Sentence-level Sentiment Analysis of [A Study in Scarlet]

XihaoCao

2021/12/8

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 manually, and the data is a txt file containing all plain text. Then I read the data into a larger character vector using the readLine function, and each element is a single sentence. Then I upload the sentence-level data into the tnum server for future use.

Check server status

There are multiple numspaces in the server, I need to set the working space to test2, and check whether my data is well stored.

```
## [1] "The following are available spaces in the server"

## Available spaces: testspace, MEPED, alion-rf, shared-testspace, test2, alion, NCNM, ED-900-Workshop,

## Numberspace set to: testspace

## [1] "Successfully find xihao's branches in the MSSP tnum server"
```

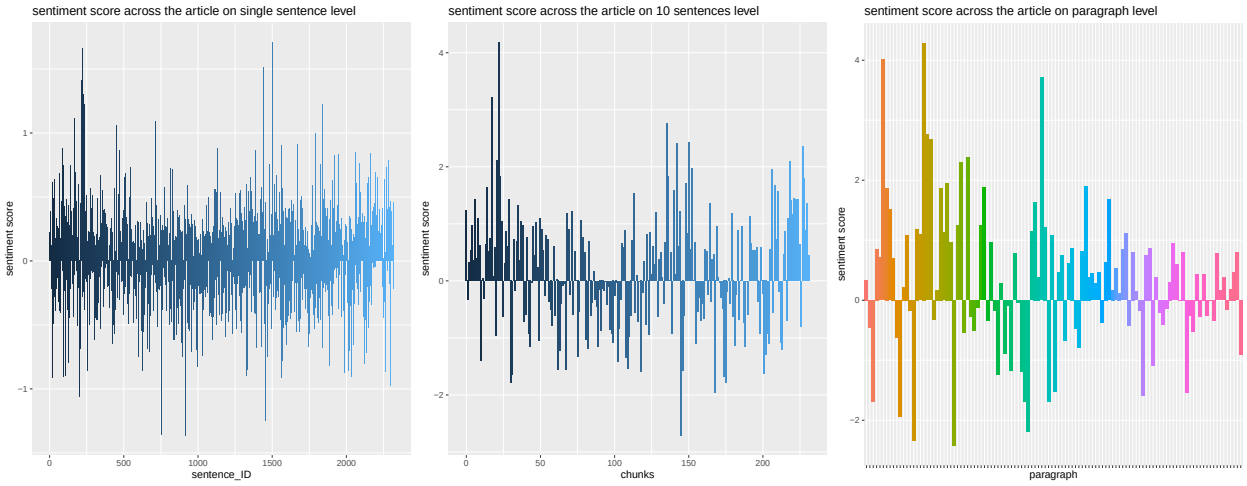
Sentiment score

After having my text stored in the server, I use the query function to get my sentence-level data. Then I use functions in the sentimentr package to calculate the sentence-level sentiment scores while making the following plots.

In the first graph, I plot the sentiment score for each sentence in the article, the x axis is the ordinal sentence ID which represents the order each sentence showing up in the story. However, since there are over 2000 sentences, the graph looks a little bit wild and we cannot find a overall pattern

In the second plot, I put 10 consecutive sentences into a chunk and plot the sum of their sentiment scores. The x axis is the chunk Id, where larger ID means the showing up later in the story. As we can see, there are less bars here and we can distinguish an overall pattern relatively easier.

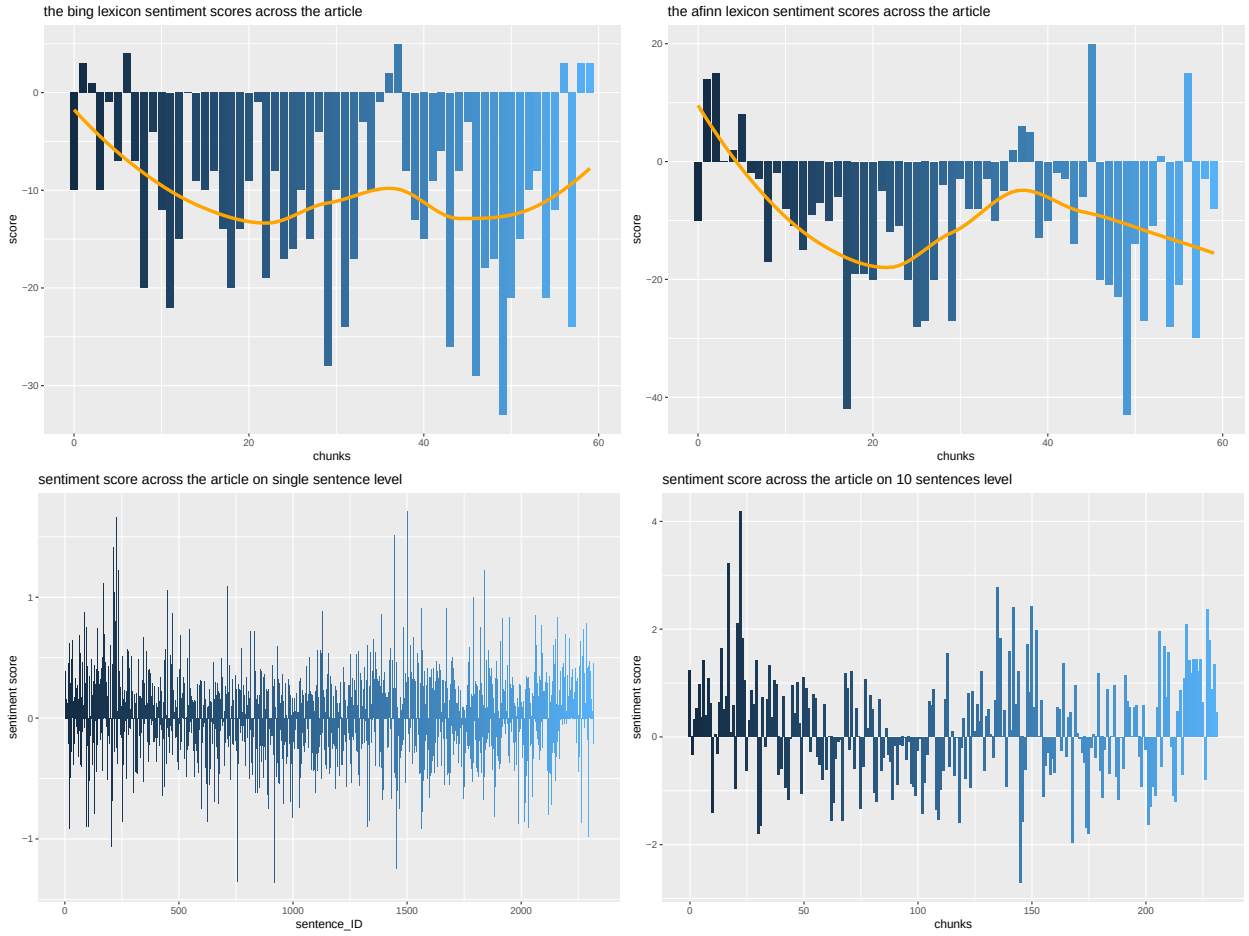
In the third graph, rather than group by consecutive 10 sentences, I group them by the paragraph they live in and make the plot. And we can see that there is not much difference between the 10-sentences graph and this one, they share the same overall pattern.



Compare the token-level and sentence-level sentiment scores across the story

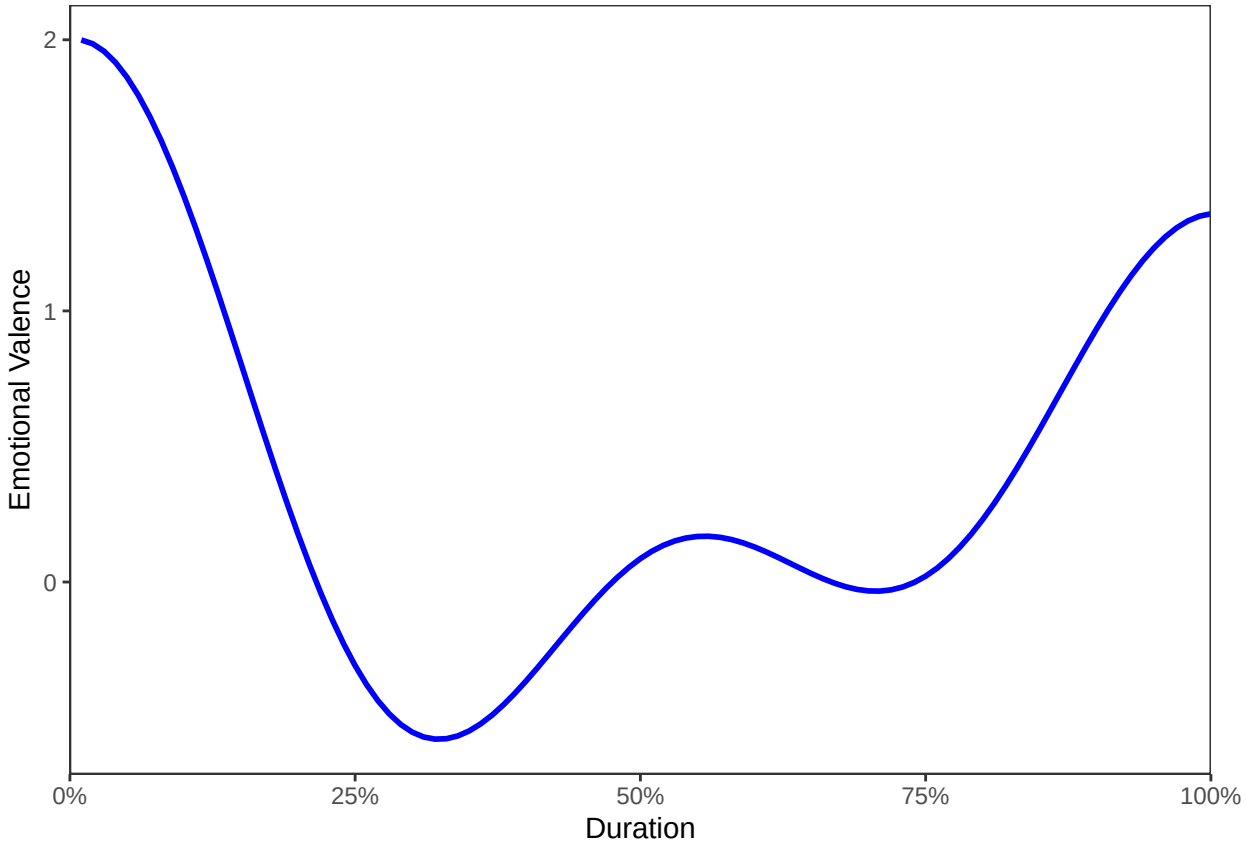
In the token-level report, I use four different lexicon engines to illustrate the sentiment scores/proportion across the story on the token-level. Now I have the sentence-level sentiment score across the story, I want

to make a comparison between the sentence-level and token-level. I will only use the bing and afinn lexicon engines since they also use numerical value to quantify the sentiment. One thing needs notice is that in the token-level graphs, each chunk contains 80 consecutive words.



It seems that the sentence-level and token-level graphs do not share the same shape. I guess the reason is that the sentimentr package has its own formula to calculate the sentiment score of a sentence rather than simply sum the sentiment scores of all words in that sentence.

Plot the Emotion Valence versus Duration



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
4. Sentimentr package manual: https://learn.bu.edu/bbcswebdav/pid-9886265-dt-content-rid-63157069_1/xid-63157069_1