Tom Sawyer Sentiment Analysis Part 1

Carolyn Wright

Figure 1: Tom Sawyer

Data Preparation

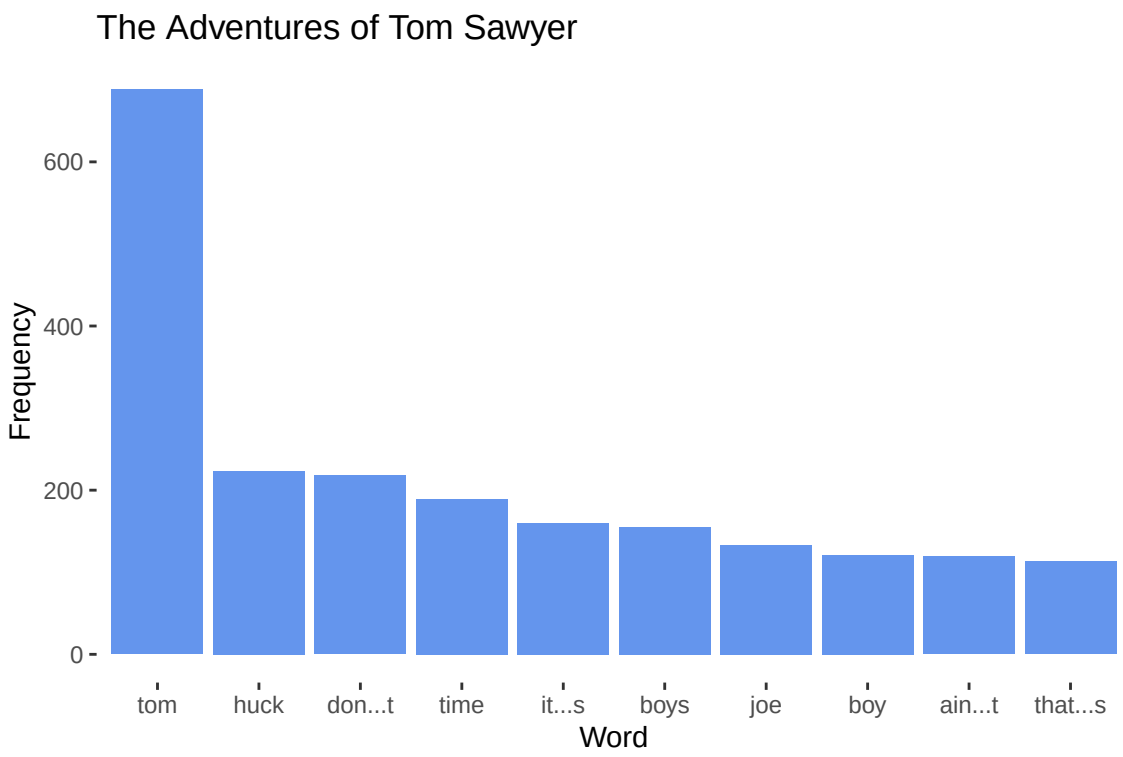
In the initial steps of this analysis I did some cleaning of the book file in order to make easier to work with. After reading in the book from Gutenberger, I manually removed the table of contents, along with any stop words throughout. My next step was to create a chapter indicator in order to do a sequential analysis of Tom Sawyer. This left me with a result that looked like the following.

	line	linenumber	chapter	word
1	460	1	1	chapter
2	460	1	1	i
3	463	4	1	tom
4	465	6	1	no
5	465	6	1	answer
6	467	8	1	tom
7	469	10	1	no
8	469	10	1	answer
9	471	12	1	what's
10	471	12	1	gone
11	471	12	1	with
12	471	12	1	that
13	471	12	1	boy
14	471	12	1	i

Figure 2: Data Screenshot

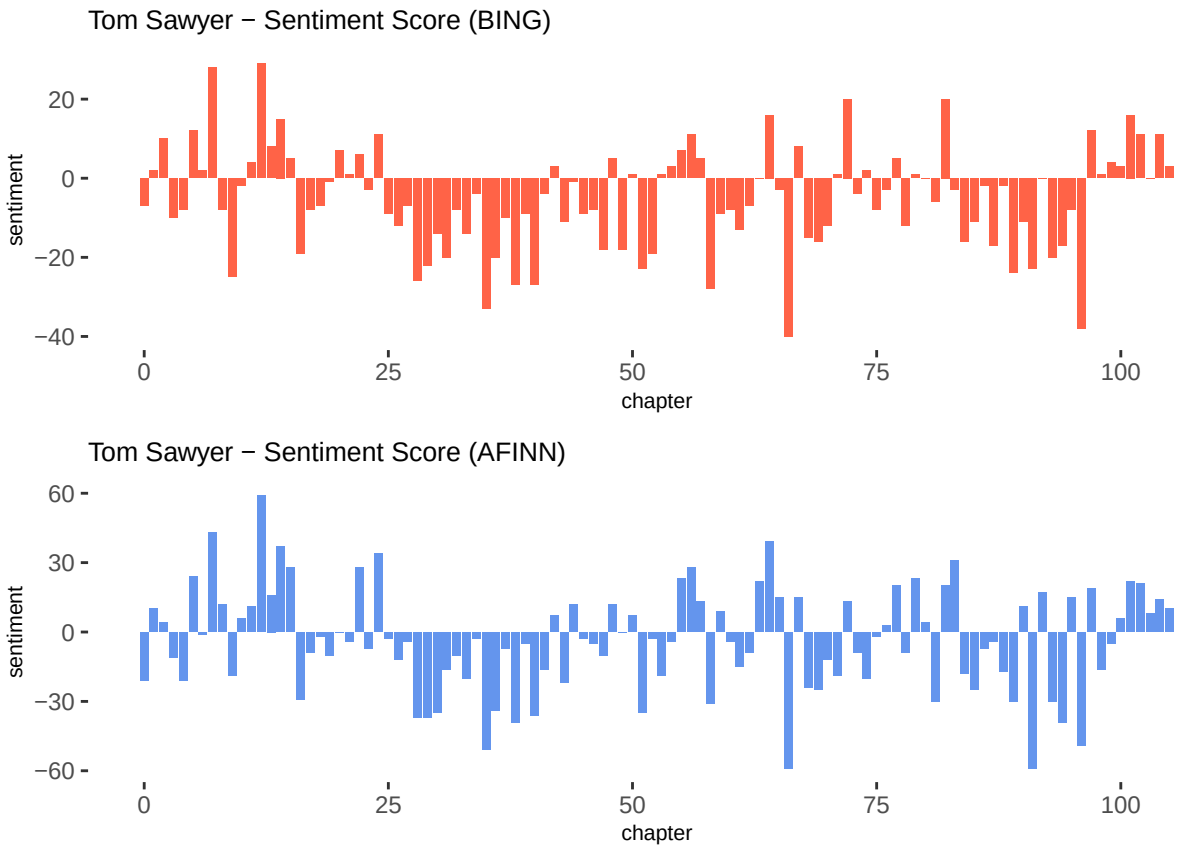
Most Commonly Used Words

From the plot below, one can see that the most commonly used words are "Tom" and "Huck". This indicates that these two characters are two of the main characters in the book. This is not all too surprising considering the name of the book.



Sentiment Analysis

The following plots show the sentiment of The Adventures of Tom Sawyer using two different lexicons, Bing and Afinn. The analysis is done at a chapter level. As one can see the two lexicons do not behave identically. The AFINN lexicon picks up on more fluctuations in the sentiment throughout the book than the BING lexicon does. However, looking at both of these analysis it is hard glean any real knowledge about what is going on in the book. We can potentially see where is conflict or something negative going on, but it does not help up to understand a more detailed level of the plot line.


Tom Sawyer − Sentiment Score (BING)
sentiment
20
0
−20
−40
0
25
50
75
100
chapter

Tom Sawyer − Sentiment Score (AFINN)
sentiment
60
30
0
−30
−60
0
25
50
75
100
chapter