STQD6114 Task 1

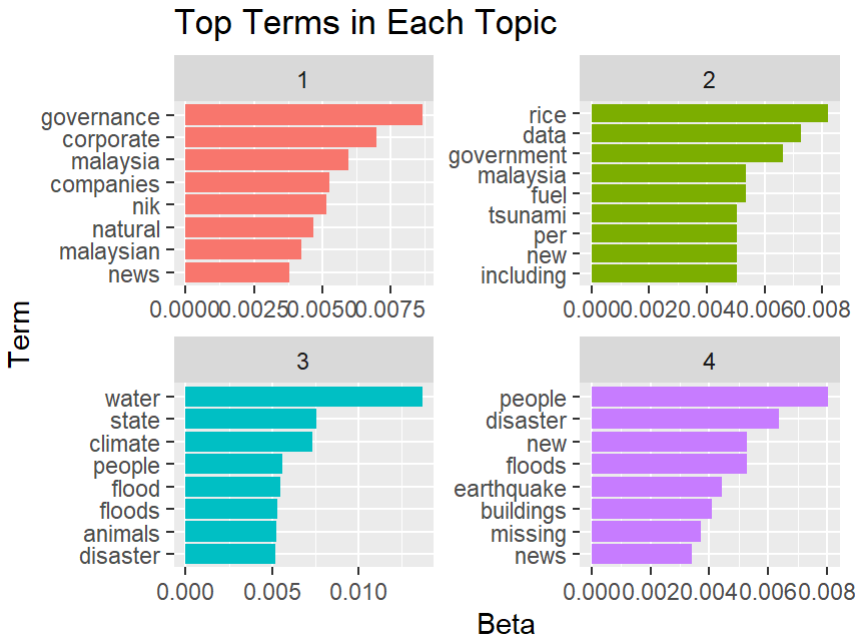
PANZHANGYU P136922

My files are all from New Straits Times, with the theme v. Natural disasters. I selected 50 news titles and text content and saved them in 50 text files in the folder Natural_disasters_news.

The first step is to extract the 8 most common terms in each topic, which can be achieved by the following code:

```
top_terms <- topic_terms %>%
  group_by(topic) %>%
  slice_max(beta, n = 8) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

Then visualize.



This visualization gives us an idea of the four topics extracted from the articles. The eight most common words in each topic are here, so let's analyze them one by one.

Overall, words like "government", "people", and "malaysia" are common in all four topics. It can be seen that these news are generally serious news about natural disasters in Malaysia with a certain degree of political significance. The "rice" in Topic 2 may be reports about food shortages caused by natural disasters. Topic 3 seems to be related to floods. Topic 4 is reports about earthquakes, floods, and building damage.

Then the topic term probability comparison:

```
beta_spread <- topic_terms %>%
  mutate(topic = paste0("topic", topic)) %>%
  pivot_wider(names_from = topic, values_from = beta) %>%
  filter(topic1 > .001 | topic2 > .001) %>%
  mutate(log_ratio = log2(topic2 / topic1))

beta_spread
#>> beta_spread
#> A tibble: 370 × 6
#>term                topic1       topic2       topic3       topic4 log_ratio
#><chr>                 <dbl>        <dbl>        <dbl>        <dbl>     <dbl>
#>   1 access          0.000812 1.89e-  3 6.84e-  4 5.32e-  4         1.22
#>2 according       0.00108    2.21e-  3 1.37e-  3 1.95e-  3         1.03
#>3 across          0.000627 1.26e-  3 1.52e-  3 1.24e-  3         1.01
#>4 activities      0.00189    6.31e-  4 6.84e-  4 1.10e-  8        -1.59
#>5 adapt           0.00108    2.78e-204 4.56e-  4 1.08e-209    -666.
#>6 added           0.00135    2.21e-  3 1.17e-  3 1.39e-  3         0.707
#>7 address         0.000754 1.26e-  3 6.84e-  4 3.81e-  5         0.744
```

#>8 addressing 0.00107 3.16e- 4 2.28e- 4 6.64e- 6 -1.76

#>9 advancements 0.00108 9.56e-205 1.18e- 6 1.76e- 4 -668.

#>10 advertisement 0.00108 3.16e-209 1.04e-206 3.61e-213 -683.

#> **i** 360 more rows

#> **i** Use `print(n = ...)` to see more rows

According to the results of beta_spread, we can see the most common terms in each topic and the difference in their probabilities between topics. For example, the term "access" has a higher probability in topic 1 (0.000812) and a lower probability in other topics (topic2: 1.89e-3, topic3: 6.84e-4, topic4: 5.32e-4), and its log_ratio is 1.22, indicating that the probability of this term in topic 1 is significantly higher than that in other topics.

And topic distribution:

```
document_topics <- tidy(lda_model, matrix = "gamma")
```

```
document_topics
```

#>> document_topics

#> A tibble: 200 × 3

#>document topic gamma

#><chr> <int> <dbl>

#> 1 1.txt 1 0.0000333

#>2 10.txt 1 0.0000350

#>3 11.txt 1 0.0000472

#>4 12.txt 1 0.0000552

#>5 13.txt 1 0.0000346

#>6 14.txt 1 1.00

#>7 15.txt 1 1.00

#>8 16.txt 1 0.0000333

#>9 17.txt 1 0.0000547

#>10 18.txt 1 1.00

#> **i** 190 more rows

#> **i** Use `print(n = ...)` to see more rows

Based on the result of document_topics, we can understand the probability of each document being assigned to different topics. For example, the probability of document "1.txt" being assigned to topic 1 is 0.0000333, while the probability of documents "14.txt" and "15.txt" being assigned to topic 1 is almost 1.00. This shows that some documents are more inclined to certain specific topics, while other documents may cover multiple topics.

Conclusion

Through topic modeling analysis, we have drawn some important findings about news coverage of natural disasters:

We identified four topics and found differences in term probabilities between them.

Some documents tend to be more specific than others, which may reflect the diversity and complexity of news coverage of natural disasters.

Further research can explore the correlation between different topics and conduct in-depth analysis of how each topic is represented in news coverage to better understand the coverage and response to natural disasters.