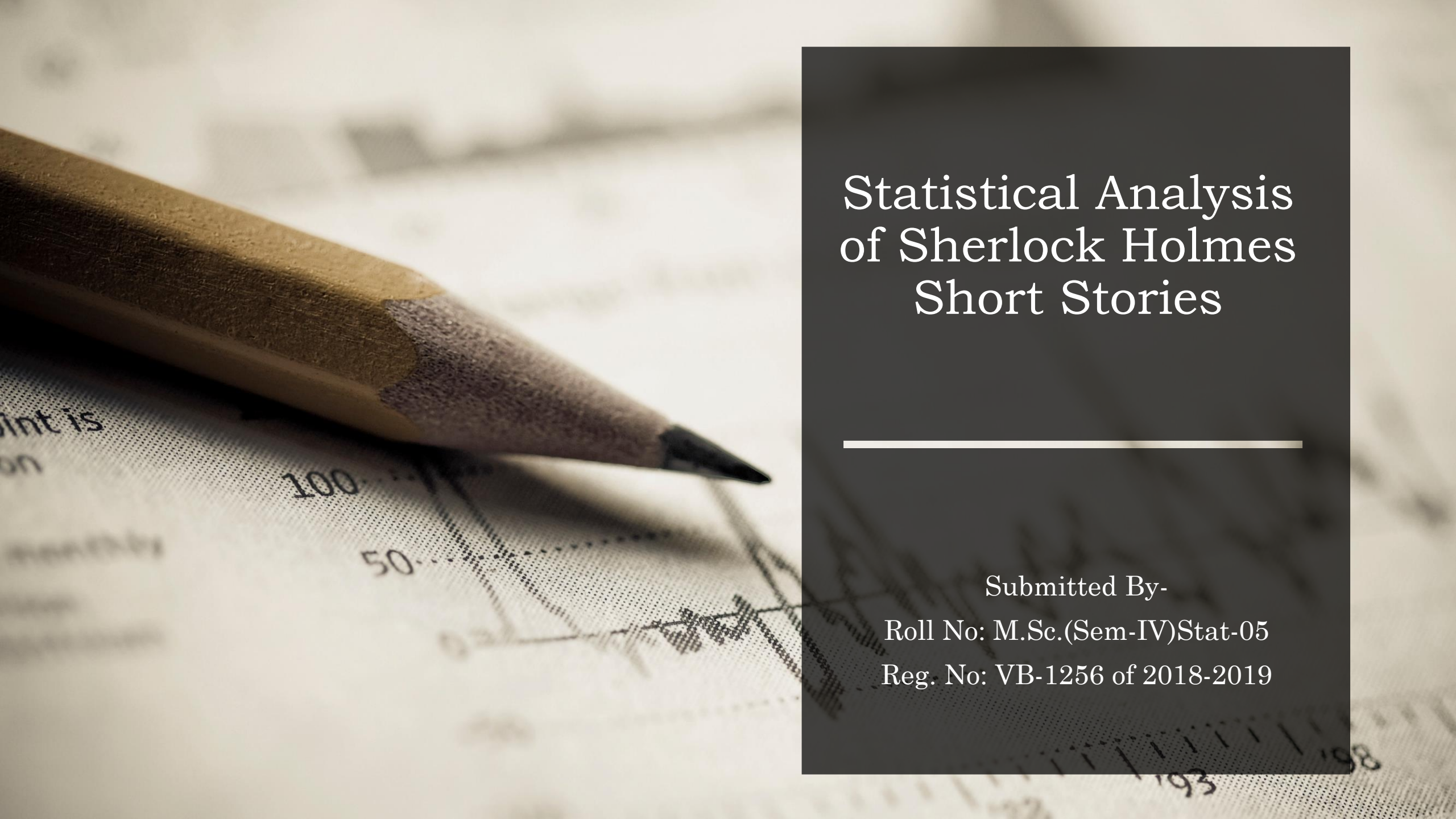


Statistical Analysis of Sherlock Holmes Short Stories

Submitted By-

Roll No: M.Sc.(Sem-IV)Stat-05

Reg. No: VB-1256 of 2018-2019

Aims and Objectives

- Classification of words in Sherlock Holmes short stories according to different sentiment class.
- Build a model for predict the quality of story.

Data Collection

Source: https://www.kaggle.com/idevji1/sherlock-holmes-stories

In the above Kaggle repository, 67 text file available. Out of these picked up only 56 short story of Sherlock Holmes written by Arthur Conan Doyle.

Two separate R-script collect the necessary data and store into two different csv file.

Description:

Dataset-I: Used for sentiment analysis	Dataset-II: Used for predictive analysis
• Dimension: 56 x 7 • Number of variables: 7 • NA values: 0 Variables: 1) Title 2) Abbreviation 3) Rank 4) Publication year 5) Collection type 6) Main story 7) Decade	• Dimension: 56 x 14 • Number of variables: 14 • NA values: 0 Variables: 1) Abbreviation 2) Collection type 3) Rank 4) Number of sentences 5) Number of words 6) Average words per sentence 7) Average character per word 8) Number of exclamation mark 9) Number of question mark 10) Number of adjective 11) Number of adverb 12) Number of pronoun 13) Number of proper noun 14) Number of verb

Sample Data

Dataset-I

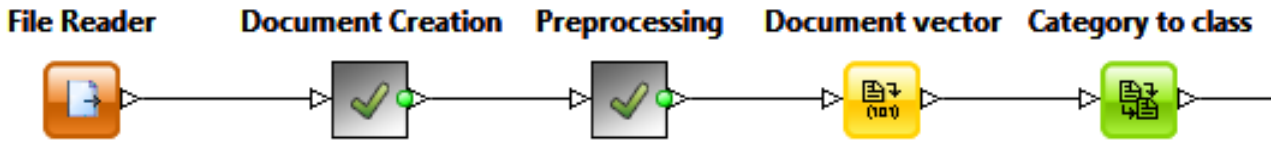
title	abbvr	rank	pub_date	collection	main_story	decade
THE ADVENTURE OF THE THREE GABLES	3gab	56	1926	casebook	I don't think that any of my adventures with Mr. Sherlock H...	1920s
THE ADVENTURE OF THE THREE GARRIDEBS	3gar	36	1924	casebook	It may have been a comedy, or it may have been a tragedy....	1920s
THE ADVENTURE OF THE THREE STUDENTS	3stu	49	1904	return	It was in the year '95 that a combination of events, into whi...	1900s
THE ADVENTURE OF THE ABBEY GRANGE	abbe	21	1904	return	It was on a bitterly cold and frosty morning during the wint...	1900s
THE ADVENTURE OF THE BERYL CORONET	bery	48	1892	adventures	'Holmes,' said I as I stood one morning in our bow-window ...	1890s
THE ADVENTURE OF BLACK PETER	blac	28	1904	return	I have never known my friend to be in better form, both me...	1900s
THE BLANCHED SOLDIER	blan	51	1926	casebook	The ideas of my friend Watson, though limited, are exceedi...	1920s
THE ADVENTURE OF THE BLUE CARBUNCLE	blue	5	1892	adventures	I had called upon my friend Sherlock Holmes upon the sec...	1890s
THE BOSCOMBE VALLEY MYSTERY	bosc	29	1891	adventures	We were seated at breakfast one morning, my wife and I, w...	1890s
THE ADVENTURE OF THE	bruc	11	1908	bow	In the third week of November, in the year 1895, a dense y...	1900s

Dataset-II

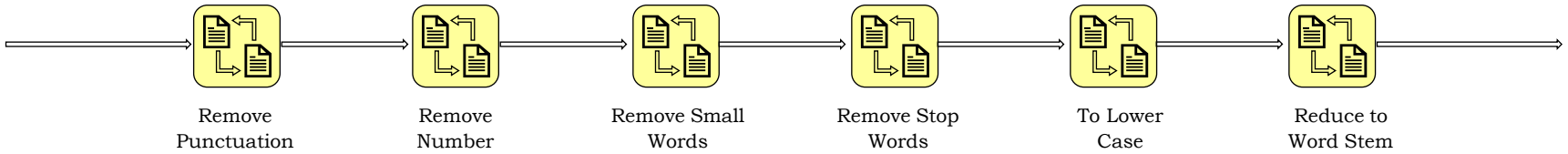
abbrv	collection	rank	nos	now	awps	acpw	em	qm	adj	adv	pron	propn	verb
3gab	casebook	56	593	6094	10.277	5.2903	36	84	368	414	1136	256	859
3gar	casebook	36	544	6242	11.474	5.3023	20	55	425	394	1040	285	804
3stu	return	49	582	6509	11.184	5.3455	23	78	459	443	1029	194	854
abbe	return	21	638	9248	14.495	5.2421	21	77	593	598	1538	271	1196
bery	adventures	48	647	9748	15.066	5.2043	43	78	539	756	1793	164	1366
blac	return	28	583	8194	14.055	5.3099	10	64	474	436	1239	328	1055
blan	casebook	51	561	7760	13.832	5.2981	15	44	532	477	1338	229	1036
blue	adventures	5	579	7886	13.62	5.3057	40	71	457	515	1200	291	1034
bosc	adventures	29	663	9689	14.614	5.2754	27	72	588	643	1558	329	1276

Data Cleaning

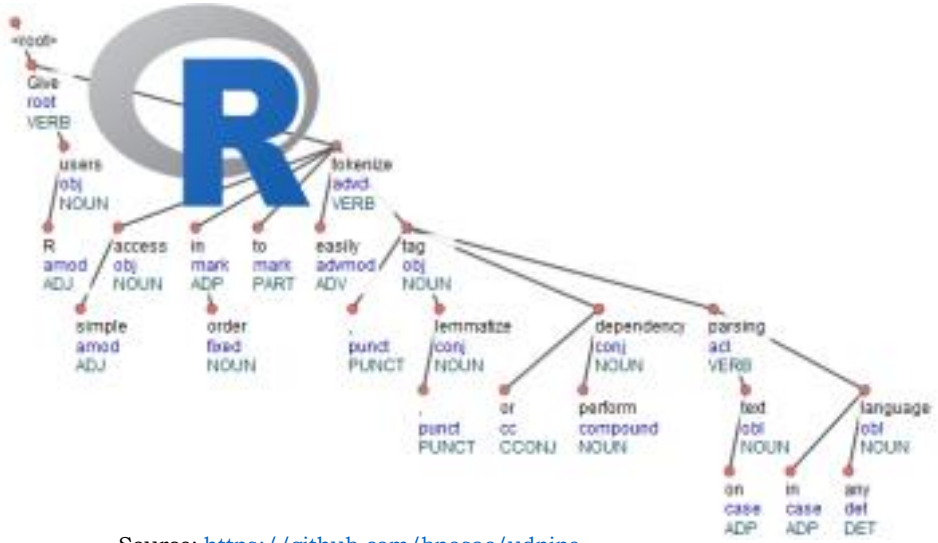
Roadmap



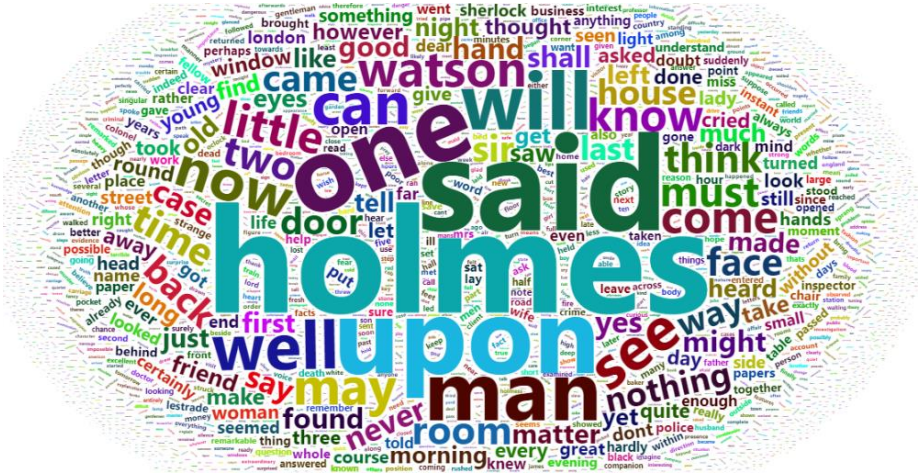
Pre-processing


Remove Punctuation
Remove Number
Remove Small Words
Remove Stop Words
To Lower Case
Reduce to Word Stem

For extract POS (Parts of Speech) tag



Source: https://github.com/bnosac/udpipe



TDM and Lexicon

TDM stands for term document matrix. Each cell of this matrix indicate the frequency of individual term occurs in corresponding document.

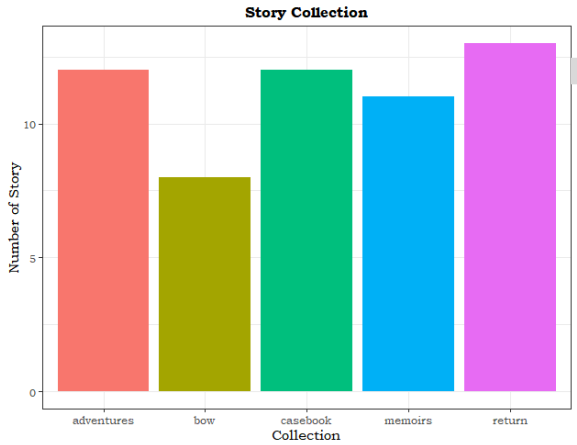
sentiment_analysis_SH_stories.R | trial.R | dtm_matrix | sholmes

Filter | Cols: 1 - 50

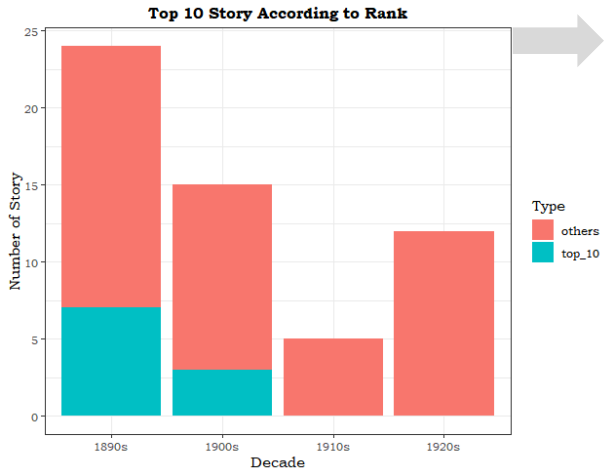
	3gab	3gar	3stu	abbe	bery	blac	blan	blue	bosc	bruc	card	chas	copp	cree	croo	danc	devi	dyin	empt	engr	fina
abandon	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0
abil	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
abl	0	2	1	0	3	3	5	1	3	2	0	0	2	6	2	7	0	1	1	1	1
abnorm	0	0	0	0	0	0	1	0	0	0	0	0	2	1	0	0	0	0	0	0	0
abroad	0	0	0	0	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0
abrupt	1	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1
absenc	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	2	0	1
absent	0	0	2	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
absolut	0	0	2	2	5	3	2	2	7	3	3	2	0	0	2	1	5	1	3	10	7

- **AFINN**: Afinn assigns words with a score that runs between -5 to 5, with negative scores indicating negative sentiments and positives scores indicating positive sentiment.
- **BING**: Bing lexicon assigns words into positive and negative categories
- **NRC**: NRC assigns words into one or more of the following ten categories: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

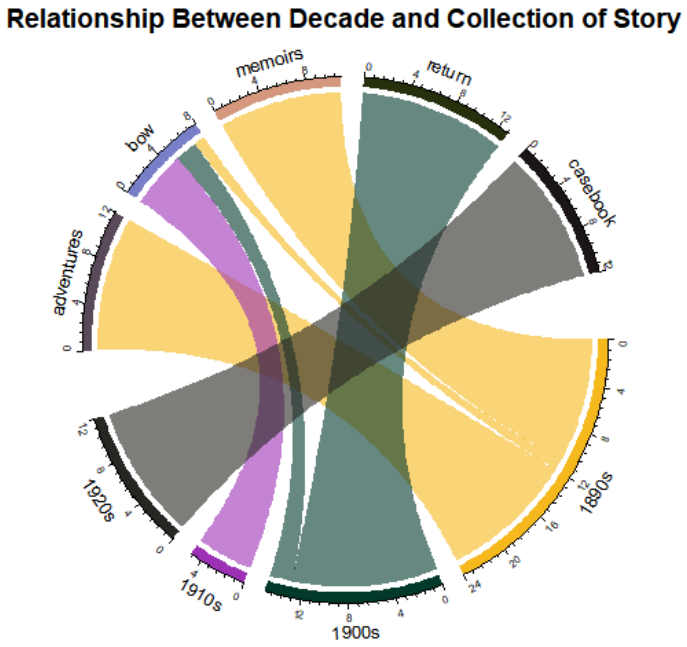
Exploratory Data Analysis



Maximum collections are return, adventure and casebook.

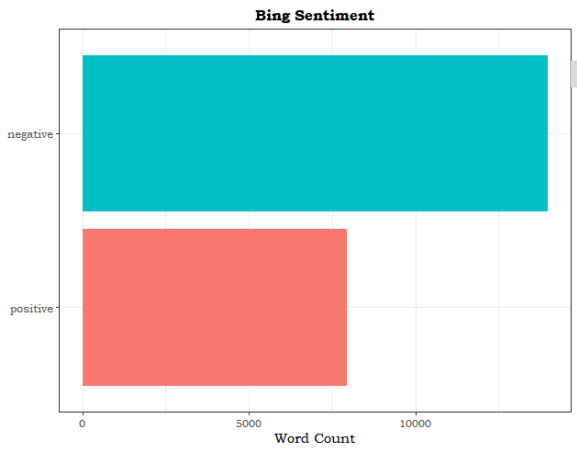


High rating stories are published in early decade.

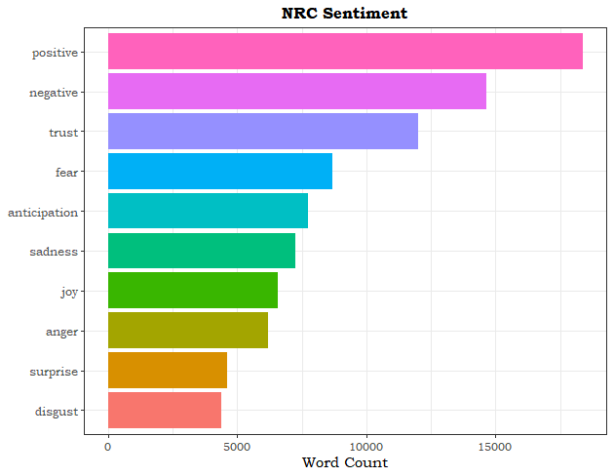


It shows that author's most active decade was 1990s and only in that decade he writes three different types of story

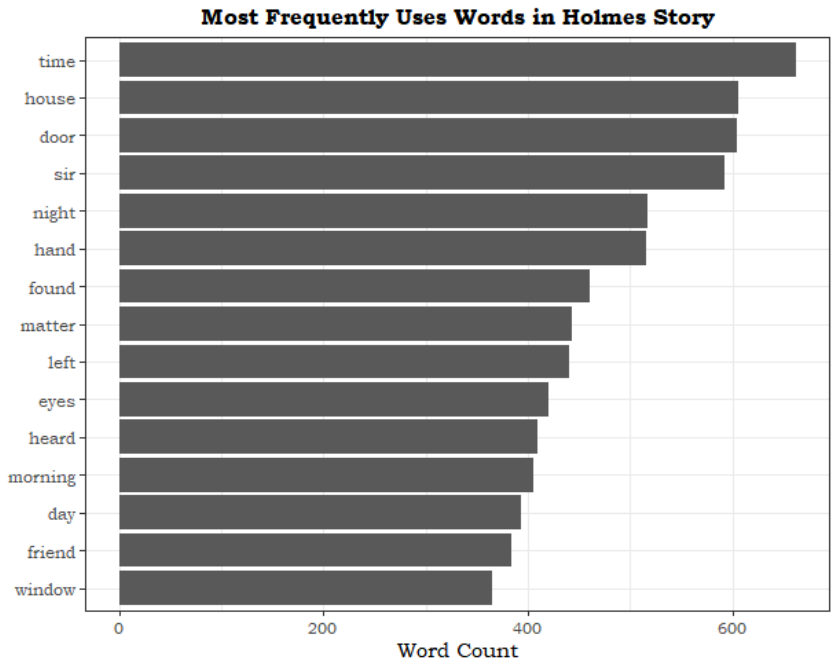
Sentiment Analysis



According to Bing lexicon, set of distinct terms have almost double negative words of positive words.

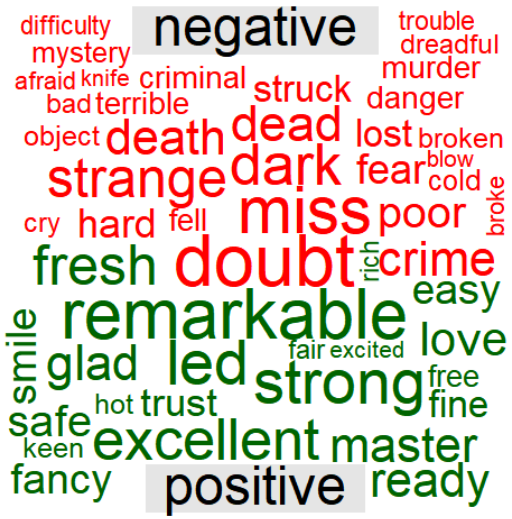


- Frequency wise order of sentiment is reliable.
- Excluding positive and negative sentiment, words with sentiment trust is more occurs than others type of sentiment.

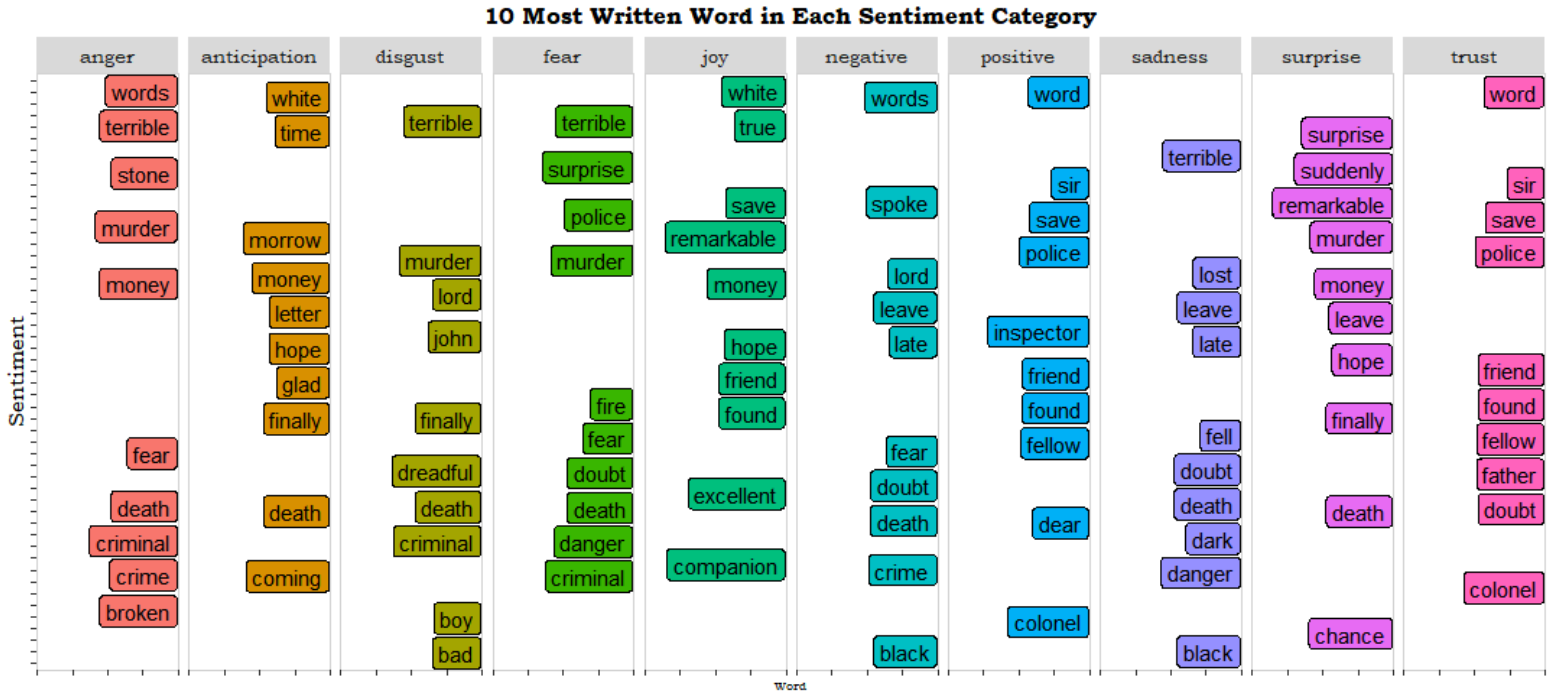


As in detective story, time, house, door, night seems to be common word.

Sentiment Analysis



Word cloud of words with frequency greater than 300. Size of words indicate their corresponding frequency.



It shows each sentiment category has 10 most written words. Although there are some intersecting terms, but these categories are depending on whole sentence sentiment.

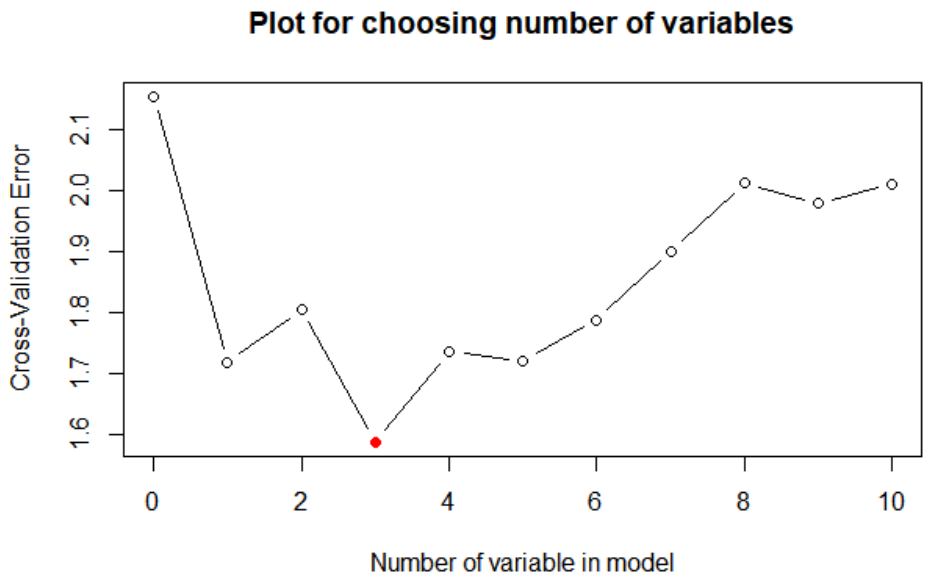
Predictive Analysis

There are 12 numerical variables that capture the statistical properties of the text and based on those variables model predict the quality of the story on a scale of 1 to 5.

Quality or ratings in scale 1 to 5 is the dependent variable and independent variables are count of words, sentences and parts of speech tag like adjectives, adverbs, verbs etc.

It use Cross-validation technique to choose the optimal number of variables that avoid overfitting our data.

Figure shows that model gives less MSE value when it takes three independent variables.



Model and Result

sentiment_analysis_SH_stories.R | model_building.R | trial.R | dtm_matrix | sholmes

Source on Save | Run | Source

```r
# Convert rank(1 to 56) into quality (1 to 5) star
data$quality <- as.numeric(cut(-data$rank, breaks = 5, labels = FALSE))

# Model with all variable
str(data)
data1 <- data[,-c(1,2,3)]
fml_full <- formula(quality~nos+now+awps+acpw+em+qm+adj+adv+pron+propn+verb)

# Cross validation folds
cv <- sample(rep(1:10, length_out = nrow(data1)))

# create model
max_vars <- 0:10
mse <- rep(0, length(max_vars))
for(i in 1:length(max_vars)){
  pred <- rep(0, nrow(data1))
  for(j in 1:10){
    train <- cv != j
    test <- !train
    m_null <- lm(quality~1, data1, subset = train)
    m_sel <- step(m_null, fml_full, steps = max_vars[i], k = 0)
    pred[test] <- predict(m_sel, data1)[test]
  }
  mse[i] <- mean((pred-data1$quality)^2)
}

# Eaxamine error vs number of variables in model
plot(max_vars, mse, type = 'b', xlab = 'Number of variable in model',
     ylab = 'Cross-Validation Error', main = 'Plot for choosing number of variables')
points(max_vars[which.min(mse)], min(mse), pch = 16, col = 'red')

m_null <- lm(quality~1, data1)
m_sel <- step(m_null, fml_full, steps = 3, k = 0)
summary(m_sel)
```

Console | Terminal | Jobs

K:/Data Science/Sherlock Holmes/

```
> summary(m_sel)

Call:
lm(formula = quality ~ now + pron + awps, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.30984 -0.68369 -0.05002  0.88887  2.26002 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.8617426  1.2782763  -1.456  0.15128    
now          0.0009946  0.0003142   3.165  0.00259 ** 
pron        -0.0043646  0.0020225  -2.158  0.03557 *  
awps         0.1746839  0.0835617   2.090  0.04148 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.178 on 52 degrees of freedom
Multiple R-squared:  0.3664,	Adjusted R-squared:  0.3299 
F-statistic: 10.02 on 3 and 52 DF,  p-value: 2.56e-05

>
```

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

K: > Data Science > Sherlock Holmes

Name	Size	Modified
sherlock_holmes_report.pdf	52.1 KB	Jun 17, 2020, 7:30 PM
sentiment_analysis_SH_stories_tidy_way.R	103 B	Jun 25, 2020, 5:40 PM
sentiment_analysis_SH_stories.R	7.3 KB	Jun 27, 2020, 12:20 PM
report		
project report.docx	763.3 KB	Jun 28, 2020, 12:39 AM

Three significant variables are number of words (now), number of pronoun (pron) and average word per sentences (awps).

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. Longer stories have better rating.

Thank You