Textmining exercise

Xuan,Megha,Yifu,Sky

November 4, 2018

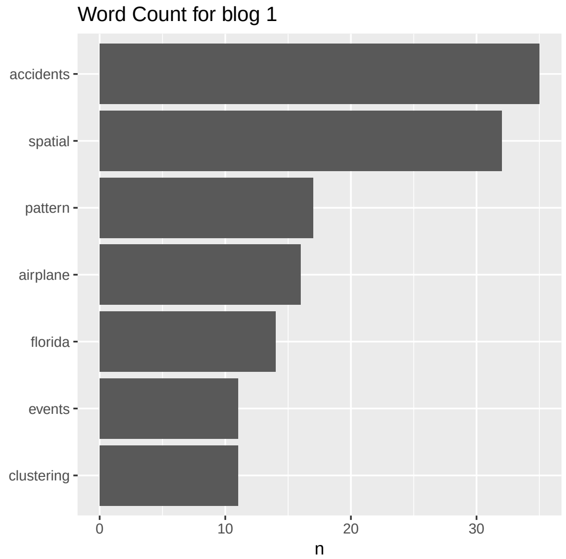
SCRAPING DATA FROM https://correlaid.org/blog/

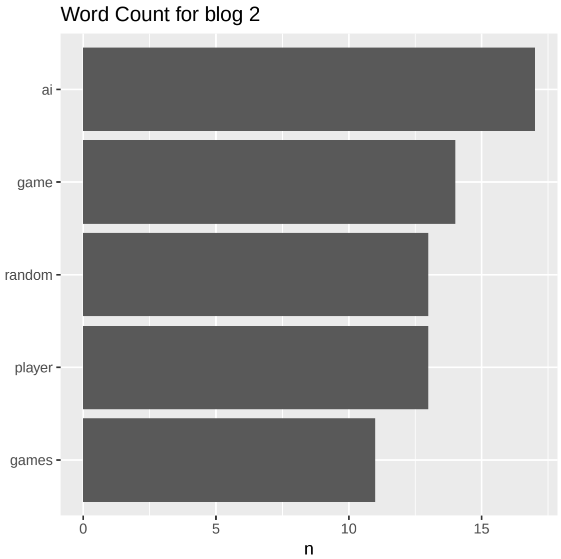
The task of this project is to select two blog entries and conduct text mining based on them.

Our first selected blog entry is about a point pattern analysis on an airplane accident in Florida. The second blog entry is about an AI model developed for game Tic Tac Toe.

GET A TIDY TEXT FORMAT & WORD COUNT

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

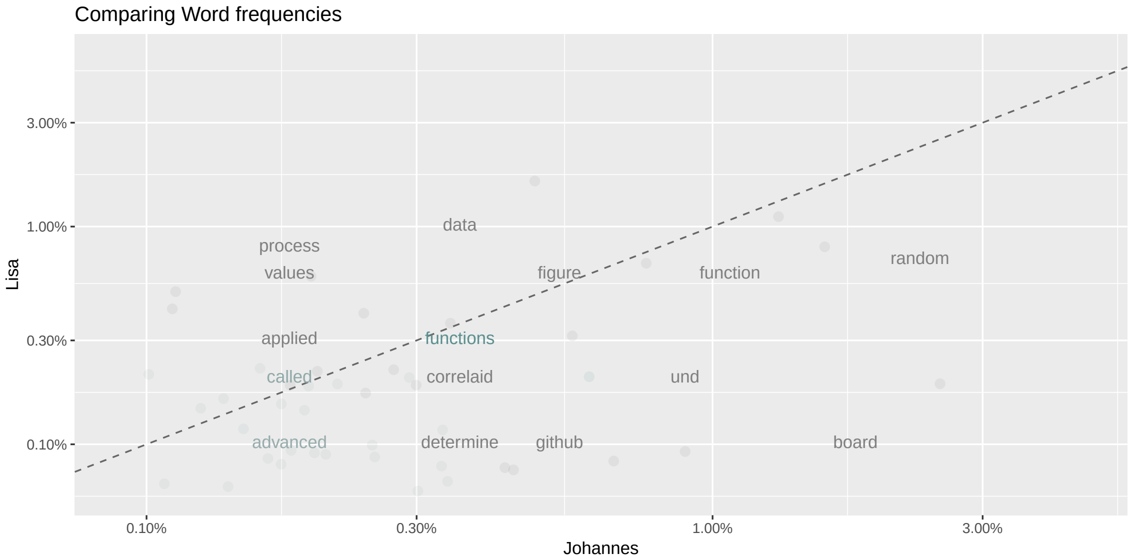




After cleaning the texts into a tidy text format, getting rid of non characters and stop words, we did word count analysis on both blogs, which gives us the words from each text with the highest frequency.

From the plot "Word Count for blog 1" we can see that, the most frequently mentioned words are "accidents", "spatial", "pattern", "airplane", "florida", etc. Those are the keywords that describe the main topic of this blog entry.

Similarly, with "Word Count for blog 2", the top words are "ai", "game", "random", etc. That is because this blog is mainly about an AI developed for a game.

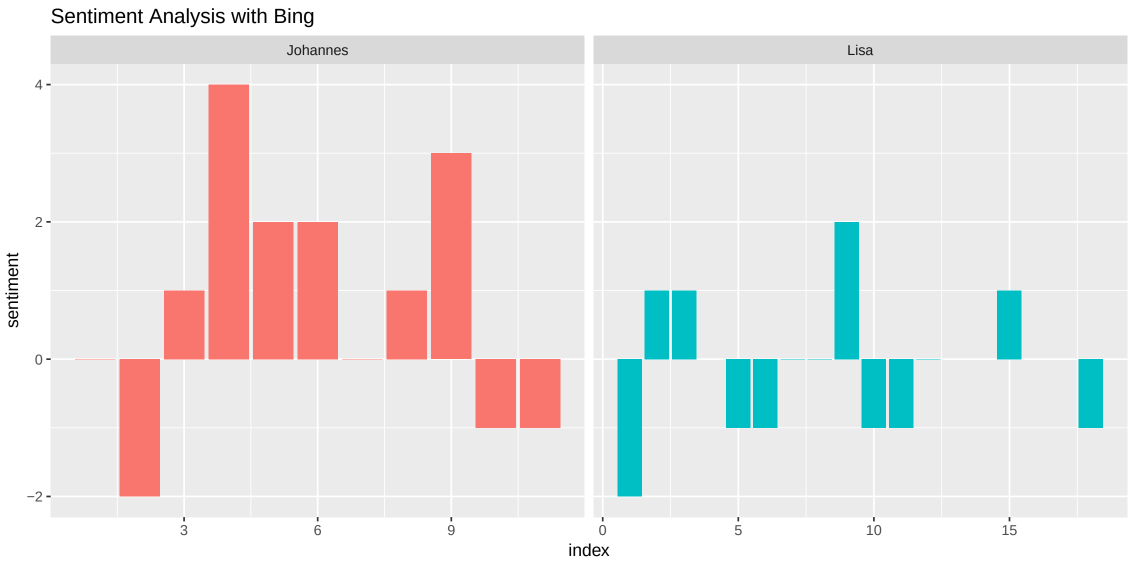


This plot compares the word frequencies from both blogs. The X axis "Johannes" refers to the author of blog 2 while the Y axis "Lisa" refers to the author of blog 1.

From the plot we could see that both blogs use words "function", "figure" at the same frequency level, largely due to the fact that both of them are science blogs. Blog 1 uses words "data", "values" more often than blog 2 while blog 2 uses words random, board more frequently than blog 1. We think this is because blog 1 is about the analysis based on a case with some data, while blog 2 is about AI program that plays a game with more randomness during the development process.

Sentiment Analysis With Tidy Data

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

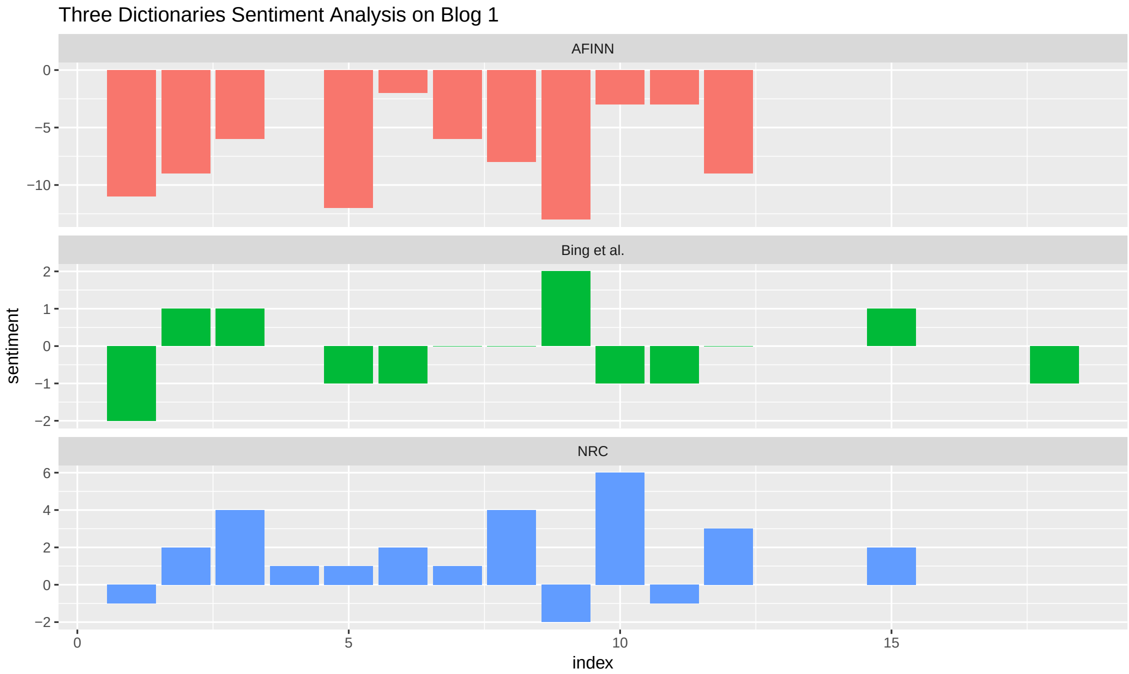


The sentiment analysis based on dictionary bing shows that blog 2 with author Johannes uses more positive words while blog 1 is more balanced with positive and negative words.

This makes sense since blog 1 is an analysis based on a case study while blog 2 is building an AI model playing a game, which seems a lot more fun.

Comparing the three sentiment dictionaries

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



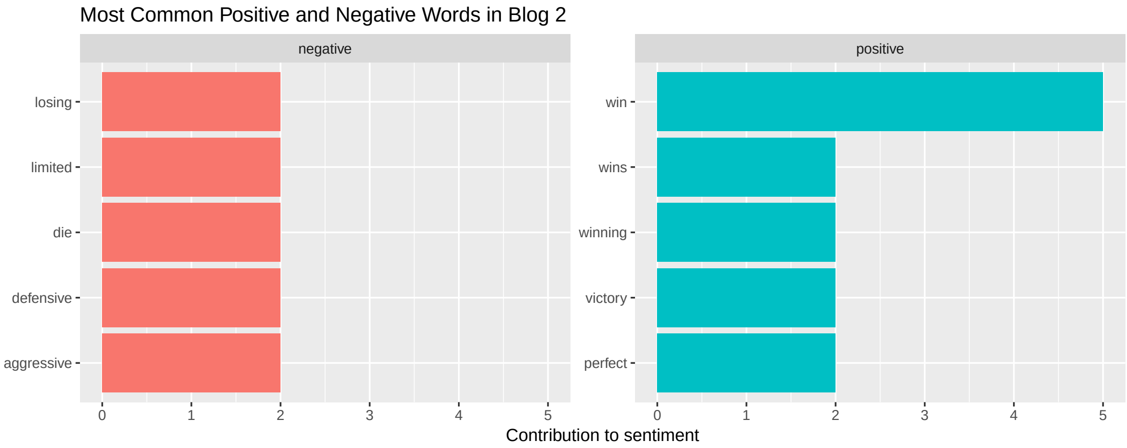
To compare Dictionary Bing with two other dictionaries AFINN AND NRC, we used blog 1 as an example.

From the plot, we could see that the analysis using AFINN dictionary gives most negative results, the analysis using NRC dictionary gives mostly positive results, while the analysis using Bing gives a result balanced from positive and negative.

Most common positive and negative words

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

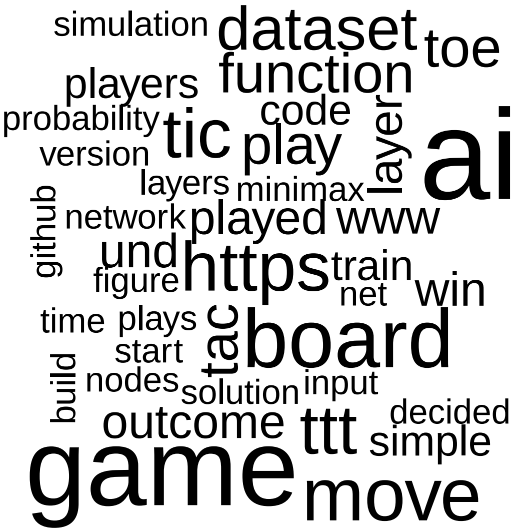
```
## Selecting by n
```



The result of the most common positive and negative words in blog 2 is obvious. Since blog 2 is about gaming AI, the common positive words would be "win" and the common negative words would be "lose" or "die" with no doubt.

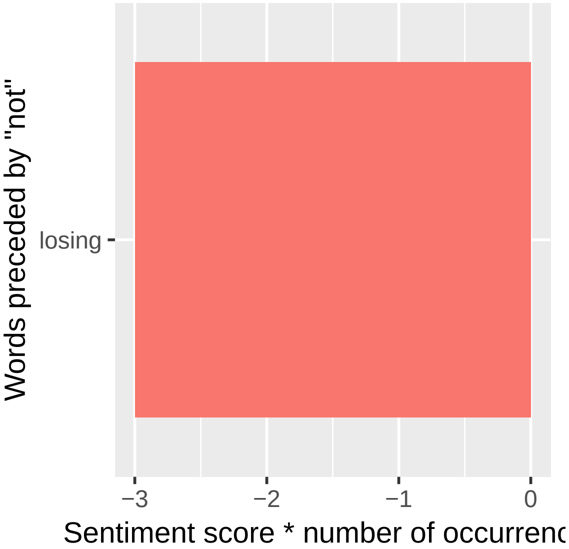
Word Cloud

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



The word cloud analysis also gives us a sense of what the blog is mainly about. Take blog 2 as an example. We can clearly see that the blog is about a gaming AI model.

Bigram sentiment Analysis



To take a further look at sentiment analysis, we took blog 2 to look at bigrams. Since the blog isn't very long, we can see from the plot that the bigram begining with "not" is mostly not winning, which is of course a negative bigram.