COMP4432 – Machine learning Project

A million headlines

20066851D Li Man Chung

18057243D Wong Man Kit

20028439D Lam Ka Chun

1. **Introduction**

**1.1 Goal**

This project is aimed at finding a way to cluster a headline dataset with over a million entries of headlines.

To do this, we should find a way to i) transform the headlines into trainable arrays, ii) cluster the headlines successfully, and iii) evaluate the performance of the clustering. We used vectorizer and dimensionality reduction method, k-means clustering method, and performance evaluation by introducing a labeled headlines dataset.

**1.2 Web source used**

i. Million headlines dataset: abcnews-date-text.csv

*Source: https://www.kaggle.com/datasets/therohk/million-headlines*

ii. Kaggle notebook for K-means clustering

*Source: https://www.kaggle.com/code/thebrownviking20/k-means-clustering-of-1-million-* *headlines/notebook*

iii. Labelled News Category Dataset: News\_Category\_Dataset\_v2.json

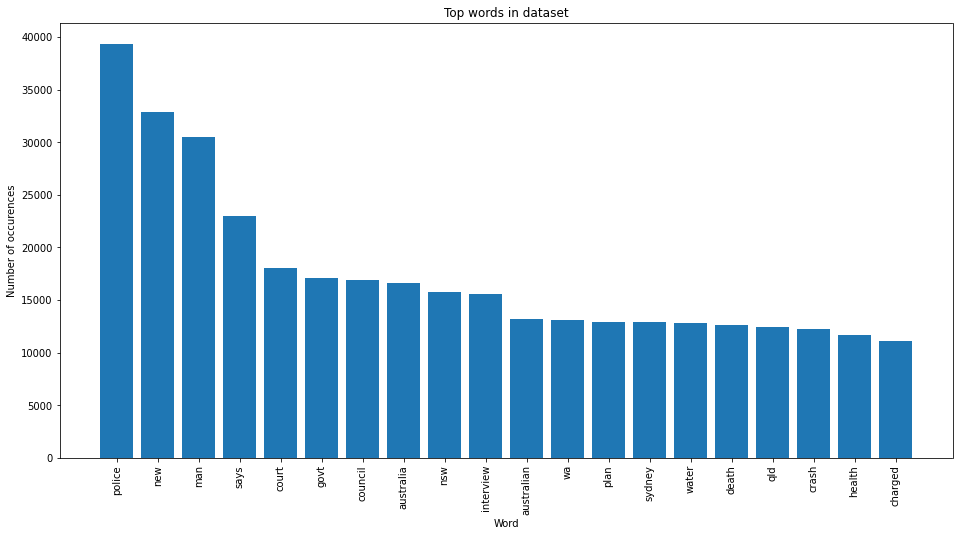
*Source: https://www.kaggle.com/datasets/rmisra/news-category-dataset*

**1.3 Libraries**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| numpy | pandas | matplotlib | sklearn | nltk | wordcloud |

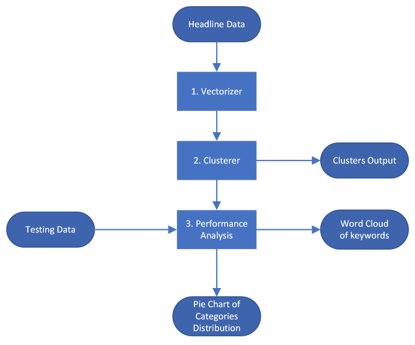
**1.4 Insights on the data**

Before everything else, we may find some insights on what the results could be from analyzing the given data. For example, we have done an analysis on the occurrence of words in the dataset.

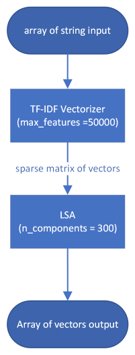


In the headlines data, the most frequently used word is “police”, which means in that specific period of 18 years, there is a high possibility that most of the news is related to the words with highest occurrence, like “police”, “new” and “man”. Since words like “new”, “man”, and “says” may not contain distinguishable information about the nature of the news, we may expect most of the headlines would be related to “police”, “court”, “government”, more specifically, law, order, and policies in societies.

1. **Overall workflow design and implementation**



**2.1 Vectorizer and Dimensionality Reduction (DR) Method implementation**



The goal of this part is to i) turn string into array of vectors and ii) reduce the dimensions of the array to reduce the training time and increase performance.

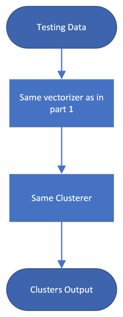
Tf-idf vectorizer is selected for string vectorizer. Stop words and tokenizer from nltk has been use along with the vectorizer.

For DR, several selections are available, includes i) assigning n\_features in the tf-idf vectorizer, ii) principal component analysis (PCA), and iii) latent semantic analysis (LSA). The final implementation was using LSA with n\_components =300, because LSA has better performance on sparse matrix (outputted by the vectorizer) than PCA with the concern of running time, and LSA perform singular value decomposition on data, instead of counting occurrence of features in tf-idf vectorizer.

**2.2 Clustering**

With the inspiration of the Kaggle notebook (Web source ii), k-means clustering is easy to implement, easy to understand and has a good performance for us to reference on, so it is selected as the classifier.

**2.3 Performance Evaluation**

 The labelled news category dataset is introduced here to give us a hint on what the clustering is doing.

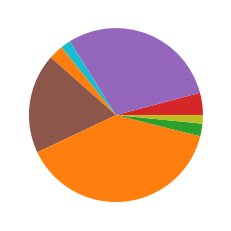
To do so, we first think of training a classifier based on the labelled data, to classify the million headlines in order to compare the clusters with the categories. We attempted it with MLPClassifier, SVMClassifier, knn-classifier. But since the training takes a lot of time, low accuracy on classification (all attempts ~30%), we decided not to take this approach.

Our final implementation is just simply cluster the testing headlines with the same process as we did to the million headlines, we want to know how the kmeans clustering model cluster different categories of news by it’s headlines. Analysis from this will be done in the next section.

**2.4 Distribution analysis of categories for each cluster**

It would be normal to think that successful clustering will have obviously differences in terms of the nature of the news, which can be seen from the distribution of categories of the labelled headlines in each cluster.

Pie chart is drawn on each cluster for how labelled categories distributed in it. For example,



Custer 0 has *wellness*, *politics*, and *healthy living* as the top-3 categories in it, it would be obvious to understand that it is a health-related cluster.

However, as you may see in the next parts, the unbalanced news category data has a result of occupation of several categories among all clusters, this is an issue that may affect the analysis of the clusters.

**2.5 Word cloud of keywords in each cluster**

一張含有 文字 的圖片

自動產生的描述By counting the occurrence of words in each cluster, we may extract the keywords for the clustering, and perform them nicely with wordcloud library. (Right figure: word cloud of cluster 0)

1. **Clustering Results and Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Keywords | Distribution | Top-5 Categories |
| 0 | |  | | --- | | 一張含有 文字 的圖片  自動產生的描述 | |  | **wellness**, politics, **healthy living**, impact, business |
| 1 | 一張含有 文字 的圖片  自動產生的描述 |  | **politics**, entertainment, the worldpost, wellness, crime |
| 2 |  | 一張含有 風車, 向量圖形 的圖片  自動產生的描述 | **politics**, entertainment, style & beauty, wellness, travel |
| 3 | 一張含有 文字 的圖片  自動產生的描述 |  | **politics**, queer voices, parenting, wellness, entertainment |
| 4 | 一張含有 文字 的圖片  自動產生的描述 |  | politics, **the worldpost**, **travel**, **world news**, **worldpost** |
| 5 | 一張含有 文字 的圖片  自動產生的描述 |  | **crime**, politics, weird news, queer voices, entertainment |
| 6 | 一張含有 文字 的圖片  自動產生的描述 |  | politics, wellness, entertainment, travel, style & beauty |
| 7 | 一張含有 文字 的圖片  自動產生的描述 |  | **crime**, politics, queer voices, black voices, entertainment |

**3.1 Distinguishing clusters**

Cluster 0 has numerous keywords related to health and a great portion of health-related categories should be classified as a health-related cluster.

Cluster 4 has many direction-related and geographical keywords, and has a great portion of worldwide news and travelling categories, it should be classified as a travel or world-news cluster.

Cluster 5 and 7 have many crime-related keywords, such as murder, assault and die, they should both be crime-related categories.

In the remaining clusters, the differences are not very obvious and they should be politics-related. There are minor differences to categorize them into sub-categories of politics, like cluster 3 may be humanity and human-rights related policy news

1. **Other information on the details of implementation**

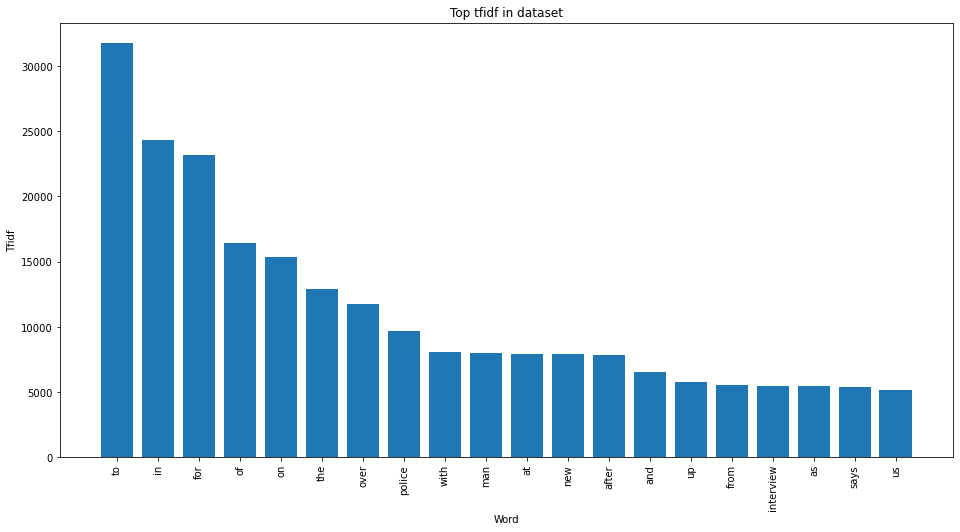
**Tf.idf matrix**

After visualizing the headline data using term frequency, tf.idf matrix is implemented for further understanding the dataset.



The function we used to calculate term frequency is modified to calculate the tf.idf.

The following graph shows the sum of tf.idf of the whole dataset.



From the graph, we discover the problem of stop words. words like “to”, “in” have no meaning and cannot represent the documents. Therefore, we implemented stop word lists in the vectorizer of the tf.idf matrix.



The stop words list is downloaded from the nltk library, we also union the punctuation with the stop words.

Moreover, we modified the tokenizer and stemmer for the vectorizer. Porter Stemmer and Regular-Expression tokenizer are used.

After stemming some words will be stemmed and grouped into meaningless words. For example, “police” will become “polic””. Although the word may cause ambiguity for the keywords for visualization, the stemming can reduce the feature’s dimensions effectively.

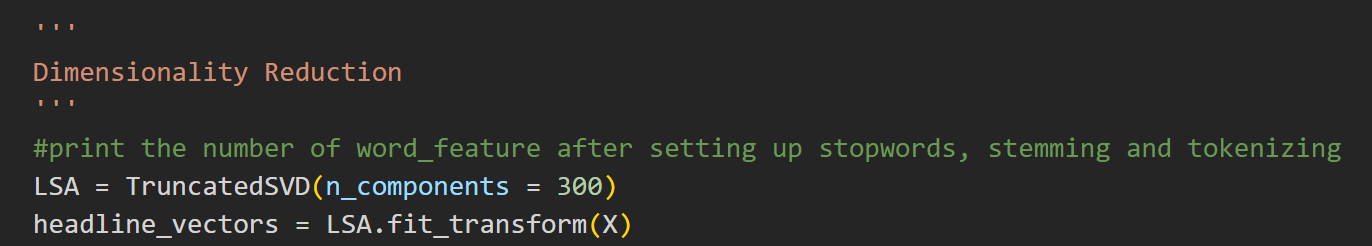
The number of features dropped from 104691 to 77058.

**Dimensionality Reduction**

Due to the large dimensions of the feature, the training time for machine learning will be large.

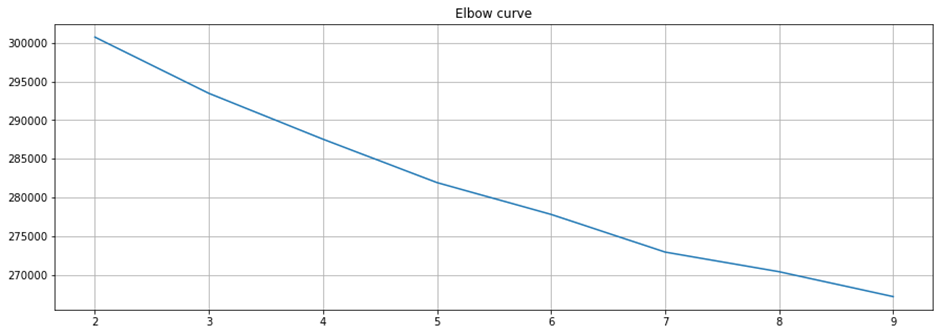
To reduce the dimensions, the max\_features for the vectorizer are set at 1000, only the top 1000 frequent words are chosen to form the matrix.

On the other hand, we performed the latent semantic analysis and reduced the components to 300.



**K means clustering**

After the feature extraction, we use K means for the implementation of the headline news clustering task. For K means clustering, the choosing of the value of K is important. We used the elbow method to examine the results of different values of k.

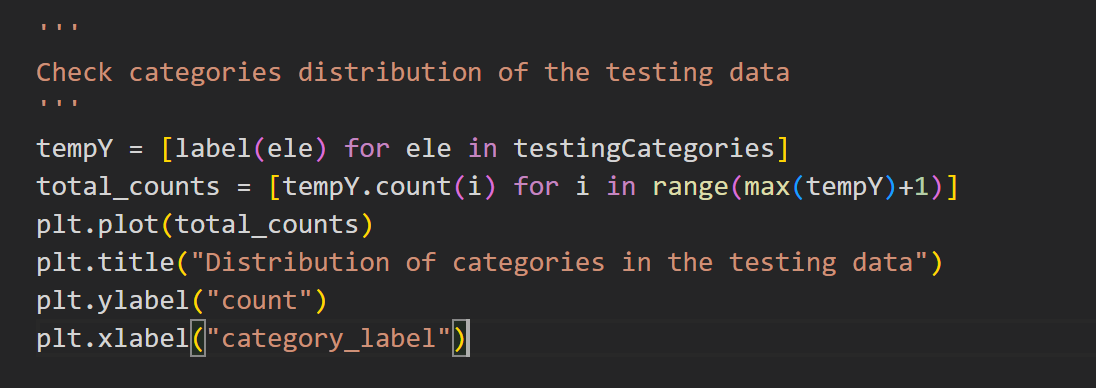


From the above graph, the elbow curve is at k=7.

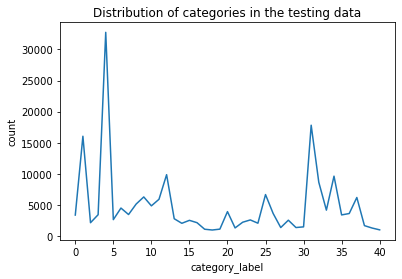
**Performance Analysis**

In addition to the original dataset, "abcnews-date-text.csv" is used as testing data for the performance analysis of the clustering method.

The following code shows the plotting of the distribution graph:



And here is the result of the distribution graph:



The distribution of categories in the testing dataset is very unbalanced, therefore most of the clusters has a great portion of results belonging to these categories. (1: Politics, 4: Entertainment,