Financial Statements Classification

Object

- Find out the profit and loss statement (PL) and balance sheet (BS) from each annual report.

Background

- Buy financial data from data vendors for business, but the cost is increasing year by year.
- Finding free historical financial data including recent data is difficult.
- Formatted free financial data is desired in financial analysis for academic use and individual investors.
- We can get financial statements for free, but the file names and formats are different in companies.
- In the future, extract each item of account from the profit and loss and balance sheet automatically.

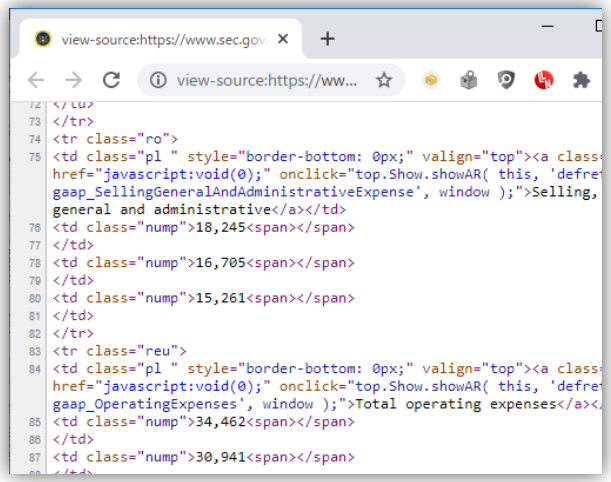
Description of Data - 1

- Financial statements from the annual reports (Form 10K) submitted to the U.S. Securities and Exchange Commission (SEC).
- Downloaded each financial statement's HTML file from EDGAR (website of SEC)
- https://www.sec.gov/Archives/edgar/data/{CIK}/{accession_number}/{file_name}
  - CIK (Central Index Key)
    - The ID for corporations and individual people.
  - accession_number
    - A unique number for each electronic submission of report.
  - file_name
    - A file name for a document. Each financial statement has a file, but the name is different in each report.
- The number of financial statements: 153,035
  - from 24,984 annual reports of 5,849 companies

Variables in Dataset - 1

- The frequency of words from each financial statement.
  - Financial statements(row): 153,035, Words(column): 13,039

HTMLs



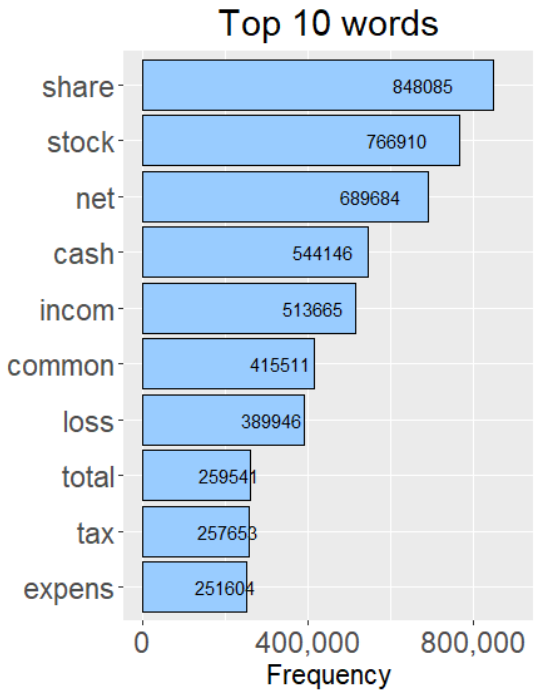
- Remove html tags
- Split phrases into words
- Remove punctuations
- Remove stop words
- Remove numbers
- Remove white space
- Stemming
- Count word frequency by financial statement

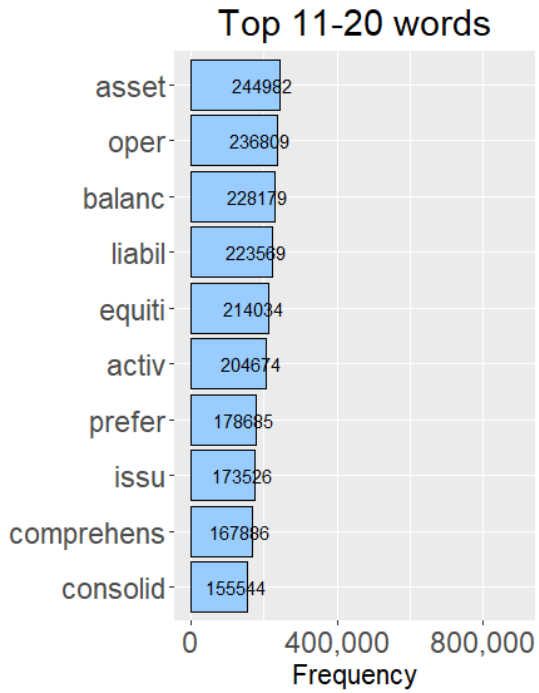
Word Frequency Table

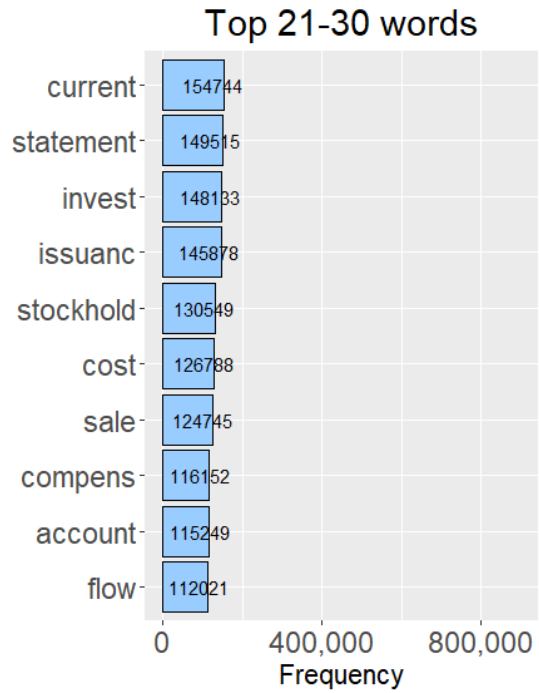
	doc_id	consolid	statement	share	net	comm
1	1	1	1	4	1	
2	2	2	2	0	6	
3	3	2	2	0	0	
4	4	1	0	4	2	
5	5	2	0	8	0	
6	6	1	1	3	6	
7	7	1	1	0	7	
8	8	1	1	4	1	
9	9	2	2	0	6	
10	10	2	2	0	0	
11	11	1	0	4	2	
12	12	2	0	8	0	

Variables in Dataset - 2

- Top 30 words by frequency in all financial statements.







Summary of Dataset

The word frequency in each financial statement.
(1st, 2nd, 3rd, 10th, 30th, 100th words by total frequency in all financial statements)

Rank	1	2	3	10	30	100
Word	share	stock	net	expens	flow	include
Min.	0.000	0.000	0.000	0.000	0.000	0.000
1st Qu.	0.000	0.000	0.000	0.000	0.000	0.000
Median	3.000	2.000	3.000	1.000	0.000	0.000
Mean	5.542	5.011	4.507	1.644	0.732	0.219
3rd Qu.	7.000	7.000	7.000	3.000	0.000	0.000
Max.	139.000	103.000	77.000	46.000	26.000	36.000

Descriptive Statistics for Each Variable

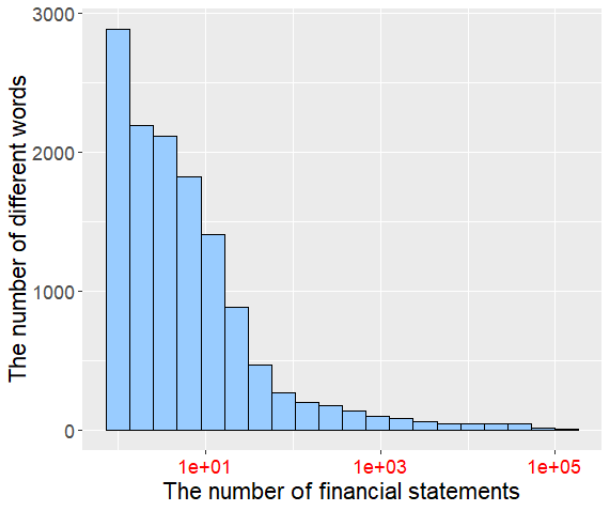
The word frequency in each financial statement.

(1st, 2nd, 3rd, 10th, 30th, 100th words by total frequency in all financial statements)

Rank	1	2	3	10	30	100
Word	share	stock	net	expens	flow	include
nbr.val	153,035	153,035	153,035	153,035	153,035	153,035
nbr.null	38,938	49,866	39,642	75,190	121,775	135,676
nbr.na	0	0	0	0	0	0
min	0	0	0	0	0	0
max	139	103	77	46	26	36
range	139	103	77	46	26	36
sum	848,085	766,910	689,684	251,604	112,021	33,566
median	3	2	3	1	0	0
mean	5.5417710	5.0113370	4.5067080	1.6440940	0.7319959	0.2193354
SE.mean	0.018330620	0.020127720	0.012331640	0.006216869	0.004440530	0.002211741
CI.mean.0.95	0.035927640	0.039449910	0.024169760	0.012184940	0.008703348	0.004334968
var	51.4215400	61.9983100	23.2719300	5.9147200	3.0175910	0.7486166
std.dev	7.1708810	7.8739010	4.8240980	2.4320200	1.7371220	0.8652263
coef.var	1.2939690	1.5712170	1.0704260	1.4792460	2.3731300	3.9447630

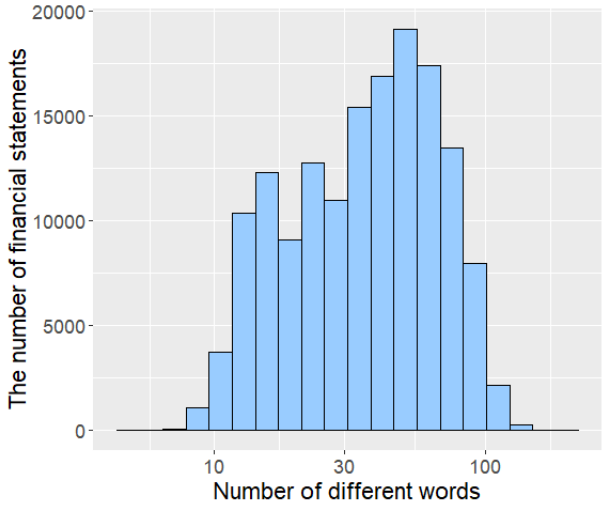
Descriptive Statistics - Histogram

The number of financial statements in which the word occurs.
(If a word occurs multiple times in one financial statement, it is counted as 1)



- A few words occur in so many financial statements, but most words occur in only a few financial statements.

The number of different words in each financial statement.



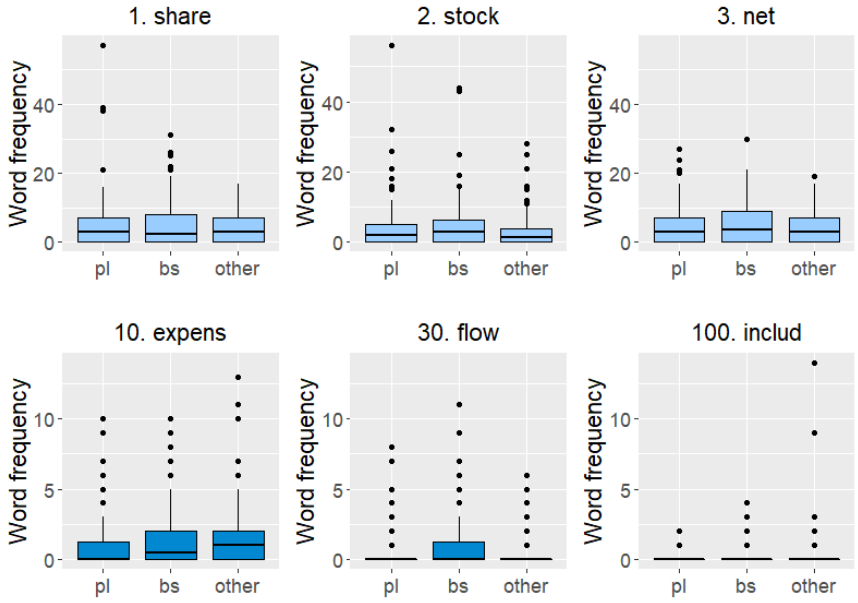
- The number of different words in one financial statement is at most 100.

Descriptive Statistics - Box Plot

The word frequency in PL, BS, and other financial statements.

(1st, 2nd, 3rd, 10th, 30th, 100th words by total frequency in all financial statements)

The distinction between PL, BS and other documents is not given, so I visually check each of the PL, BS, and other financial statements to pick 100 from each and graph them within the sample.

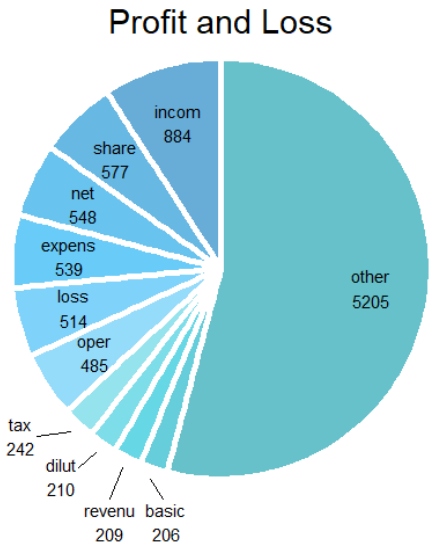


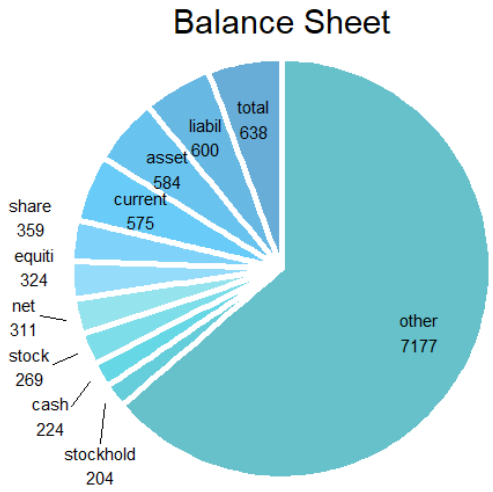
- Top 3 and 10th words
  - There are some small differences between financial statements.
- 30th and 100th words
  - There are less differences in between financial statements.
  - The frequency is so low.
- Top 10 to 30 word could be used for predictions, but over 30 might be less information.

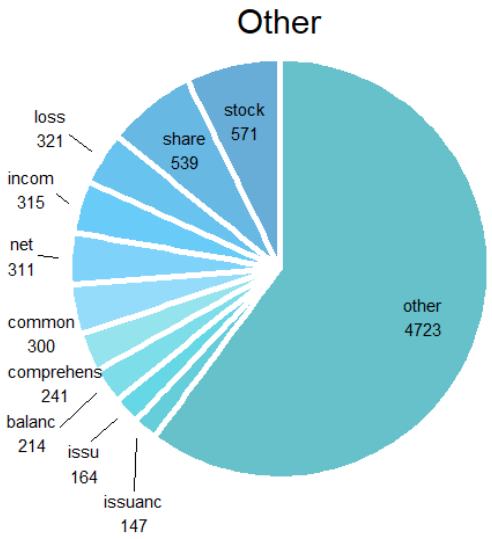
Descriptive Statistics - Pie Chart

The word frequency in sample financial statements.

The distinction between PL and BS and other documents is not given, so I visually check each of the PL, BS, and other financial statements to pick 100 from each and graph them within the sample.





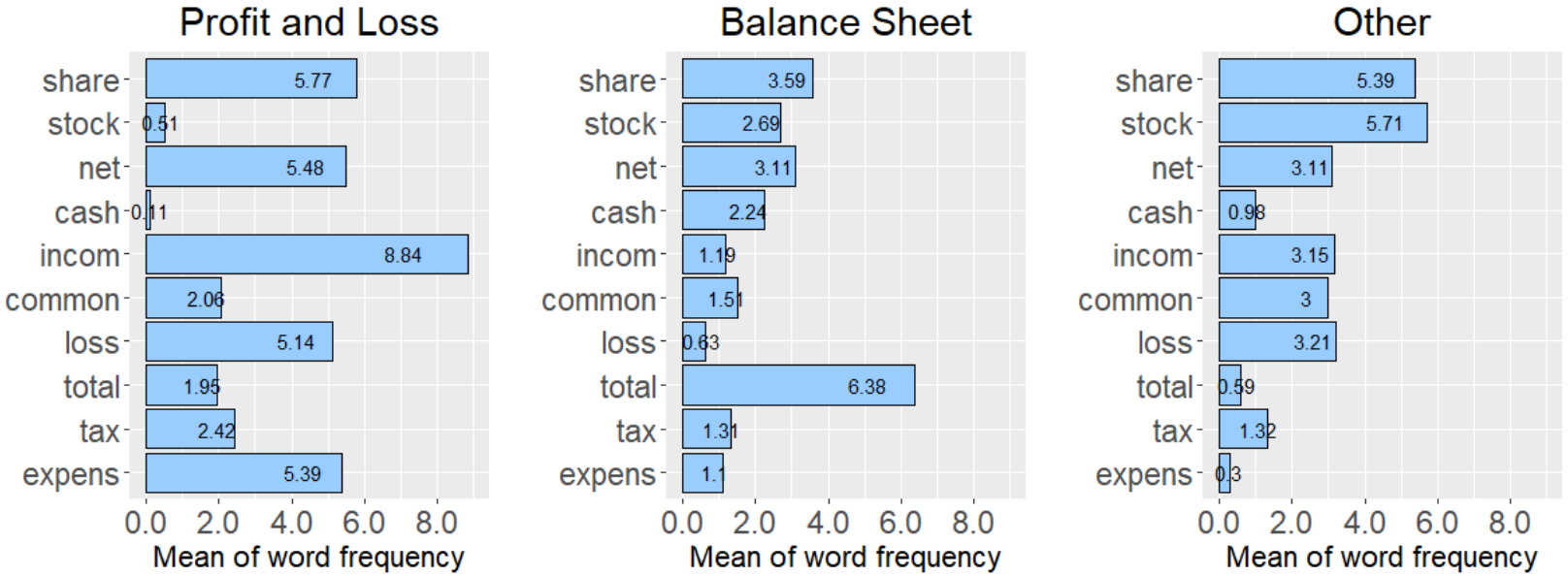


- The rank of top 10 words are different from each other.
- The top 10 words account for around 40% in each financial statement.
- The union of these top 10 words could be used for predictions.

Overall Mean for Word Frequency

Top 10 words by frequency in all financial statements.

The distinction between PL and BS and other documents is not given, so I visually check each of the PL, BS, and other financial statements to pick 100 from each and graph them within the sample.



- The means of the top 10 word frequencies are different each other.
- At least the top 10 words can be used for prediction.

Predictive Modeling - Naive Bayes

- Classify financial statements into three categories, profit and loss, balance sheet, and the other using Naive Bayes.
- Two types of tests and two models.
  - Test 1
    - Sample: 300 (PL:100, BS:100, Other:100)
    - Training: 80%
    - Test: 20%
    - Check the accuracy of the prediction.
  - Test 2
    - Population: 153,035
    - Training: 300 (PL:100, BS:100, Other:100)
    - Test: 152,735 (No distinction between PL and BS and other documents)
    - Check the number of financial statements classified as PLs and BSs in each annual report. The expected results are 1 for both PLs and BSs. (In most cases, it is assumed that the annual report has only one PL and BS.)
  - Apply two models for each test.
    - Model 1: Union of top 10 words in PLs, BSs and the others. (23 words)
    - Model 2: Top 23 words in all financial statements.

Comparing the 2 Models

Word	Model 1	Model 2	Rank in all financial statements	Rank in PLs	Rank in BSs	Rank in Others
share	✓	✓	1	2	5	2
stock	✓	✓	2		8	1
net	✓	✓	3	3	7	5
cash	✓	✓	4		9	
incom	✓	✓	5	1		4
common	✓	✓	6			6
loss	✓	✓	7	5		3
total	✓	✓	8		1	
tax	✓	✓	9	7		
expens	✓	✓	10	4		
asset	✓	✓	11		3	
oper	✓	✓	12	6		
balanc	✓	✓	13			8
liabil	✓	✓	14		2	

Word	Model 1	Model 2	Rank in all financial statements	Rank in PLs	Rank in BSs	Rank in Others
equiti	✓	✓	15		6	
activ	✓		16			
prefer	✓		17			
issu	✓	✓	18			9
comprehens	✓	✓	19			7
consolid	✓		20			
current	✓	✓	21		4	
statement	✓		22			
invest	✓		23			
basic		✓	81	10		
dilut		✓	253	8		
issuanc		✓	468			10
revenu		✓	798	9		
stockhold		✓	869		10	

Naive Bayes, Test 1

- Sample: 300 (PL:100, BS:100, Other:100)
- Training: 80%
- Test: 20%
- Check the accuracy of the prediction.
- Result
  - Model 1 (Union of top 10 words)
    - Correct : 18 + 22 + 19 = 59
    - Incorrect : 1
    - Accuracy: 59 / (59 + 1) = 98.3%

		Actual		
		PL	BS	Other
Predict	PL	18	0	1
	BS	0	22	0
	Other	0	0	19

    - An Other was miss-classified as PL.
  - Model 2 (Top 23 Words)
    - Correct : 17 + 23 + 10 = 50
    - Incorrect : 10
    - Accuracy: 50 / (50 + 10) = 83.3%

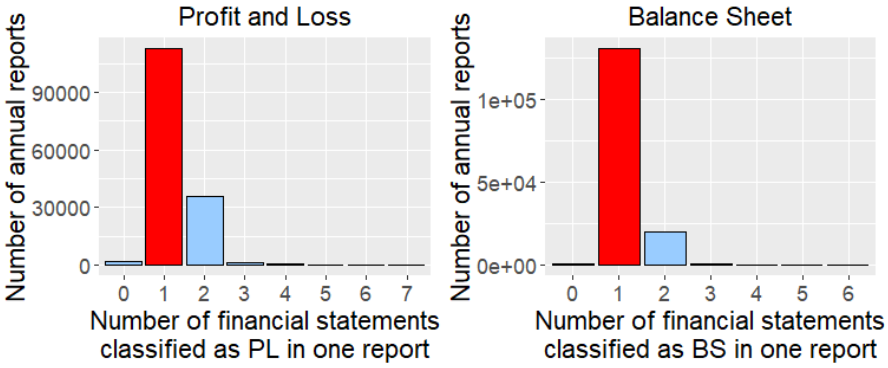
		Actual		
		PL	BS	Other
Predict	PL	17	0	10
	BS	0	23	0
	Other	0	0	10

    - 10 Others were miss-classified as PL.

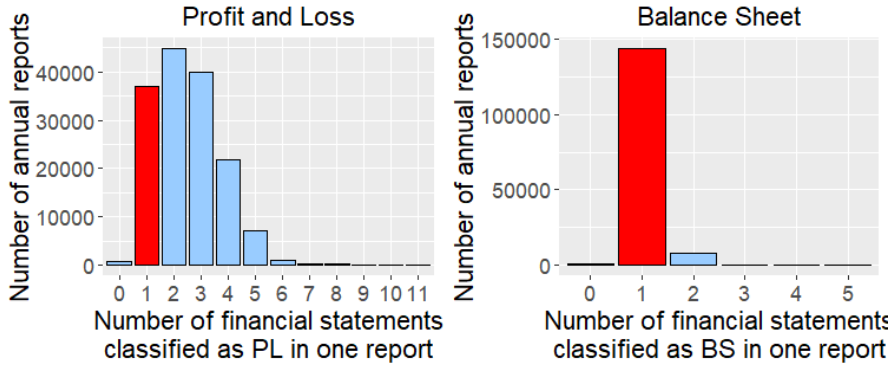
Naive Bayes, Test 2

- Population: 153,035
- Training: 300 (PL:100, BS:100, Other:100)
- Test: 152,735 (No distinction between PL and BS and other documents)
- Check the number of financial statements classified as PLs and BSs in each annual report. The expected results are 1 for both PLs and BSs. (In most cases, it is assumed that the annual report has only one PL and BS.)
- Result
  - Number of financial statements classified as PL and BS in one report.

Model 1 (Union of top 10 words)



Model 2 (Top 23 Words)



- The result of Model 1 is better than Model 2.

Libraries & Functions

Library	Function	Purpose
RMySQL	dbConnect, dbGetQuery	• Read texts of financial statements and some attributes from a DB
tidytext	unnest_tokens	• Split a phrase into token Remove punctuations
dplyr	anti_join	• Remove stop words from financial statements
	group_by, summarise, ungroup	• Count word frequencies by financial statements
SnowballC	wordStem	• Word stemming
base	grep, gsub	• Remove numbers from financial statements • Remove white spaces from words
	summary	• Summarize the dataset
	predict	• Prediction
pastecs	stat.desc	• Descriptive statistics
E1071	naiveBayes	• Creating classifier model for the Naïve Bayes algorithm
ggplot2	ggplot	• All plottings
	geom_histogram	• Histogram
	geom_boxplot	• Box plot
	geom_col, coord_polar	• Pie chart
	geom_bar	• Bar chart
ggrepel	geom_text_repel	• Text labels in pie chart

Conclusion

- The frequencies of words in PL, BS, and other financial statements are not the same. It can be said that each financial statement has its own characteristics.
  - Word information can be used for automatic classification.

- There are 13,039 kinds of words in all financial statements, but informative words for predictions are at most 30.

- The Model 1 (Union of top 10 words in PLs, BSs and the others) could classify financial statements in high accuracy than the Model 2 (Top 23 words in all financial statements)
  - The characteristical words from each financial statement help classification.
  - Adding more characteristical words may improve accuracy.

- The 300 training sets for 152,735 predictions were able to leave some results.
  - Increase sample data for prediction may improve accuracy.

- This time only word frequency was used for analyses, but using other information may improve accuracy.
  - Break Other into specific categories such as shareholders equity, cash flow, and so on.
  - Use word location information that considers the format of financial statements.
  - Add HTML headwords with weight.