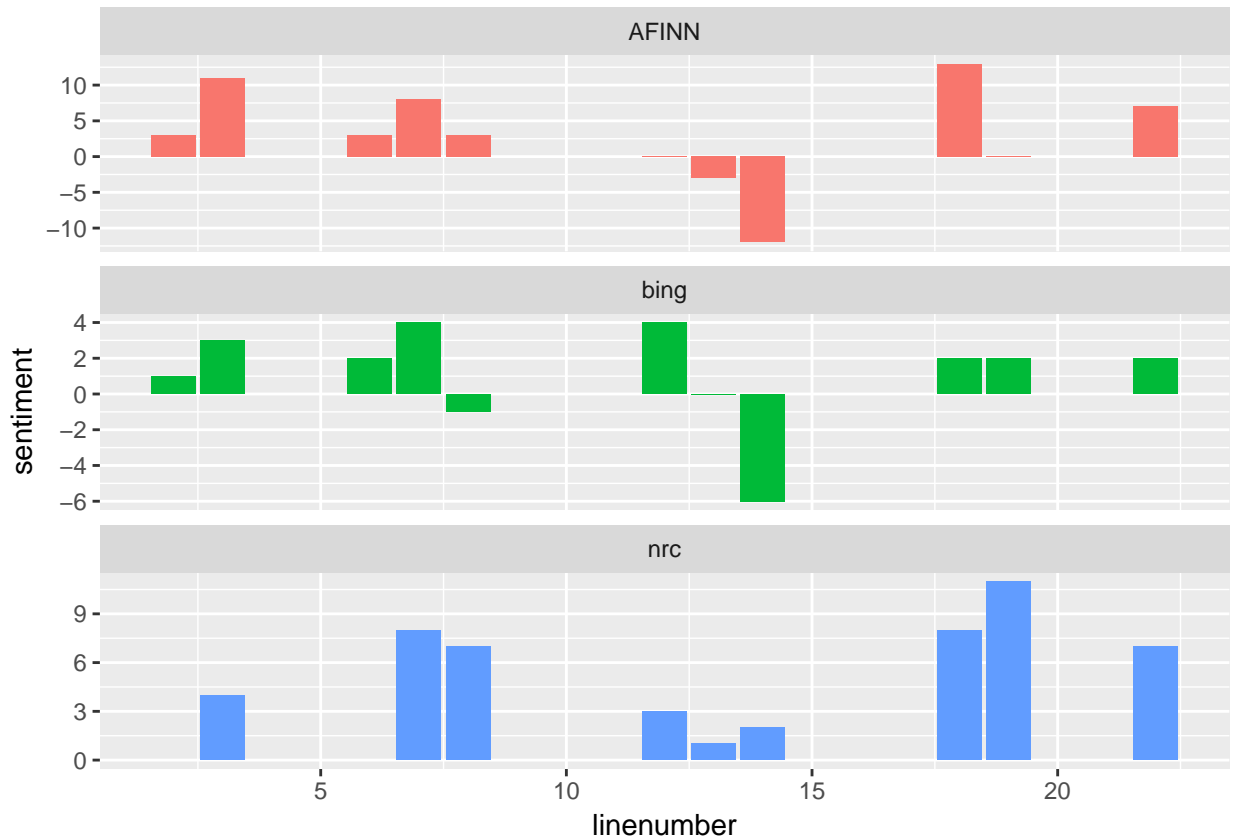


# Textming

*Siwei Hu, Yifeng Luo, Hao Qin, David Anderson*

*November 4, 2018*

```
## Loading required package: xml2
##
## Attaching package: 'rvest'
## The following object is masked from 'package:purrr':
##
##   pluck
## The following object is masked from 'package:readr':
##
##   guess_encoding
## Joining, by = "word"
## Joining, by = "word"
## Joining, by = "word"
## Joining, by = "word"
## Joining, by = "word"
```

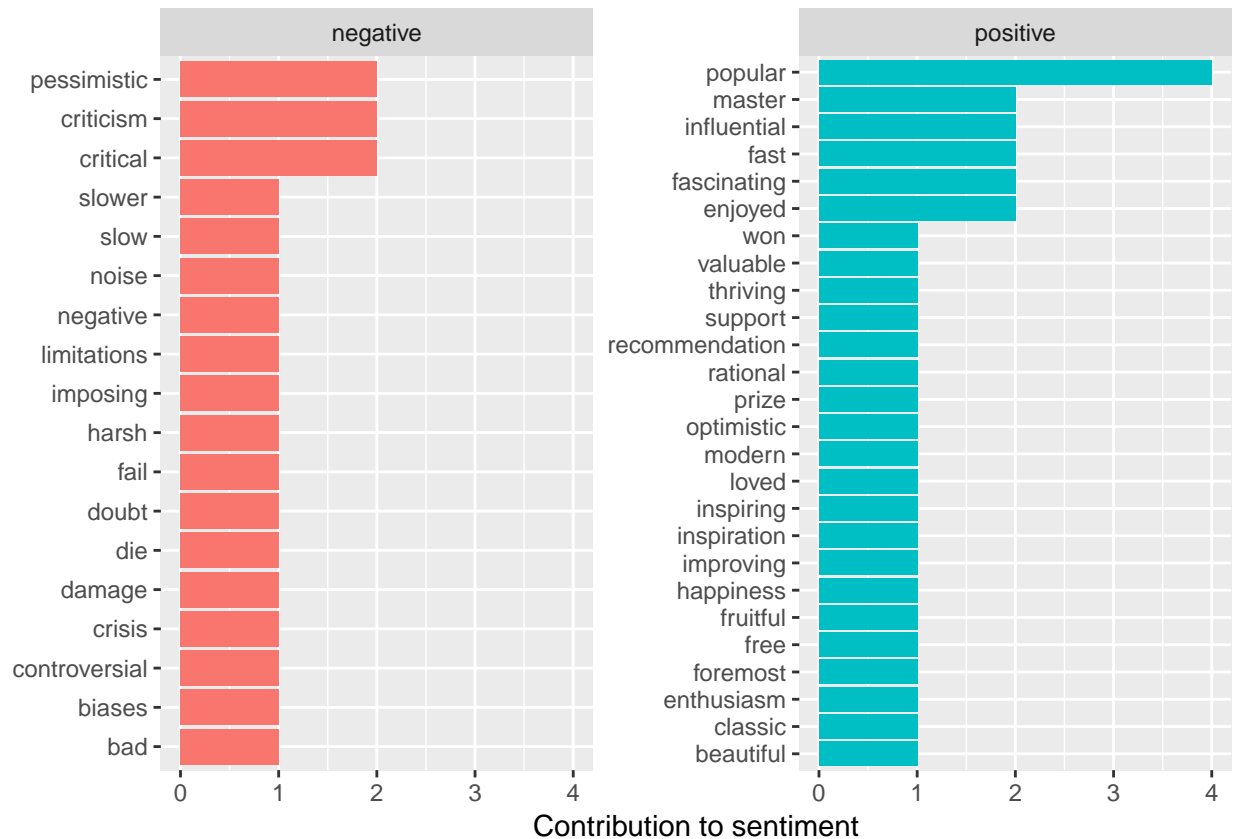


Sentiment analysis for these paragraphs: From AFINN and bing, they have really similar sentiment fluctuate(positive,negative and neutral). since nrc recognize the words as “fear”,“happy”,“postive” and so on, so it will not easily to decide a word is positive or negative. However, for this data science article, we do not

need too much different emotion word to justify it. Positive or negative is better. So i think Bing and AFINN can simply display the sentiment of the article. Basically, it displayed a positive emotion to the audience.

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

```
## Selecting by n
```



These two charts provide us with the top 10 words contributing to sentiment. Because it's a short article, so many words only appeared once, making the top 10 rank less useful. However, we can still identify the most significant words. In the positive chart, 'popular' was used 4 times. I can understand this because it's an article recommending a book for data scientists. In the negative chart, 'pessimistic', 'criticism', and 'critical' all appeared twice.

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

```
##
```

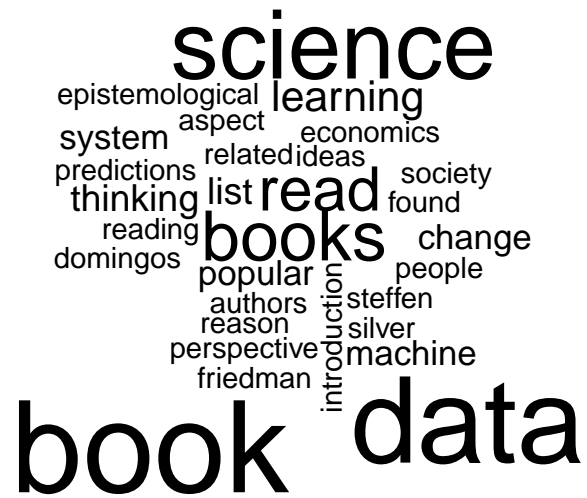
```
## Attaching package: 'reshape2'
```

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

```
##
```

```
## smiths
```

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

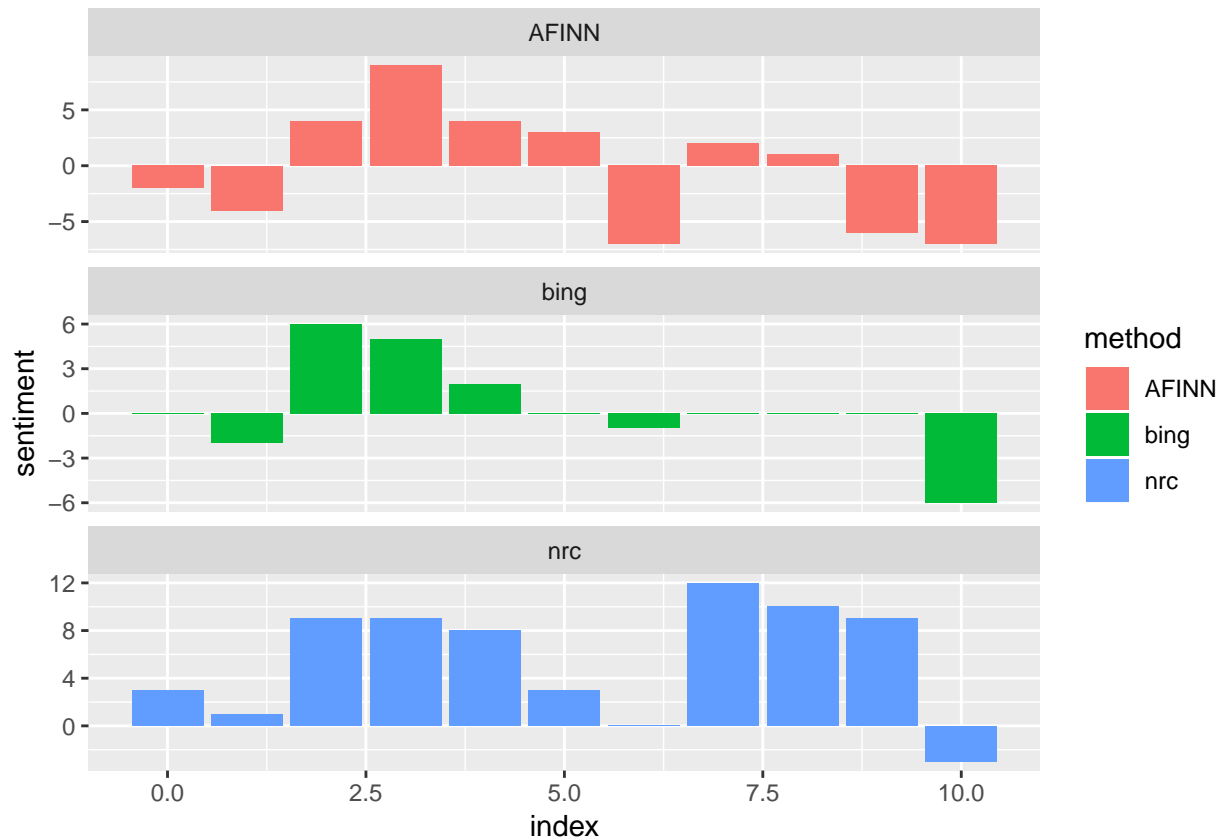


```
## Using n as value column: use value.var to override.
```



i draw  
two wordcloud to help my audience to understand what is the high frequency words and which side they belong.

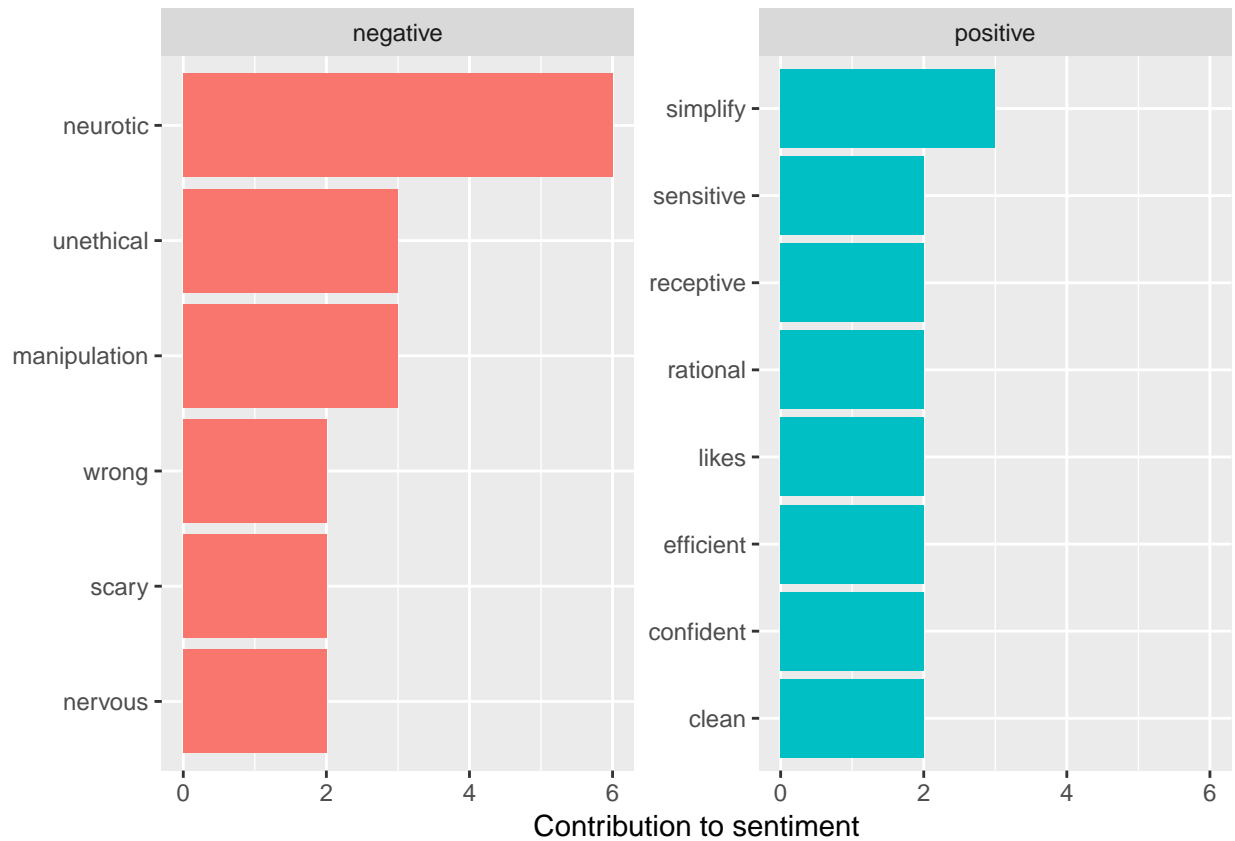
```
## Joining, by = "word"  
## Joining, by = "word"  
## Joining, by = "word"  
## Joining, by = "word"  
## Joining, by = "word"
```



In the article “RECONSTRUCTING CAMBRIDGE ANALYTICA’S PSYCHOLOGICAL WARFARE TOOL”, I set each two paragraphs give us a value of sentiment. The AFINN method gave us the clearest influence of emotion, there is no blank (which means the value equal to 0 which is neutral) for each two paragraphs. Bing method shows several natural paragraphs and two extreme values positive 6 and negative -7. The analysis of AFINN and Bing is similar except the paragraphs 15 and 16. However, the method NRC gives a very positive tendency of sentiment for this article. From all three methods, the first 10 paragraphs, the author displays a positive attitude, however, at the end of the article, he showed the audience a negative attitude.

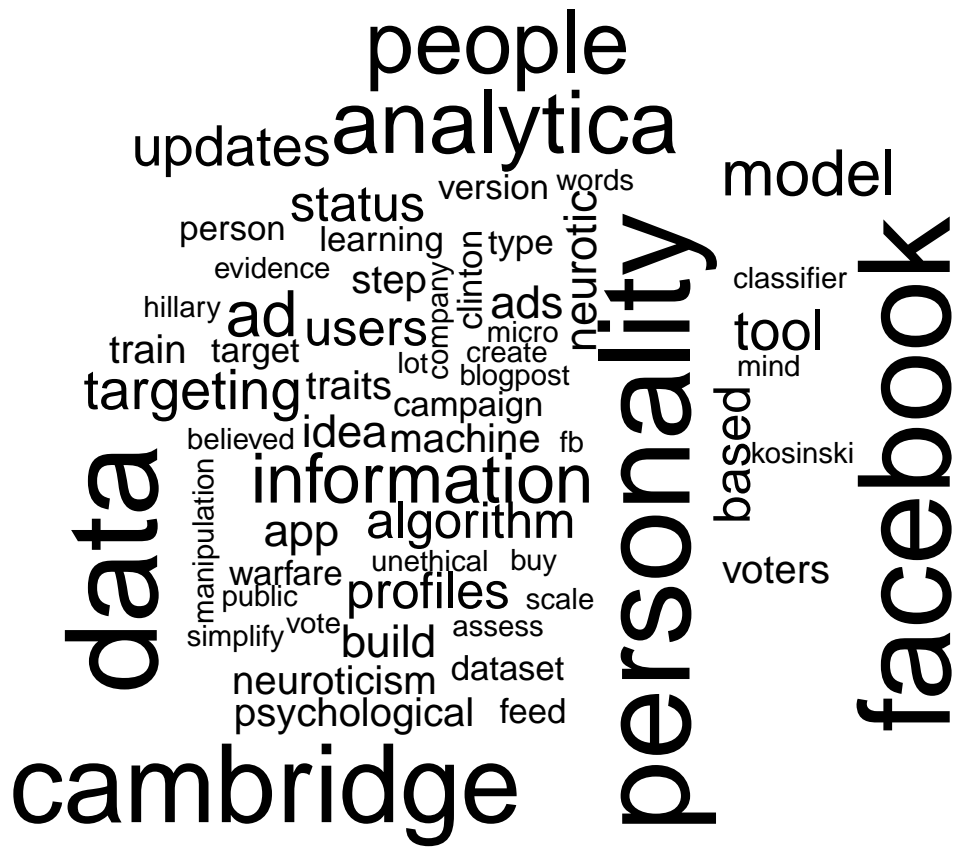
```
## Joining, by = "word"
```

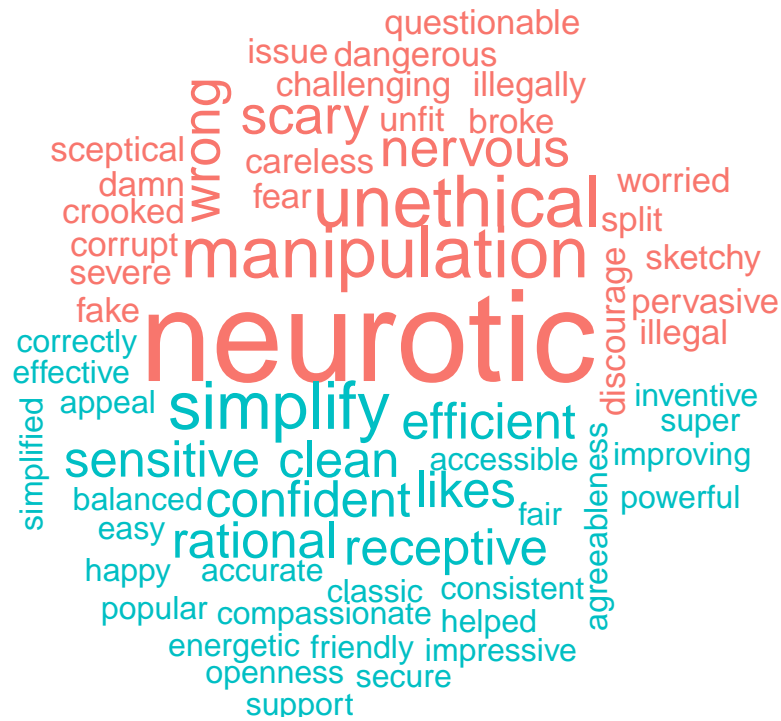
```
## Selecting by n
```



From two chart, neurotic is the most negative word in article. Then word “unethical” and “manipulation” both appeared 3 times. These word helped author to explain about how facebook use their impact to affect people during the election. The most word in postive part actually is trump. In the article, it means the president. So i filter this word which is error in this analysis. So we can see the positive words here. From two word frequency charts, the negative words display stronger emotion than positive words do.

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

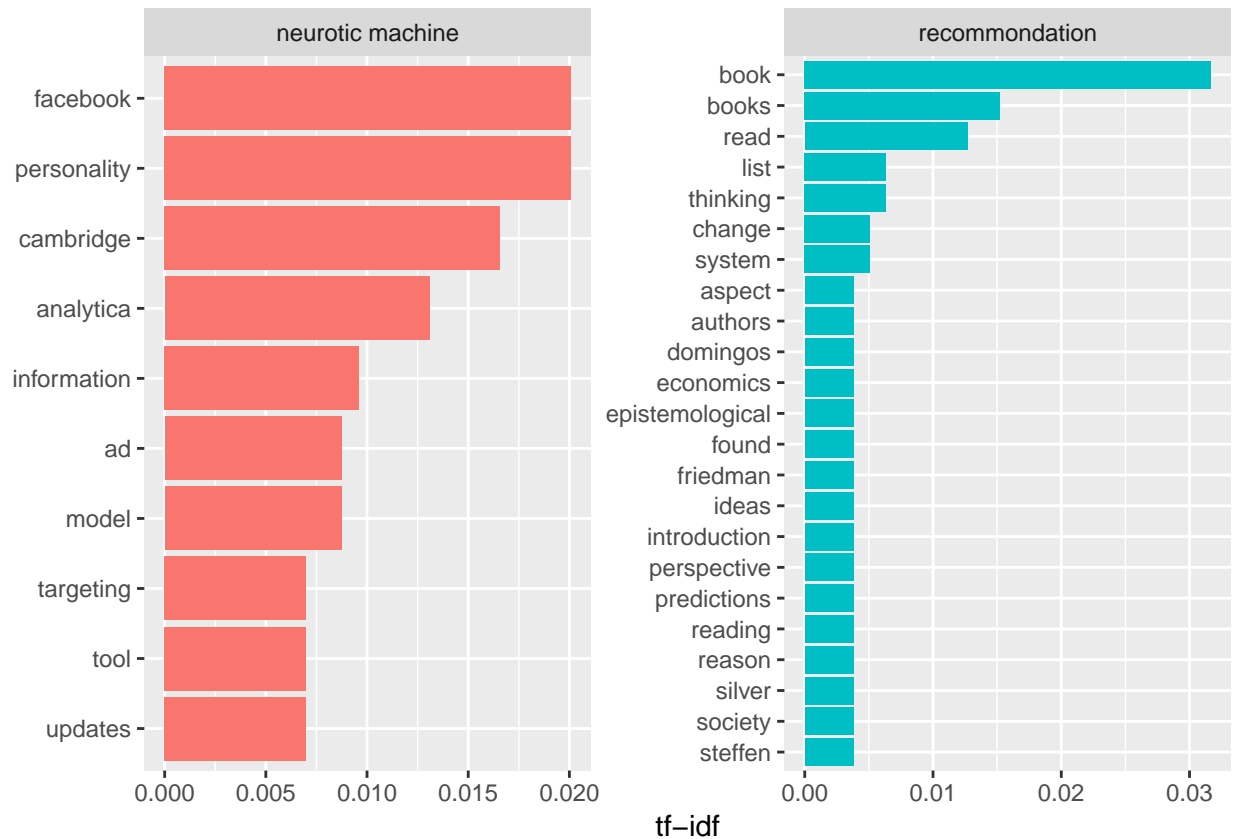




The wordcloud also help to support what i said in the above word frequency chars. The negative word display higher frequency and stronger emotion.

## Selecting by tf\_idf





tf\_idf charts showed us which words are more important in one article. In the article “RECONSTRUCTING CAMBRIDGE ANALYTICA’S”PSYCHOLOGICAL WARFARE TOOL“, company name and words like”information“,”ad” are more important. These are core vocabulary for this article.They can help people to understand what happen easier. In the article “DATA SCIENCE IS NOT JUST ABOUT DATA SCIENCE”, book and books become the core words. I agree with this because the subtitle of it is “7 POPULAR SCIENCE BOOKS THAT YOU HAVE TO READ”.