FIN3210 Week 5 Assignment Report
Ma Kexuan 120090651

Abstract

This report prepared the word cloud, sentiment variable and fog index of a news article, as well as analyzing the Dataset of Tesla.

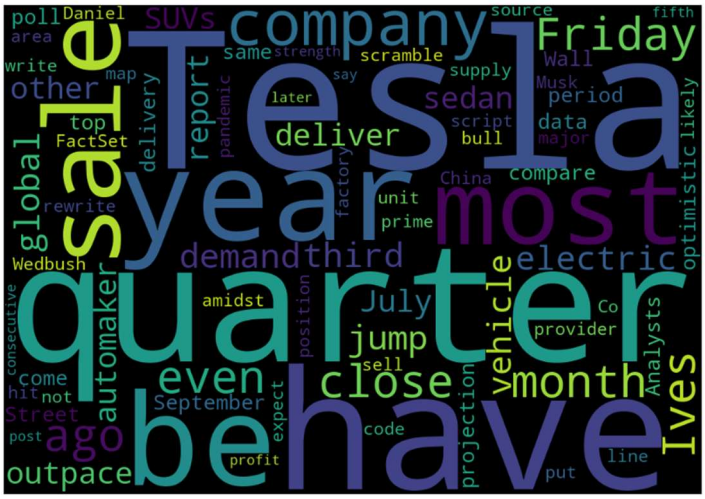
Data Preprocessing

The preprocessing procedures and some interpretations of the code are described in each code blocks in the appendix, please check.

Questions

Q1. Present word cloud

The word cloud is shown in the picture below. The news article was originally about the quarterly sales of Tesla. As we can directly see the picture, “Tesla, quarter, sale, company”, those words has the largest font size, which means that they are frequently used in the original article. Through this word cloud, we can see the essential information lucidly.



Q2. Calculate the news sentiment variable using Loughran and McDonald Sentiment Word Lists

By using the positive and negative words included in Loughran and McDonald Sentiment Word Lists, and since I have dealt with the original article into worked clean corpus, then I use the sentiment formula described in tutorial to calculate the overall sentiment of the whole passage. The result is 0.020204, which indicates that the sentiment of the article is quite neutral, and a little bit positive.

Q3. Calculate the Fog index

The Gunning Fog Index is used to measure the difficulty level of reading a passage, the more the Index value, means that the article is more difficult to read. Another interpretation is that the Index value is corresponded to the number of years that is needed to study in order to comprehend the whole article well. The Fog Index of the given article is 12.9282, which means that the article of Tesla is quite hard to comprehend, it approximately needs 13 years of study to better read the whole article.

Q4. Using the data set of Tesla, a) report the summary statistics of sentiment, novelty, and impact; b) present the correlation coefficient among sentiment, novelty, and impact; and c) show the frequency and fraction of top 10 news categories.

a) The summary statistics of sentiment, novelty and impact are provided below, 'Sentiment' had an average score of around 53.85 with a standard deviation of 14.82. 'Novelty' had an average of 28.88 with a broader spread (standard deviation: 38.54). 'Impact' had an average score of 45.28 with a standard deviation of 10.00.

	Sentiment	Novelty	Impact
count	1292.000000	1292.000000	1292.000000
mean	53.845975	28.877709	45.277864
std	14.824257	38.537946	9.996097
min	2.000000	0.000000	13.000000
25%	40.000000	0.000000	39.000000
50%	50.000000	3.000000	45.000000
75%	64.000000	56.000000	52.000000
max	100.000000	100.000000	77.000000

b) The correlation coefficient is provided below, the correlation between 'Sentiment' and 'Novelty' was 0.1656, suggesting a weak positive relationship. 'Sentiment' and 'Impact' had a correlation of -0.1310, indicating a weak negative association. The correlation between 'Novelty' and 'Impact' was -0.0648, suggesting a very weak negative relationship.

	Sentiment	Novelty	Impact
Sentiment	1.000000	0.165557	-0.130972
Novelty	0.165557	1.000000	-0.064844
Impact	-0.130972	-0.064844	1.000000

c) The frequency and fraction of top 10 news categories is provided below, we can see that terms relevant to stocks are frequently shown in the Tesla Dataset, since its stock price is quite volatile and it is genuinely a growing tech company, which makes the result more percipient.

	Category	Frequency	Fractions
0	stock-loss	409	0.316563
1	stock-gain	232	0.179567
2	product-release	111	0.085913
3	business-contract	63	0.048762
4	capital-increase	55	0.042570
5	legal-verdict-favored	42	0.032508
6	price-target-upgrade	40	0.030960
7	fundraising	33	0.025542
8	acquisition-interest-acquirer	33	0.025542
9	product-price-cut	30	0.023220