STATISTICAL ANALYSIS OF SHERLOCK HOLMES SHORT STORIES

A project report submitted for the partial fulfillment of the Master of Science in Statistics, Visva-Bharati, Santiniketan.

Submitted by
Name of the Student: Krishnakanta Maity
Exam Roll No: MSC-STAT-SEM-IV-05
Department of Statistics,
Visva-Bharati, Santiniketan
2020

DECLARATION

I, Mr Krishnakanta Maity, hereby under take that the dissertation entitled ***"Analysis of Sherlok Holmes Short Stories"*** is the result of the Post Graduate project work carried out by me, at the Department of Statistics, Visva-Bharati, Santiniketan. I have collected the data from primary or/and secondary sources as per the requirement of my dissertation, which are appropriately referred in the report. All the computations involved in this dissertation are result of my own calculations on the data that I collected.

The formulae that are used in the dissertation are acknowledged providing appropriate reference of the source from which they are obtained. No part of the dissertation is been submitted to any other institution for the purpose of any degree/diploma.

(Krishnakanta Maity)

June 28, 2020

ACKNOWLEDGEMENT

It is great pleasure for me to undertake this project. I am grateful to my project guide Prof. Tirthankar Ghosh.

This project would not have completed without his enormous help and worthy experience. Whenever I was in need, he was there behind me.

Although, this project report has been prepared with utmost care and deep routed interest. Even then I accept respondent and imperfection.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Sherlock Holmes is a fictional private detective created by British author Sir Arthur Conan Doyle (Doyle, 1993). Holmes is known for his proficiency with observation, deduction, forensic science and logical reasoning that borders on the fantastic, which he employs when investigating cases for a wide variety of clients.

1.1 Background

This project was motivated by my desire to investigate the sentiment analysis (Hawkins & Niblock, 2011) field of machine learning since it allows to approach natural language processing (Camilleri, 2019) which is very popular now. Following my previous experience where it was about classifying tweets according to their positive and negative sentiments, I applied the same idea with Sherlock Holmes 56 short stories and try to build a model for classify the story into some collection type of story.

1.2 Literature

This project is divided into three part in which we will use R to perform a variety of analytic tasks on 56 short stories of Sherlock Holmes by the author Arthur Conan Doyle. In the first part we do text mining and exploratory analysis, in the second part we will do sentiment analysis and in the last part we will do predictive analytics using machine learning algorithm.

1.3 Objectives

Before dive into predictive analysis we are interested in how the sentiments are flows in the stories. What are the frequent words and what are the sentiment behind these words? What are the facts? Lastly, we are trying to build a model for classify the stories into some collection type of story.

Chapter 2

DATA SET

2.1 Data Source

All 56 stories are available in very popular repository Kaggle with .txt file extension (Data Link).

2.2 About Data

Full dataset has 67 .txt file, from these files we select 56 short stories which are written by author Arthur ConanDoyle. Each text file has full name of story, author name and a common license.

2.3 Pre-processing

It is a very important as we have only raw text files. We extract title of the story, abbreviation and main story part from each file. After that we collect some related information about stories from different websites such as publication year and month (Knight, 2003), types of story, rank of story (Stock, 1999). Using the very popular udpipe parts of speech model (Straka et al., 2016) for R, we count different parts of speech.

2.4 Making Final Dataset

We created two datasets, one for sentiment analysis purpose and another for predictive analysis. Initial variables are title, abbreviation, published year, collection type of story and rank. In dataset-I, we include variable main story, which contain raw text of each story. In dataset-II, we have counts of all the parts of speech tag e.g. how many adjective, adverb, verb etc are there in each story.

Chapter 3

METHODOLOGY

3.1 NLP and Related Task

Natural language processing (NLP) is one methodology used in mining text. It tries to decipher the ambiguities in written language by tokenization, clustering, extracting entity and word relationships and using algorithms to identify themes and quantify subjective information.

Lexical complexity can mean different things in different contexts. Some measures are following:

- **Word Frequency**: Number of words per story.
- **Word Length**: Average length of individual words in a text.
- **Lexical Diversity**: Number of unique words used in a text.
- **Lexical Density**: The number of unique words divided by the total number of words.

First of all, we create a corpus (collection of documents) of 56 stories and tokenize each story. Then we remove the punctuations and numbers form all the stories. Remove all the unnecessary English stop words (like an, the etc.) and undesirable words which are commonly occurs and unnecessary to our analysis (e.g. Holmes, like). Pictorially after tokenization, we pre-process the data as follows(See 3.1).

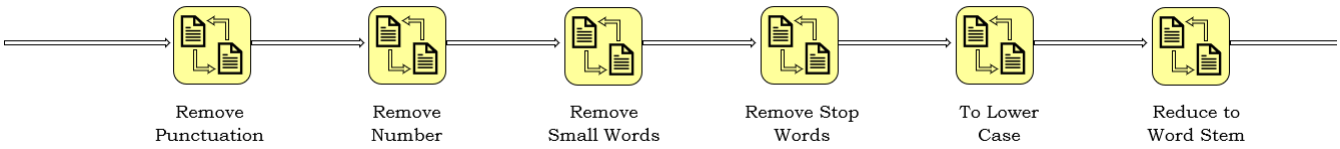


Figure 3.1: Roadmap of cleaning text

By looking at word-cloud we create a set of undesirable words. Stemming is very useful things for text analysis as this process stem all formats of a word into the main form word. Stemming refers to removing suffixes from words to get the common origin.

In story, individual word frequencies carry a great deal of importance, whether it be repetition or rarity. Both affect memorability of the entire story itself.

3.1.1 TF-IDF

TF stands for 'term frequency' and IDF is 'inverse document frequency', which attaches a lower weight for commonly used words and a higher weight for words that are not used much in a collection of text. When we combine TF and IDF, a term's importance is adjusted foe how rarely it is used. The assumption behind TF-IDF (Ramos et al., 2003) is that terms that appears in many documents. The formula can be summarized below:

- **Term Frequency**: Number of times a term occurs in a document. By term document matrix, we define term frequency. In term document matrix T, t_{ij} be word frequency of the jth term of ith document.
- **Document Frequency**: Number of documents that contain each word.
- **Inverse Document Frequency**: Reciprocal of document frequency.
- **TF-IDF**: TF*IDF.

The IDF of any term is therefore a higher number for words that occur in fewer of the documents in the collection.

3.1.2 Lexicon

In this project we use pre define lexicon provided by 'tidytext' package (Silge & Robinson, 2016). Tidytext package includes a dataset called 'sentiments' (Silge & Robinson, 2017), which provides several distinct lexicons. These lexicons are dictionaries of words with an assigned sentiment category or value. Three general purpose lexicons:

- **AFINN**: Afinn lexicon (Nielsen, 2011) assigns words with a score that runs between -5 to 5, with negative scores indicating negative sentiments and positives scores indicating positive sentiment. There are 1597 negative and 879 positive words stored.

Table 3.1: Words count of Afinn lexicon

Words in Lexicon	Positive	Negative
2467	879	1597

- **BING**: Bing lexicon (Ding et al., 2008) assigns words into positive and negative categories. Bing define 4782 negative and 2006 positive words.

Table 3.2: Words count of Bing lexicon

Words in Lexicon	Positive	Negative
6785	2006	4782

- **NRC**: NRC assigns (Mohammad & Turney, 2013) words into one or more of the following ten categories: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

Table 3.3: Words count of NRC lexicon

Words in Lexicon	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
8265	1247	839	1058	1476	689	1191	534	1231

Chapter 4

DATA ANALYSIS

4.1 Text Mining and Exploratory Analysis

Since one of our target questions is to look for high ranking story published across the time and dataset contains individual release years, so we create a bucket and group the years into decade. Similarly, we create a filter of top 10 high ranked story. Before getting into text mining, start with a basic view of what our data holds at the story collection.

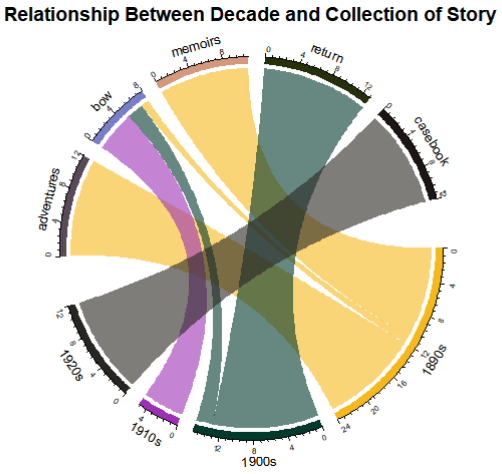


Figure 4.1: Chord plot by decade and collection

Fig: 4.1 and Fig: 4.2 clearly show author's most active decade was 1990s and only in that decade he writes three different types of story.

From the Fig: 4.3, high ranked stories are written in first two decade also these stories are published in start of decade.

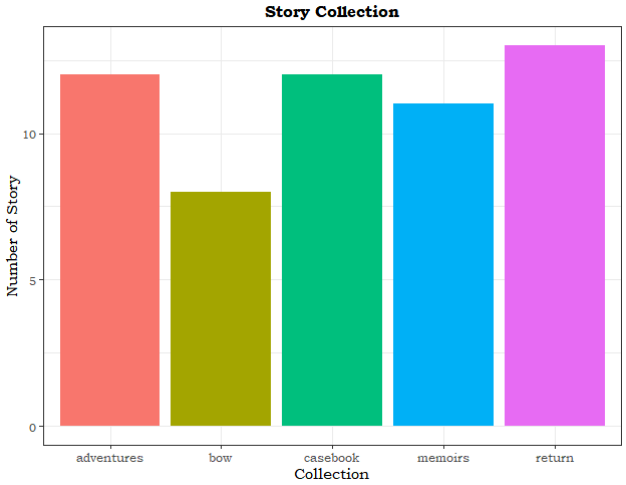


Figure 4.2: Barplot of number of story by collection type

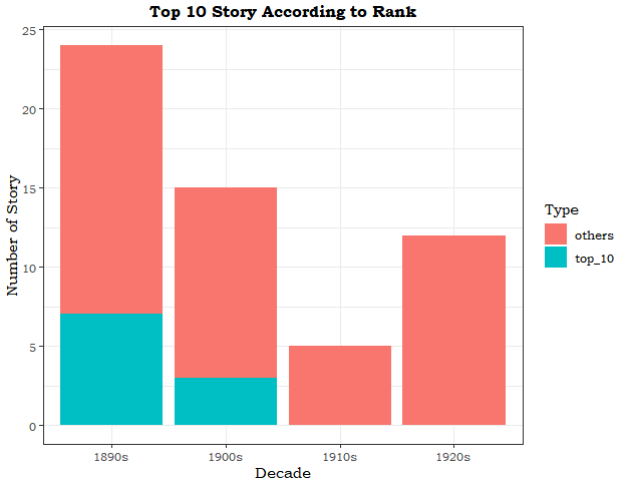


Figure 4.3: Top 10 ranked story by decade

Before evaluate the word frequency we set our undesirable set of words by looking at word cloud (See Fig: 4.4) and remove them e.g. holmes, said, one etc.

In order to do a simple evaluation of the most frequently used words in the full set of stories, we count the occurrence of each terms and plot them. As in detective story, time, house, door, night seems to be common word.

4.2 Sentiment Analysis

Classify words into positive and negative category by 'bing' lexicon, we compare two categories(See Fig: 4.6(b)). Adventure or detective stories have death,

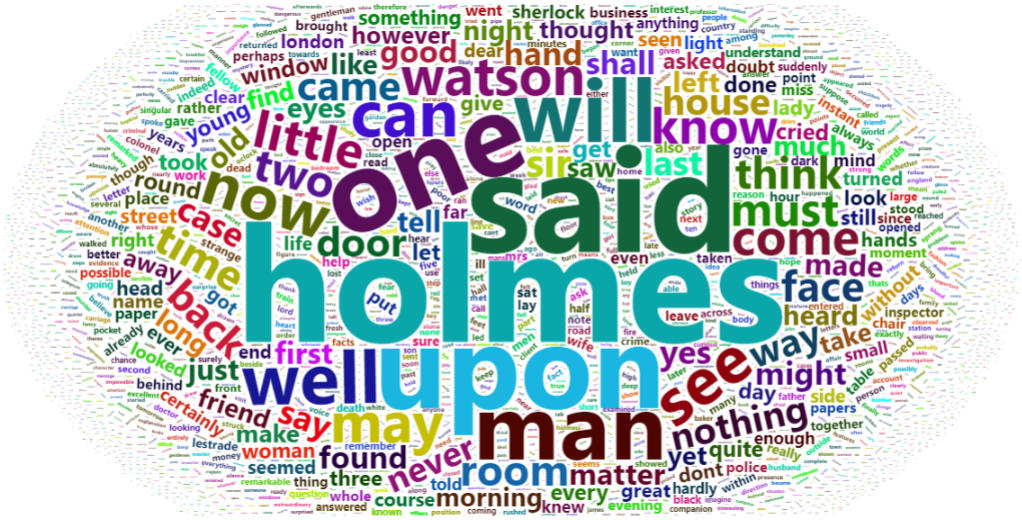


Figure 4.4: Word cloud before removing undesirable set of words

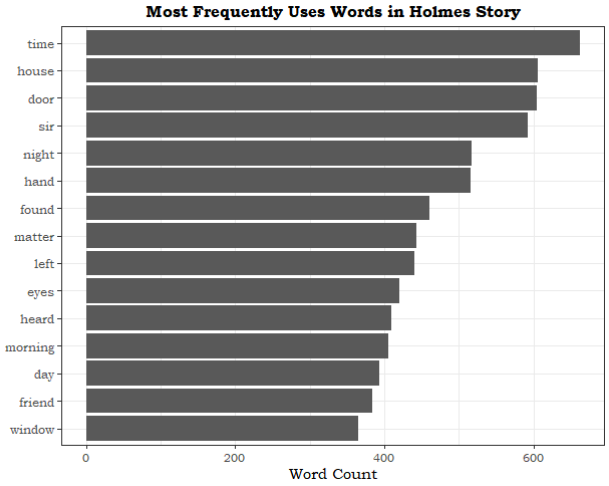
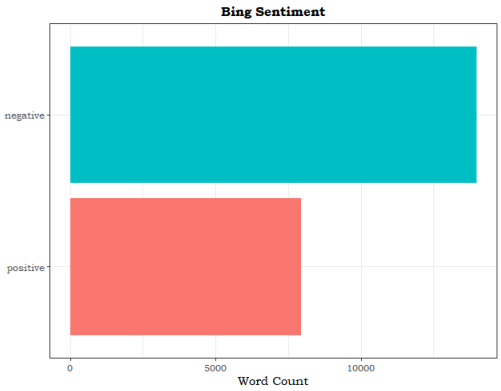


Figure 4.5: 15 most frequent words

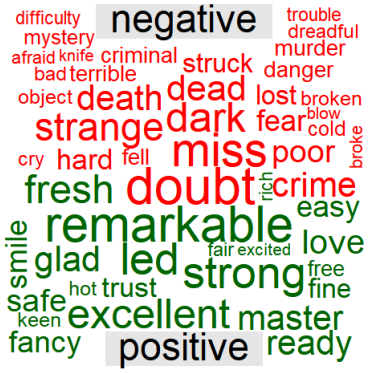
miss, doudt, crime negative sentiment terms and excellent, remarkable etc positive sentiment terms. Also stories have more negative terms than positive terms (See Fig: 4.6(a)).

By looking at NRC sentiment plot (See Fig: 4.7) we can see that excluding positive and negative words, words with sentiment trust is more than others type of sentiment. Trust is more suitable for any kind of detective stories.

In Fig: 4.8, each sentiment category has 10 most written words. Although there are some intersecting terms, but these categories are depending on whole sentence sentiment.



(a) Sentiment classification by Bing lexicon



(b) Word cloud of negative and positive words

Figure 4.6: Sentiment with Bing lexicon

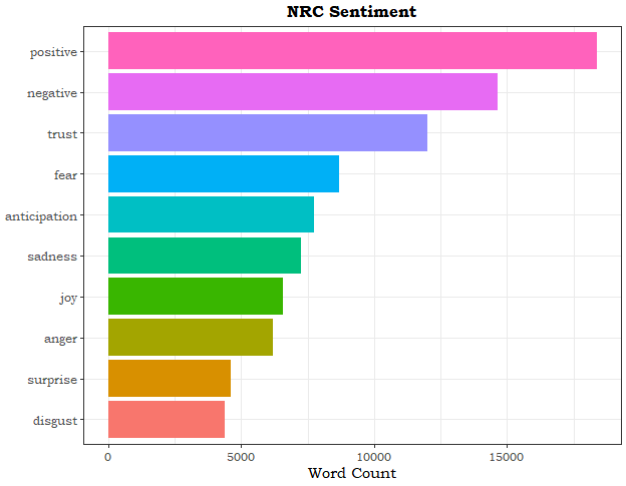


Figure 4.7: Words count by NRC lexicon

4.3 Predictive Analysis

Here we will do regression modelling. We need to decide which variables to include in the regression. We have 12 variables that capture the statistical properties of the text and we would like to use these variables to predict the quality of the story on a scale of 1 to 5. However, we only have 56 observations (i.e. stories), so forcing all 12 variables into the model would almost certainly result in unreliable parameter estimates for some variables. Here we use cross-validation technique to choose the optimal number of variables that avoid overfitting our data. We use R's stepwise variable selection to create models of increasing complexity, and then check the cross validated mean-squared error of each of these models.

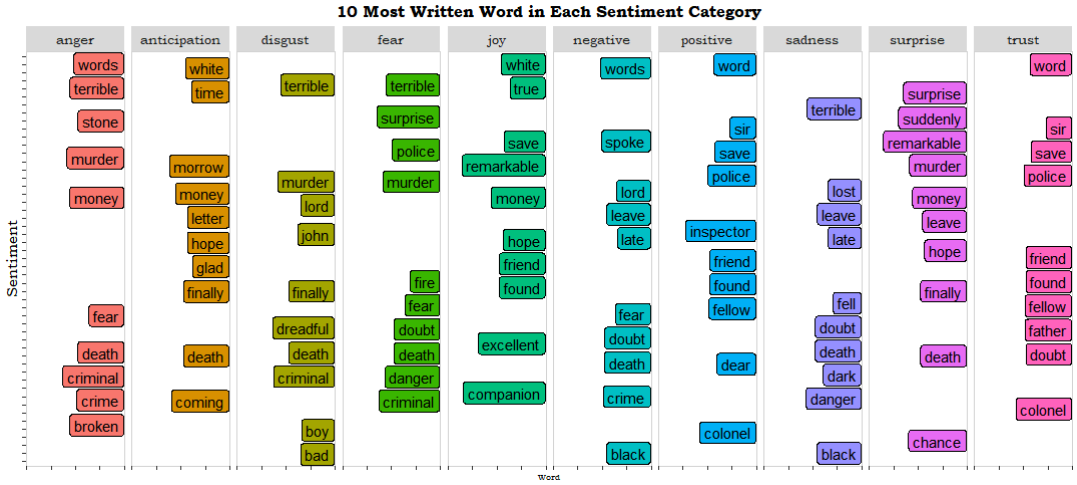


Figure 4.8: Top 10 words in each category by NRC lexicon

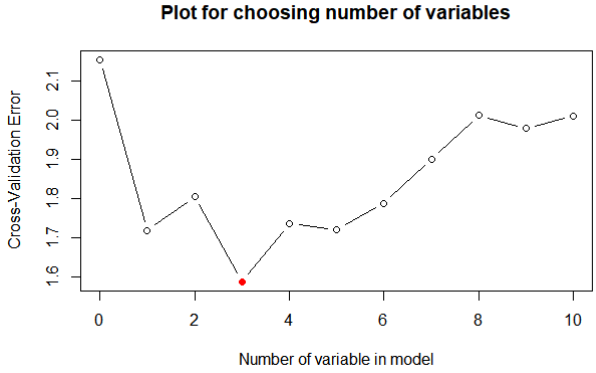


Figure 4.9: Plot for number of variable we sholud take into model with minimum MSE

From the Fig: 4.9, we pick up 3 variables as it gives us less MSE value. Three variables give us the best performance.

Now we have to choose those three variables which contributed more to our model.

Our model suggests, there are some strong relationship between the statistical properties of the text and the quality of Sherlock Holmes story. Three variables are number of words (now), number of pronoun (pron) and average word per sentence (awps).

From the model we conclude that rating of story is related to number of words and average number of words per sentence. Longer stories have better rating.

4.4 Conclusion

Readers are love to read the 1990s published story which are mostly casebook and adventures. Time, door, night, heard, friend etc frequent term indicate the stories are related to detective sense.

Our predictive model captures 33% variability, which is quite low but three variables gives us minimum MSE value. Model says that quality of story is highly related to longer stories i.e. number of words and average words per sentences. Also quality depends on number of pronoun i.e. how much conservation going on between two person.

References

Camilleri, L. (2019). *Natural language processing for sentiment analysis* (B.S. thesis). University of Malta.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 231–240).

Doyle, A. C. (1993). *The case-book of sherlock holmes*. Wordsworth Editions.

Hawkins, S., & Niblock, S. (2011). *Prince: The making of a pop music phenomenon*. Ashgate Publishing, Ltd.

Knight, S. T. (2003). *Crime fiction 1800-2000: Detection, death, diversity*. Palgrave Macmillan.

Mohammad, S. M., & Turney, P. D. (2013). Nrc emotion lexicon. *National Research Council, Canada*, *2*.

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133–142).

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in r. *Journal of Open Source Software*, *1*(3), 37.

Silge, J., & Robinson, D. (2017). *Text mining with r: A tidy approach. "* O'Reilly Media, Inc.".

Stock, R. (1999). Rating the canon. *The Baker Street Journal*, *49*, 5–11.

Straka, M., Hajic, J., & Straková, J. (2016). Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the tenth international conference on language resources and evaluation (lrec'16)* (pp. 4290–4297).