# Token-level Sentiment Analysis of [A Study in Scarlet]

XihaoCao

2021/12/6

## Introduction

A Study in Scarlet is an 1887 detective novel written by Arthur Conan Doyle. The story marks the first appearance of Sherlock Holmes and Dr. Watson. Conan Doyle wrote the novel at the age of 27, and the story attracted little public interest when it first appeared. But after Sherlock Holmes and Dr. Watson become one of the most famous and iconic literary detective characters, this story gets more and more popular. One thing needs notice is that this book is one of only four full-length novels in the original canon.

## Synopsis

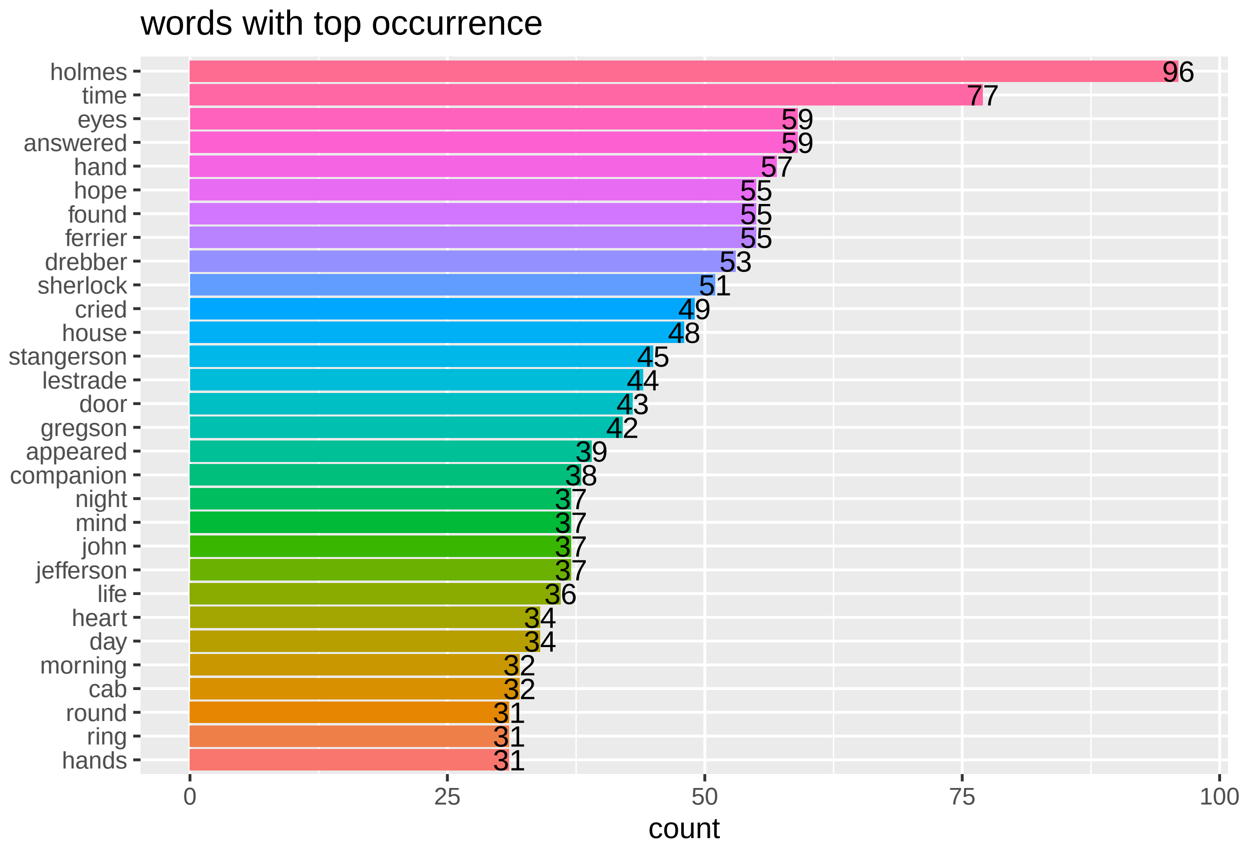
The novel is split into two quite separate halves. The first part is told in first person by Sherlock Holmes' friend Dr Watson, and describes his introduction in 1881 to Holmes through a mutual friend and the first mystery in which he followed Holmes' investigations. The mystery revolves around a corpse found at a derelict house in Briton, London with the word "RACHE" scrawled in blood on the wall beside the body. After detailed investigation and thorough detection, Holmes gets the clues and reveal the story behind the crime which is told in the second part of the book. The second half of the story is called The Country of the Saints and jumps to the United States of America and the Mormon community, and incorporating a depiction of the Danites, including an appearance by Brigham Young in a somewhat villainous context. It is told in a third person narrative style, with an omniscient narrator, before returning in the last two chapters to Watson's account of Holmes' investigation, and then Holmes own explanation of his solution. In these two chapters the relationship between the two halves of the novel becomes apparent. The motive for the crime is essentially one of lost love and revenge.

## Data cleaning and organization

I download the book from the Gutenberg Book Project using the built-in function of Gutenberg package. The initial data is stored in tibble on paragraph level, and it only contains the text information. Then I tokenize the paragraph while adding the number of line and chapter each single word lives in, and deleting all 'stop words'. Thus the final organized data is on token-level without any 'stop words', ex: the, a, an, that, etc..

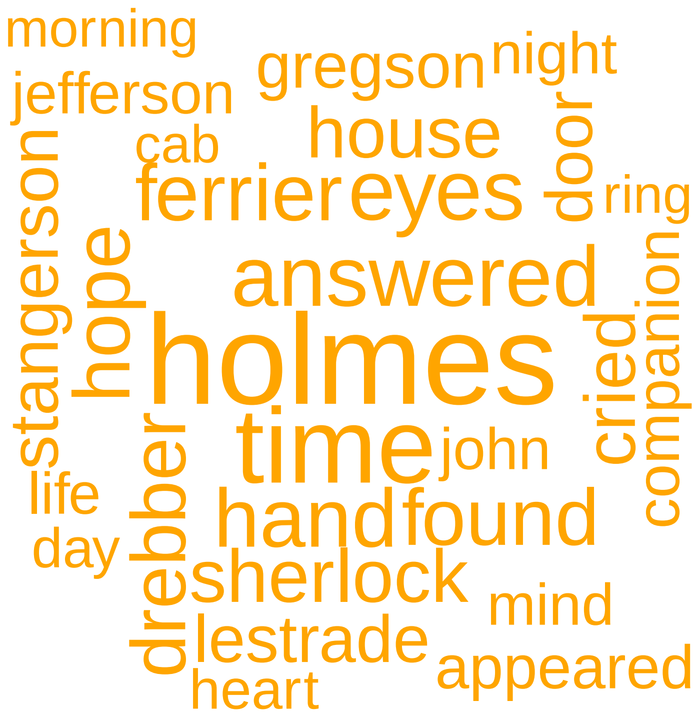
## Words Occurrence Frequencies

In order to explore which words show up most frequently in the story, I plot a bar chart illustrating words with at least 30 occurrences.

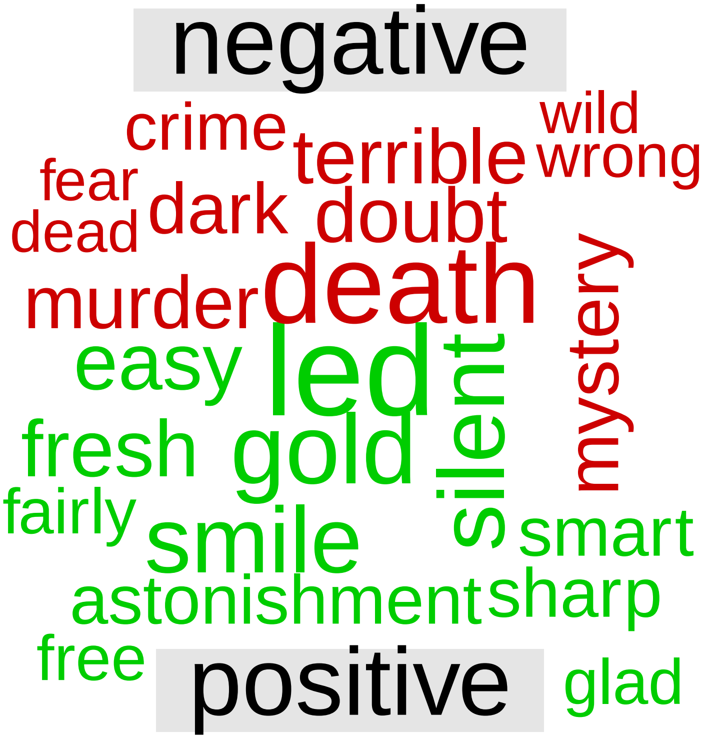


## Cloud Graphs

I also use the cloud graph to illustrate the word occurrence. Bigger the word is in the first cloud graph, more occurrences it has in the story. And we can see that this graph is completely consistent with the bar chart above.



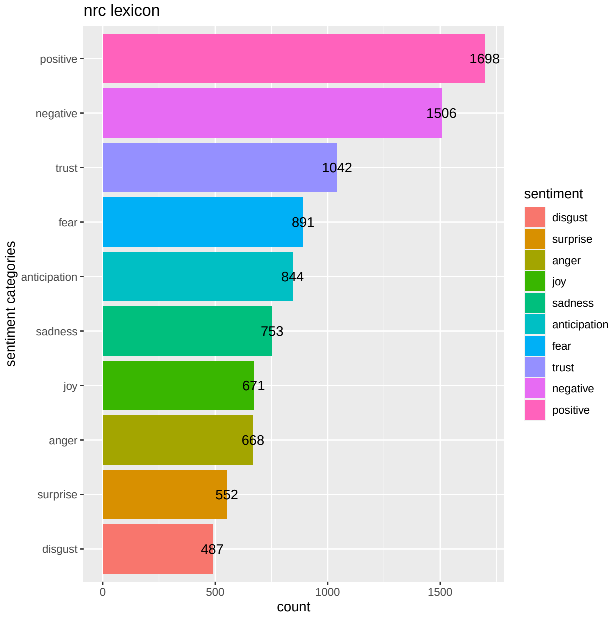
The second cloud graph uses the Bing lexicon to classify words into positive and negative types, and words that cannot be matched up with the Bing lexicon will be dropped. The upper part are all negative words, and the lower part are negative, and same as the first cloud graph, bigger the word is, more occurrences it has in the story.

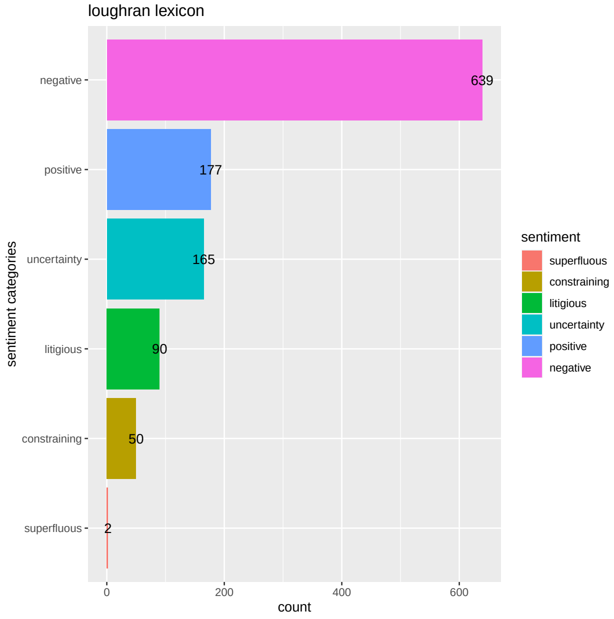


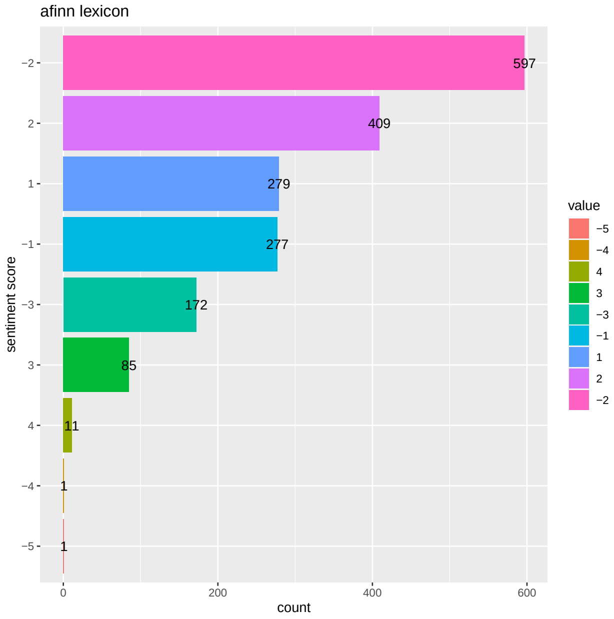
## Size of sentiment categories in each lexicon engine

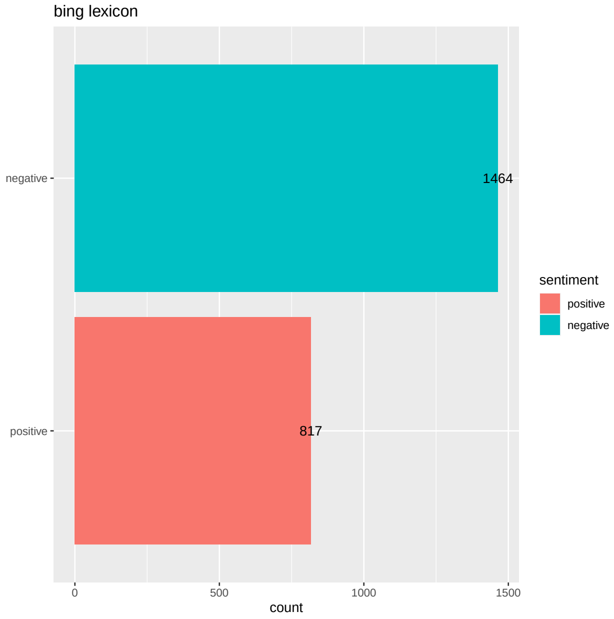
I choose four lexicon engines to analyze, they are one of the most popular and common-used engine for the text mining, they are afinn, NRC, bing, and loughran. Then I use the bar charts to illustrate the size of sentiment categories in each four lexicon engine. Both NRC lexicon and Loughran lexicon have multiple sentiment categories, while Bing lexicon only has positive, negative two categories, and afinn lexicon uses integers range from -5 to 4 to quantify how positive a word is.

We can see that except for the NRC lexicon, all other three are basically consistent which show that there are more negative words than positive ones in this story. However, since the NRC lexicon have multiple categories, where fear, sadness, anger can also be considered negative in other lexicons, I think the NRC is also consistent with other three engines.







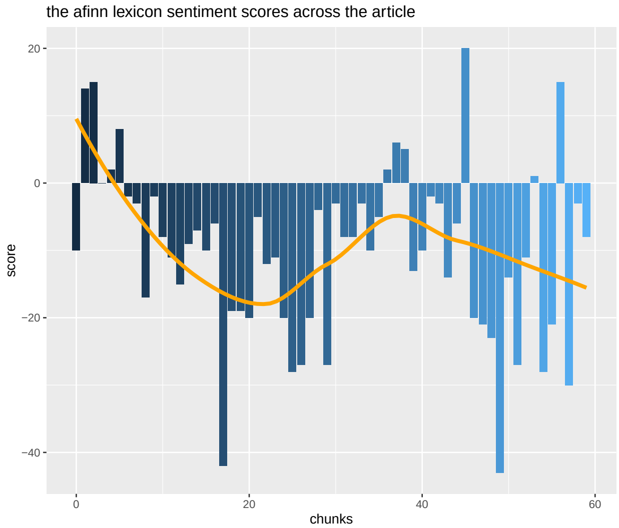


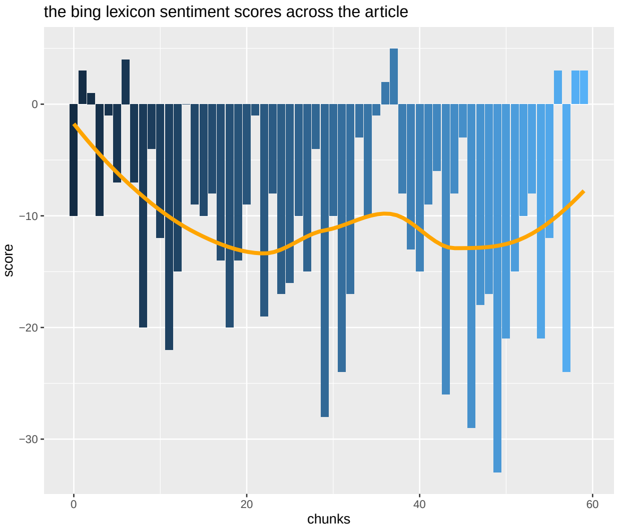
## Sentiment across the story

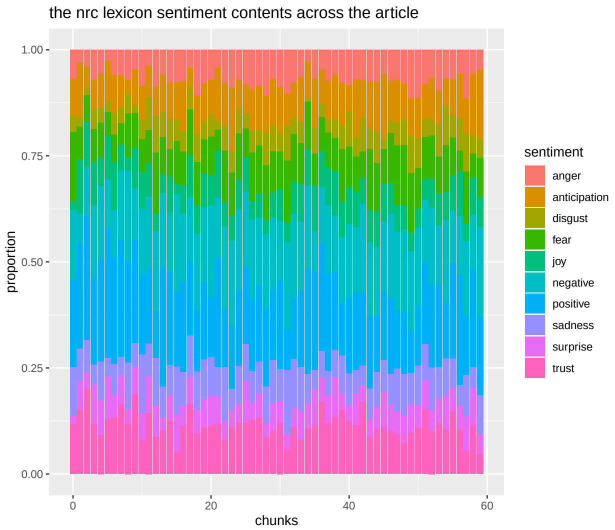
Then I try to compare the four different lexicon engines across the story. Since the data now are on the word-level, I manually group each 80 consecutive words into a chunk, and then evaluate the sentiment score/proportion of chunks across the story.

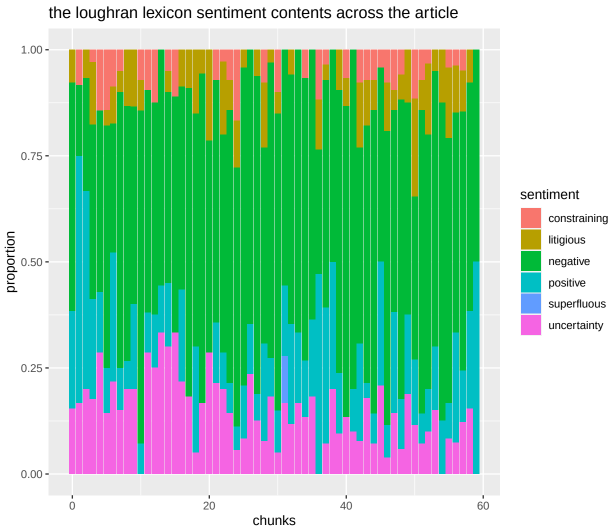
Both affine and Bing lexicon use numerical value to quantify how positive a word is, I can set a variable called sentiment score which is the sum of the sentiment values of all words in a chunk. While Loughran and nrc have multiple sentiment categories, and there is not a numerical value we can use to evaluate each chunk, thus I can only use a proportion bar chart to show the content of each chunk across the article.

For affine and Bing lexicon, we can see that most of the time the story has negative sentiment, which is reasonable, since this is a detective fiction where most words are classified as negative. We can also see that there are two negative peaks at around 20, 50 chunks. They are consistent with the story plot where Holmes finds the murderer in the middle of the first part of the story, and the murderer reviews the criminal process in the middle of the second part. And the chunk around 38 has small scores values which is the conjuction of two parts of this story. NRC and Loughram are kind of hard to analyze, maybe we can define a way to evaluate them numerically later.









# Reference

1. Wikipedia: https://en.wikipedia.org/wiki/A_Study_in_Scarlet
2. Baker Street website: https://bakerstreet.fandom.com/wiki/A_Study_in_Scarlet
3. TextMining book: https://www.tidytextmining.com/sentiment.html