Question: The question: In the Reuters C50 text corpus, what words are the most frequently used after clustering the documents?

# The Reuters corpus

## Load the dataset

### Preprocessing

Approach:

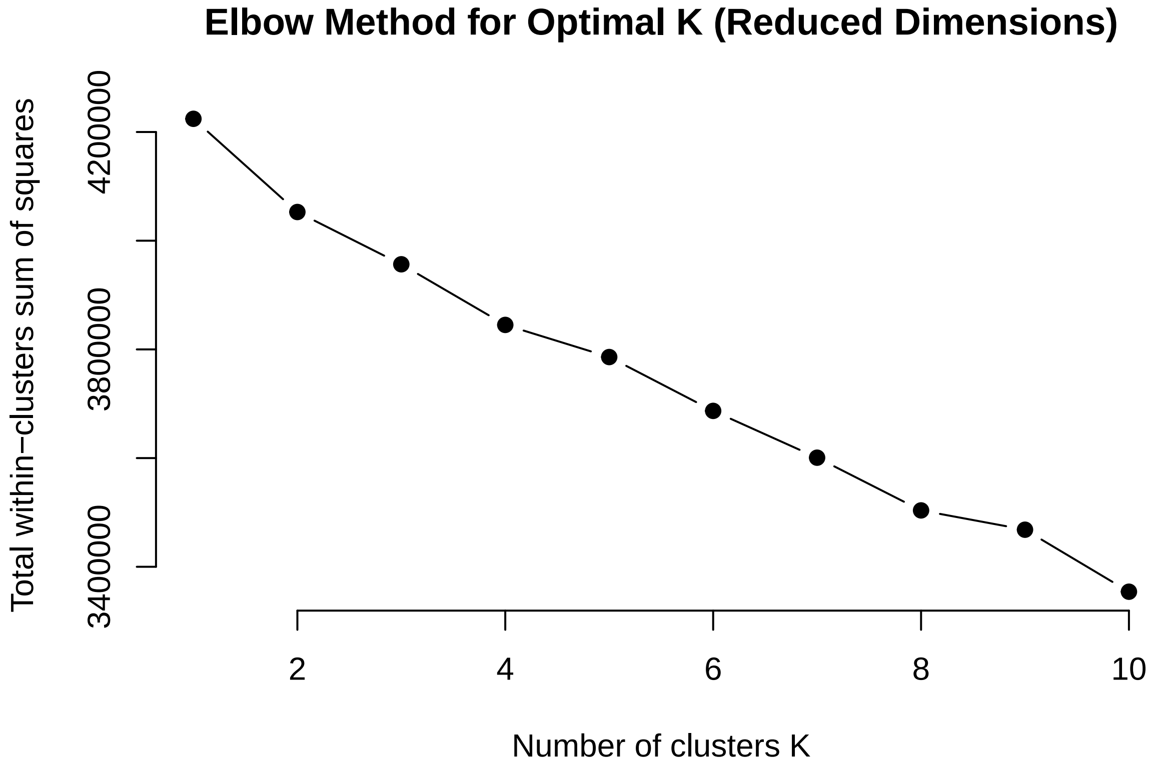
For this problem, I only used the training set because we are trying to analyze the existing data and find patterns in it.

I first preprocessed the data to prepare it for analysis. First, I converted all the text to lowercase for uniformity. Then, removed common words like "the" and "and" as well. I also removed punctuation, white spaces, and numbers to examine only the words. I also stemmed the words to their root form to reduce some dimensionality by grouping similar words. Finally, I used a tf-idf transformation to adjust term frequencies and reduce the influence of common words and highlight distinctive words.

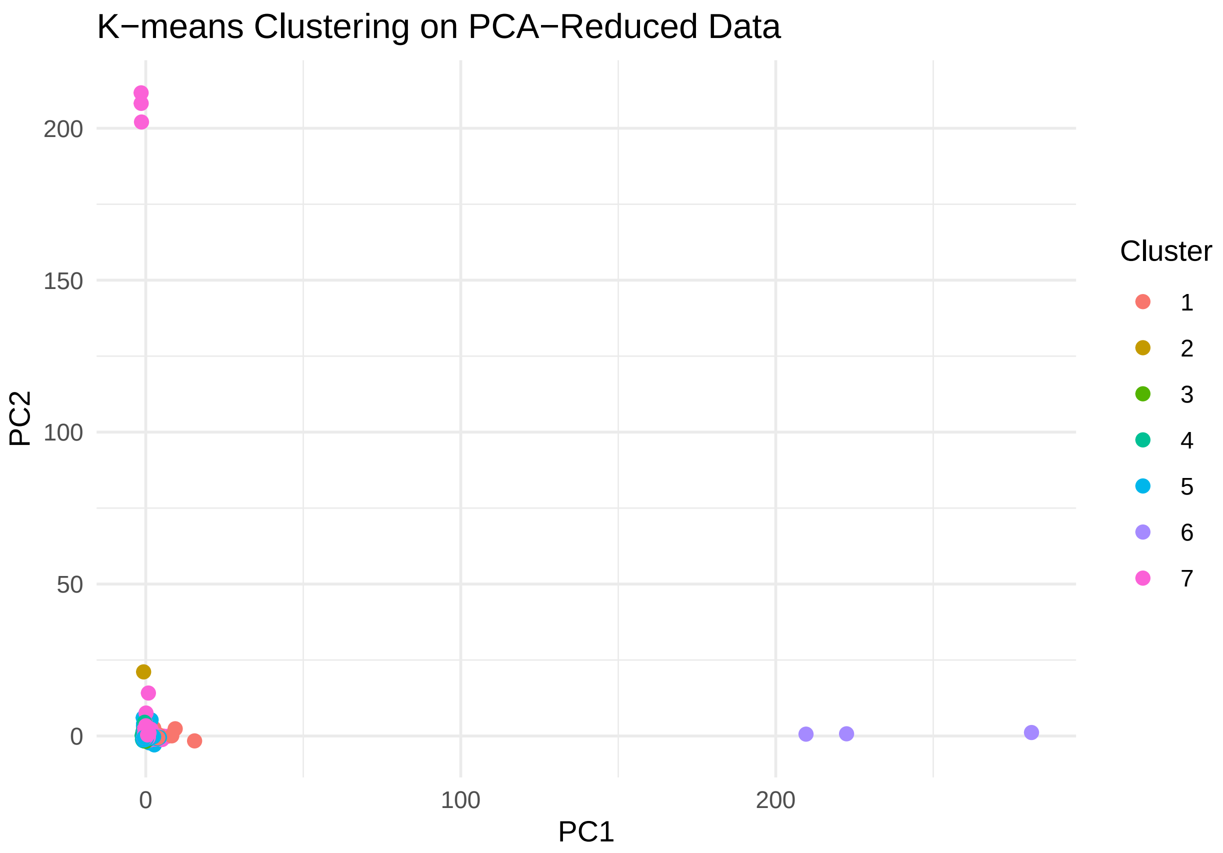
For analysis, the approach I used was both PCA and K-means clustering. Due to the high dimensionality of the dataset, I sought to reduce it significantly using PCA, reducing it to 50 components only. I then constructed an elbow plot find the optimal k for k-mean clustering, which was determined to be k = 7. Using k = 7, I conduct k-means clustering to determine the most frequently used words in each of the seven clusters. Additionally, I included word clouds to better illustrate the most common words in each cluster.

## Results

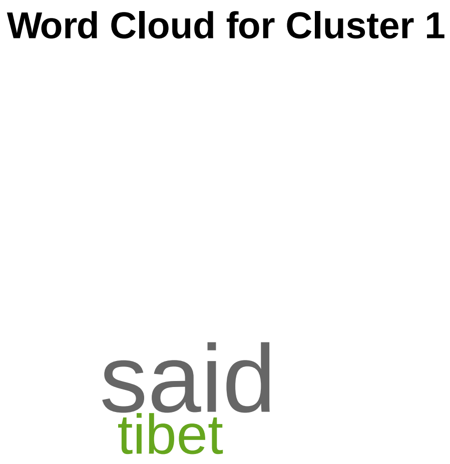
### K-means clustering



```
## The optimal number of clusters (K) is: 9
```

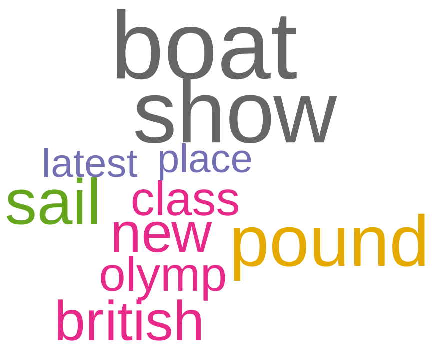


```
## Cluster sizes:

##
##    1    2    3    4    5    6    7
##   52    1    1  286 1737    3  420

## The optimal number of clusters (K) is:

## # A tibble: 82 x 3
## # Groups:   cluster [7]
##    cluster term     frequency
##      <int> <chr>        <dbl>
##  1       1 said          6.63
##  2       1 tibet         3.42
##  3       1 region        2.15
##  4       1 lama          2.06
##  5       1 dalai         2.02
##  6       1 offici        1.98
##  7       1 china         1.79
##  8       1 year          1.77
##  9       1 colombia      1.63
## 10       1 beij          1.62
## 11       2 boat         10
## 12       2 show          9
## 13       2 pound         7
## 14       2 sail          6
## 15       2 new           5
## 16       2 british       5
## 17       2 class         4
## 18       2 olymp         4
## 19       2 equip         3
## 20       2 place         3
## 21       2 year          3
## 22       2 say           3
## 23       2 tradit        3
## 24       2 latest        3
## 25       2 compet        3
## 26       2 cost          3
## 27       2 hightech      3
## 28       2 bullimor      3
## 29       2 dinghi        3
## 30       3 gum          11
## 31       3 chad         10
## 32       3 arab         10
## 33       3 world         6
## 34       3 year          6
## 35       3 produc        6
## 36       3 food          6
## 37       3 output        6
## 38       3 fao           6
## 39       3 said          5
## 40       3 tonn          5
## 41       4 said          7.80
## 42       4 will          3.53
## 43       4 compani       3.37
## 44       4 comput        3.07
## 45       4 new           2.80
## 46       4 internet      2.63
## 47       4 year          2.49
## 48       4 corp          2.21
## 49       4 analyst       2.03
## 50       4 servic        1.96
## 51       5 said          8.20
## 52       5 year          2.66
## 53       5 percent       2.64
## 54       5 compani       2.62
## 55       5 million       2.48
## 56       5 will          2.37
## 57       5 market        2.26
## 58       5 analyst       1.89
## 59       5 share         1.87
## 60       5 bank          1.73
## 61       6 czech         7.33
## 62       6 spain         6
## 63       6 team          5.33
## 64       6 pavel         5
## 65       6 two           4.33
## 66       6 first         3.67
## 67       6 raul          3.33
## 68       6 play         3
## 69       6 striker      3
## 70       6 win          2.33
## 71       6 berger       2.33
## 72       6 game         2.33
## 73       7 said         7.17
## 74       7 china        4.21
## 75       7 hong         2.92
## 76       7 kong         2.89
## 77       7 beij         2.14
## 78       7 year         2.01
## 79       7 chines       1.95
## 80       7 will         1.81
## 81       7 offici       1.66
## 82       7 state        1.49
```

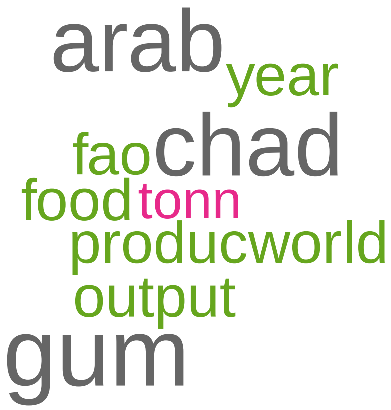


Word Cloud for Cluster 1

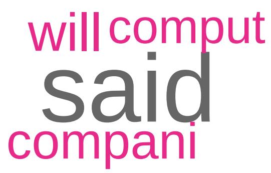
## Word Cloud for Cluster 2



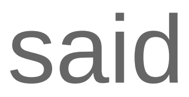
## Word Cloud for Cluster 3



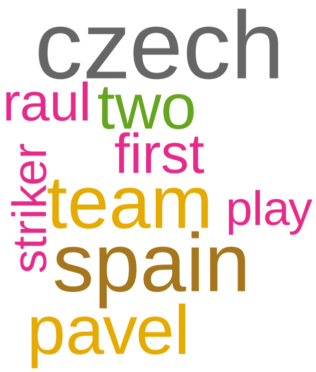
## Word Cloud for Cluster 4

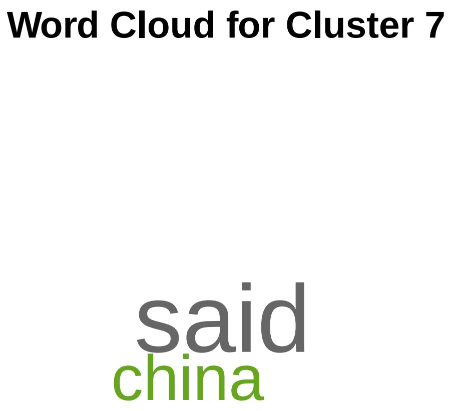


**Word Cloud for Cluster 5**



## Word Cloud for Cluster 6





Word Cloud for Cluster 7

Conclusion:

Looking at the word clouds for the seven clusters, said seems to be extremely common. This seems to suggest that much of the text pertains to quotes, perhaps quoting people.

Cluster 1: The most frequent words seem to pertain to countries, such as Tibet, China, and Colombia.

Cluster 2: The most frequent words seem to pertain to boats, with words like “boat” and “sail”, and Britain, with words like “british” and “pound”.

Cluster 3: The most frequent words seem to pertain to countries in Africa/Middle East, with words like “chad” and “arab” and “world” the most common.

Cluster 4: The most frequent words seem to pertain to companies, with terms like “compani” “corp” and “analyst” being the most common.

Cluster 5: The banking industry seems to be relevant with words like “bank”, “market”, and “share” appearing.

Cluster 6: The most frequent words appear to be pertain to countries, with “czech” and “spain” as number 1 and 2. Also, there appears to be a relation to sports, with words like “team”, “play”, “striker”, and “win”.

Cluster 7: The most frequent words appear to be related to China, with “china”, “hong”, “kong”, “beij”, “chines” appearing as frequent words.

If presenting this data to stakeholders, one can conclude that the most common themes in the Reuters C50 text corpus are various countries, the economy, and corporations. As such, it may provide valuable information about events pertaining to certain countries (particularly China, Tibet, Colombia, and Chad), markets and stocks, and the business/working world.