fast Data-Story 439 0.005494505 0.6931472
## 435          post originally Data-Story 439 0.005494505 0.0000000
## 436              cross post Data-Story 439 0.005494505 0.0000000
## 437          insightful post Data-Story 439 0.005494505 0.0000000
## 438              original blog Data-Story 439 0.005494505 0.0000000
## 439              blog post Data-Story 439 0.005494505 0.0000000
##          tf_idf
## 1      -0.008959475
## 2      -0.008959475
## 3      -0.008959475
## 4          0.002697071
## 5          0.002697071
## 6          0.002697071
## 7          0.002697071
## 8          0.002697071
## 9          0.002697071
## 10     -0.004274756
## 11     -0.008959475

```

12 -0.007840090
13 -0.008959475
14 0.002697071
15 0.002697071
16 0.002697071
17 0.002697071
18 -0.007840090
19 0.000000000
20 -0.001577685
21 0.002697071
22 0.000000000
23 0.000000000
24 -0.007840090
25 0.002697071
26 -0.001577685
27 -0.004274756
28 -0.007840090
29 0.002697071
30 -0.007840090
31 -0.004274756
32 0.002697071
33 0.002697071
34 -0.003565334
35 -0.007840090
36 -0.007840090
37 0.002697071
38 0.002697071
39 -0.004274756
40 0.002697071
41 0.002697071
42 -0.007840090
43 -0.003565334
44 0.002697071
45 0.002697071
46 0.002697071
47 0.002697071
48 0.000000000
49 0.002697071
50 0.002697071
51 -0.001577685
52 0.002697071
53 -0.007840090
54 0.002697071
55 0.002697071
56 -0.007840090
57 0.002697071
58 0.002697071
59 0.002697071
60 -0.003565334
61 0.002697071
62 -0.001577685
63 -0.003565334
64 -0.004874564
65 -0.001577685

66 0.000000000
67 0.002697071
68 0.000000000
69 0.002697071
70 -0.004274756
71 -0.003565334
72 0.000000000
73 -0.004274756
74 0.002697071
75 0.002697071
76 -0.008959475
77 0.002697071
78 -0.004874564
79 0.000000000
80 0.002697071
81 0.002697071
82 -0.004274756
83 0.002697071
84 0.002697071
85 -0.001577685
86 -0.003565334
87 0.000000000
88 -0.004274756
89 -0.001577685
90 -0.007840090
91 -0.004874564
92 0.002697071
93 -0.001577685
94 0.002697071
95 -0.008959475
96 0.002697071
97 -0.004274756
98 -0.008959475
99 -0.004274756
100 -0.001577685
101 0.000000000
102 -0.004874564
103 -0.007840090
104 0.002697071
105 -0.008959475
106 -0.008959475
107 0.002697071
108 -0.003565334
109 -0.008959475
110 -0.008959475
111 0.002697071
112 0.002697071
113 0.002697071
114 0.000000000
115 0.002697071
116 0.002697071
117 0.002697071
118 0.002697071
119 0.002697071

120 -0.001577685
121 0.002697071
122 0.002697071
123 0.002697071
124 -0.008959475
125 0.002697071
126 0.002697071
127 0.002697071
128 0.002697071
129 -0.008959475
130 0.002697071
131 0.002697071
132 -0.002697071
133 -0.008959475
134 -0.003565334
135 0.002697071
136 0.002697071
137 0.002697071
138 -0.008959475
139 0.002697071
140 0.002697071
141 0.000000000
142 0.002697071
143 -0.008959475
144 0.000000000
145 0.002697071
146 0.002697071
147 -0.007840090
148 -0.004274756
149 -0.002697071
150 -0.004274756
151 -0.002697071
152 0.002697071
153 0.002697071
154 0.002697071
155 -0.008959475
156 0.002697071
157 -0.002697071
158 0.002697071
159 0.002697071
160 0.002697071
161 0.002697071
162 -0.007840090
163 0.002697071
164 0.002697071
165 0.002697071
166 0.002697071
167 0.002697071
168 0.002697071
169 0.000000000
170 0.000000000
171 0.002697071
172 0.002697071
173 -0.002697071

174 -0.001577685
175 -0.001577685
176 0.002697071
177 0.002697071
178 0.002697071
179 -0.004274756
180 -0.001577685
181 -0.008959475
182 -0.004274756
183 -0.001577685
184 0.002697071
185 -0.002697071
186 0.002697071
187 -0.002697071
188 -0.001577685
189 0.002697071
190 0.002697071
191 0.002697071
192 -0.002697071
193 -0.001577685
194 -0.004274756
195 0.002697071
196 -0.001577685
197 0.002697071
198 0.002697071
199 -0.004274756
200 0.002697071
201 -0.001577685
202 0.002697071
203 0.002697071
204 0.002697071
205 0.000000000
206 0.000000000
207 -0.001577685
208 -0.003565334
209 -0.001577685
210 -0.003565334
211 -0.007840090
212 -0.004874564
213 0.002697071
214 0.002697071
215 0.002697071
216 -0.003565334
217 0.002697071
218 -0.004274756
219 0.002697071
220 0.002697071
221 0.002697071
222 -0.004874564
223 -0.004874564
224 0.002697071
225 0.002697071
226 0.000000000
227 0.000000000

228 0.002697071
229 0.000000000
230 0.002697071
231 0.002697071
232 -0.003565334
233 -0.003565334
234 -0.008959475
235 -0.003565334
236 -0.003565334
237 0.002697071
238 0.002697071
239 -0.004274756
240 0.002697071
241 0.002697071
242 0.002697071
243 0.002697071
244 0.002697071
245 0.002697071
246 0.002697071
247 0.002697071
248 -0.001577685
249 0.002697071
250 0.002697071
251 0.000000000
252 0.002697071
253 0.002697071
254 0.000000000
255 0.000000000
256 0.000000000
257 0.000000000
258 0.003808501
259 0.003808501
260 0.003808501
261 0.003808501
262 0.003808501
263 0.003808501
264 0.003808501
265 0.003808501
266 0.000000000
267 0.003808501
268 0.003808501
269 0.003808501
270 0.003808501
271 0.003808501
272 0.003808501
273 0.003808501
274 0.003808501
275 0.003808501
276 0.003808501
277 0.003808501
278 0.000000000
279 0.003808501
280 0.003808501
281 0.003808501

```
## 282 0.000000000
## 283 0.003808501
## 284 0.003808501
## 285 0.003808501
## 286 0.003808501
## 287 0.003808501
## 288 0.003808501
## 289 0.003808501
## 290 0.003808501
## 291 -0.002227830
## 292 -0.005034564
## 293 0.003808501
## 294 0.003808501
## 295 0.000000000
## 296 0.003808501
## 297 0.003808501
## 298 0.000000000
## 299 0.003808501
## 300 0.000000000
## 301 0.003808501
## 302 0.003808501
## 303 0.003808501
## 304 0.003808501
## 305 0.003808501
## 306 0.003808501
## 307 0.003808501
## 308 0.003808501
## 309 0.000000000
## 310 0.000000000
## 311 0.003808501
## 312 0.003808501
## 313 0.003808501
## 314 0.000000000
## 315 0.003808501
## 316 0.003808501
## 317 0.003808501
## 318 0.003808501
## 319 -0.002227830
## 320 -0.005034564
## 321 0.000000000
## 322 0.003808501
## 323 0.003808501
## 324 0.003808501
## 325 0.003808501
## 326 0.003808501
## 327 0.003808501
## 328 0.003808501
## 329 0.003808501
## 330 0.003808501
## 331 0.003808501
## 332 0.003808501
## 333 0.003808501
## 334 0.003808501
## 335 0.003808501
```

336 0.003808501
337 0.003808501
338 0.003808501
339 0.003808501
340 0.003808501
341 0.003808501
342 0.003808501
343 0.000000000
344 0.003808501
345 0.003808501
346 0.003808501
347 0.003808501
348 0.003808501
349 0.003808501
350 0.003808501
351 -0.005034564
352 0.003808501
353 0.003808501
354 0.003808501
355 0.003808501
356 0.003808501
357 0.003808501
358 0.003808501
359 0.003808501
360 0.003808501
361 0.003808501
362 0.003808501
363 0.003808501
364 0.000000000
365 0.003808501
366 0.003808501
367 0.003808501
368 0.000000000
369 0.003808501
370 0.003808501
371 0.000000000
372 0.003808501
373 0.003808501
374 0.003808501
375 0.003808501
376 0.003808501
377 0.003808501
378 0.000000000
379 0.003808501
380 0.003808501
381 0.003808501
382 -0.002227830
383 0.000000000
384 0.003808501
385 0.003808501
386 0.003808501
387 0.000000000
388 0.003808501
389 0.003808501

```

## 390 0.003808501
## 391 0.003808501
## 392 0.003808501
## 393 0.003808501
## 394 -0.005034564
## 395 -0.005034564
## 396 0.000000000
## 397 0.003808501
## 398 0.003808501
## 399 0.003808501
## 400 0.003808501
## 401 0.000000000
## 402 0.000000000
## 403 0.003808501
## 404 0.000000000
## 405 0.003808501
## 406 0.003808501
## 407 0.003808501
## 408 0.003808501
## 409 0.003808501
## 410 0.003808501
## 411 0.003808501
## 412 0.003808501
## 413 0.000000000
## 414 0.003808501
## 415 0.003808501
## 416 0.003808501
## 417 0.000000000
## 418 0.003808501
## 419 0.003808501
## 420 0.003808501
## 421 0.003808501
## 422 0.003808501
## 423 0.003808501
## 424 0.003808501
## 425 0.003808501
## 426 0.003808501
## 427 0.003808501
## 428 0.003808501
## 429 0.000000000
## 430 0.003808501
## 431 0.003808501
## 432 0.003808501
## 433 0.003808501
## 434 0.003808501
## 435 0.000000000
## 436 0.000000000
## 437 0.000000000
## 438 0.000000000
## 439 0.000000000

```

Using bigrams to do sentiments analysis. If we do seperate analysis on both article about “not” words.

```

## # A tibble: 1 x 4
##   word2 score      n contribution

```

```
##      <chr> <int> <int>      <int>
## 1 true      2      3          6
```

In this P-Value article, only one word is followed by “not”.

```
## # A tibble: 0 x 4
## #   ... with 4 variables: word2 <chr>, score <int>, n <int>,
## #     contribution <int>
```

And in this Data to Story article, on word is followed by “not”.

Network of Bigrams

```
##
## Attaching package: 'igraph'

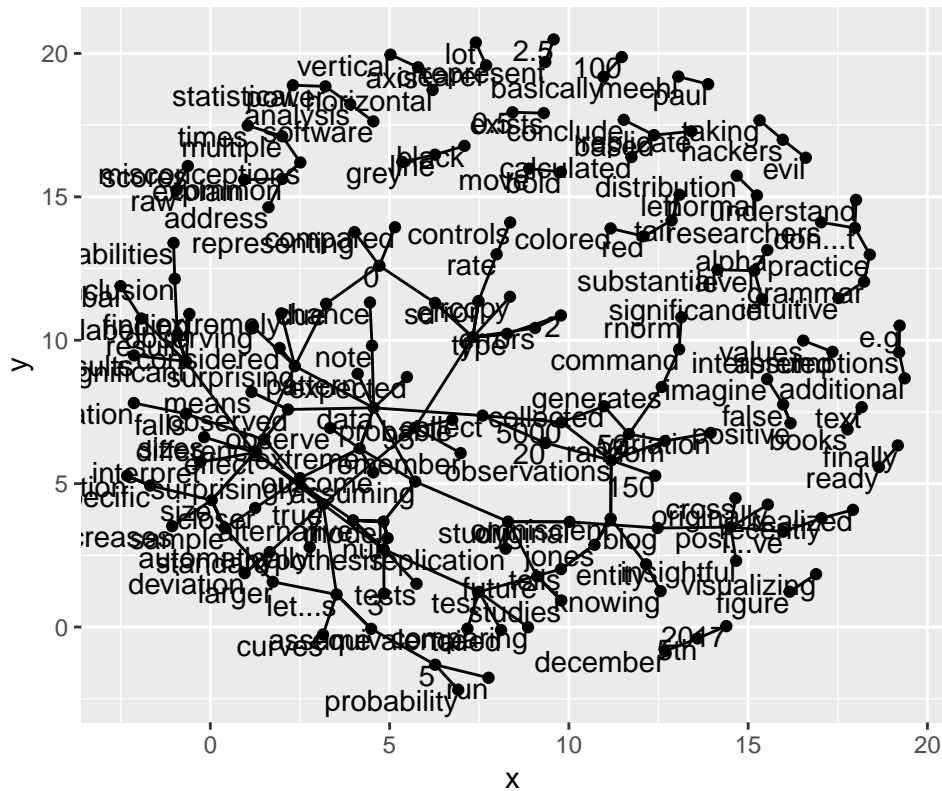
## The following object is masked from 'package:tidyr':
##
##     crossing

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

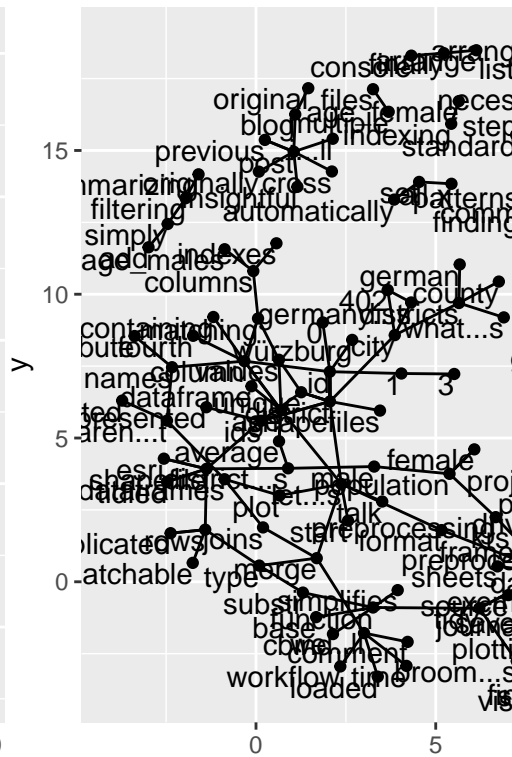
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

Network Plot for P-Value Article



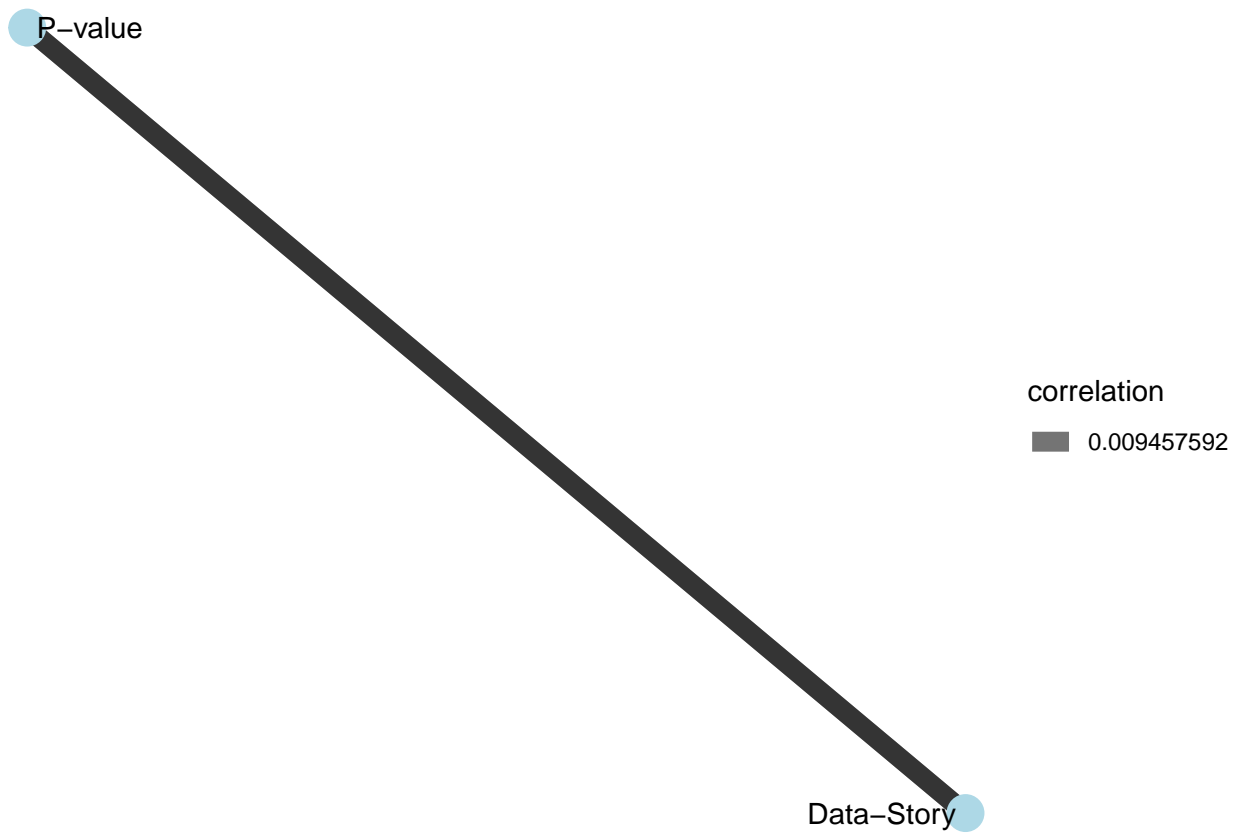
Network Plot for Data to Sto



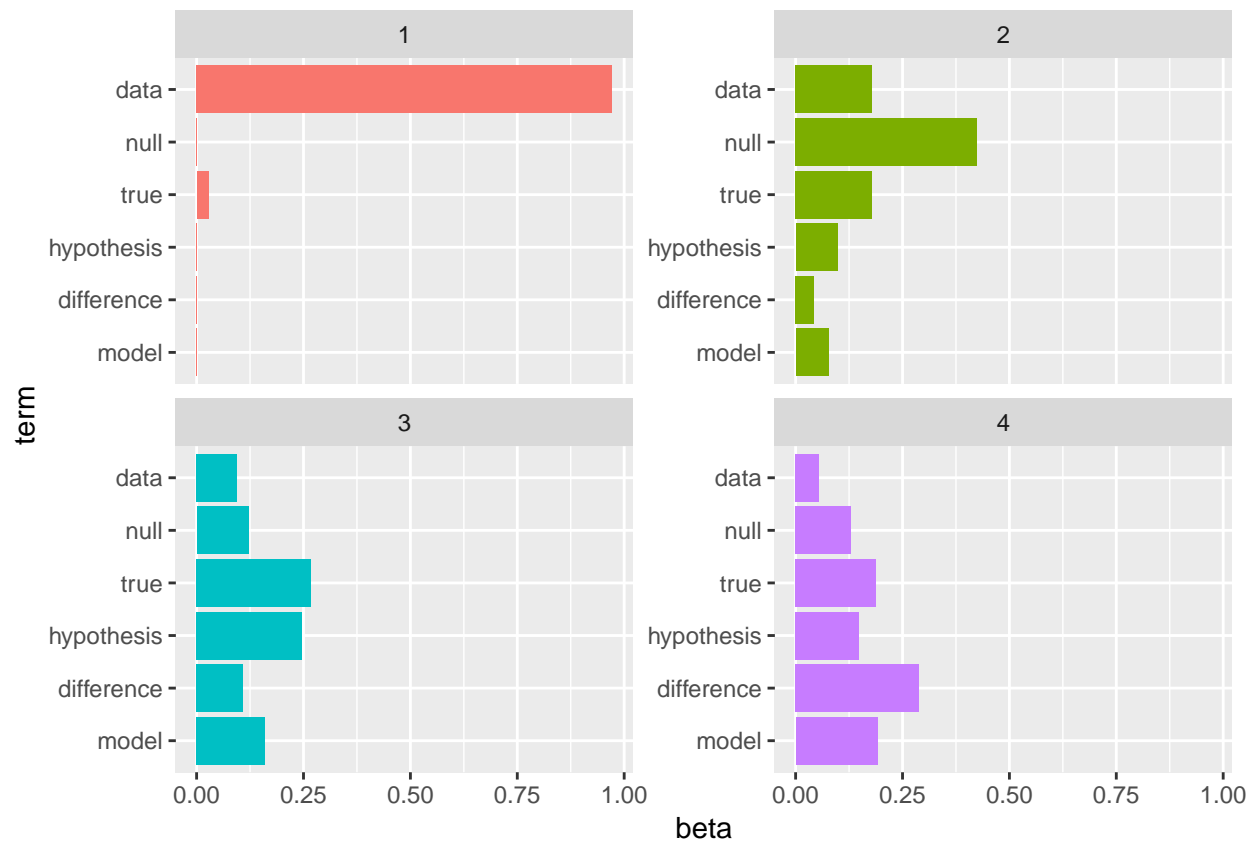
```
## # A tibble: 103 x 4
## # Groups:   word [?]
##   word          n article sentiment
##   <chr>        <int> <chr>   <chr>
## 1 critical      1 P-value negative
## 2 clearer       1 P-value positive
## 3 prefer        1 P-value positive
## 4 easy          1 P-value positive
## 5 reject        1 P-value negative
## 6 evil          1 P-value negative
## 7 misconception 1 P-value negative
## 8 sneaky        1 P-value negative
## 9 broken        1 P-value negative
## 10 correctly    1 P-value positive
## # ... with 93 more rows

## # A tibble: 1,433 x 6
## # Groups:   word [524]
##   word          n article      tf   idf   tf_idf
##   <chr>        <int> <chr>    <dbl> <dbl>   <dbl>
## 1 tutorial      1 Data-Story 0.000341 0.693 0.000237
## 2 steps         1 Data-Story 0.000341 0.693 0.000237
## 3 driven        1 Data-Story 0.000341 0.693 0.000237
## 4 morgenpost    1 Data-Story 0.000341 0.693 0.000237
## 5 exemplary     1 Data-Story 0.000341 0.693 0.000237
## 6 journalism    1 Data-Story 0.000341 0.693 0.000237
## 7 workflow      1 Data-Story 0.000341 0.693 0.000237
## 8 bbsr          1 Data-Story 0.000341 0.693 0.000237
## 9 commented     1 Data-Story 0.000341 0.693 0.000237
## 10 organized    1 Data-Story 0.000341 0.693 0.000237
## # ... with 1,423 more rows

## # A tibble: 2 x 3
##   item1      item2      correlation
##   <chr>    <chr>          <dbl>
## 1 Data-Story P-value      0.00946
## 2 P-value   Data-Story    0.00946
```



```
## # A tibble: 211 x 4
##   word      n article word_total
##   <chr> <int> <chr>      <int>
## 1 data      18 P-value        53
## 2 data      18 P-value        53
## 3 data      18 P-value        53
## 4 data      18 P-value        53
## 5 data      18 P-value        53
## 6 data      18 P-value        53
## 7 data      18 P-value        53
## 8 data      18 P-value        53
## 9 data      18 P-value        53
## 10 data     18 P-value        53
## # ... with 201 more rows
```



We can see that we have the same common words amongst all of the topics in the LDA.