BA820 Final Deliverable

Team member:

Yantao Wang, Shiqi Sun, Chuning Chen, Chenli Qiu

Business Problem

Mashable is an American culture, tech, science and social good digital media platform, news website and multi-platform media and entertainment company. It won the 1st International Open Web Awards to recognize the best online communities and services on 27 November 2007. Mashable has more than 6,000,000 twitter followers and over 3,200,000 fans on Facebook.

Mashable has many channels to provide diversified online services such as video, entertainment, culture, tech, science and social good. In these years, Mashable also has a huge user group from Twitter and Facebook, meanwhile, there are numerous resources on the website. Users with different interests, races and nationalities are surfing Mashable everyday.

We are a start-up company which wants to create a platform for news and media service. At this time, we do not know which kinds of articles and news should we focus on. In addition, we also want to understand what determines the popularity. What's more, we also need to decide the target customers for our platform. As a result, we want to analyze the popularity of different articles of Mashale to improve our company services' exposure and ensure our products can be viewed by more people.

Dataset

Our UCI Online News Popularity dataset was downloaded from Kaggle.
https://www.kaggle.com/thehapyone/uci-online-news-popularity-data-set
The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content is publicly accessed and retrieved using the provided URLs in the dataset.

In our analysis, we are planning to focus on some relevant attributes of the articles in the dataset. Some of these attribute information are listed as follows:

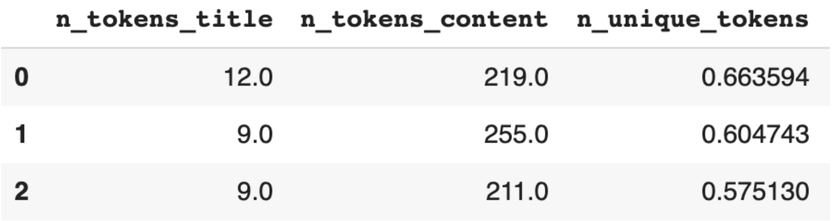
- n_tokens_title: Number of words in the title
- n_tokens_content: Number of words in the content
- n_non_stop_unique_tokens: Rate of unique non-stop words in the content
- num_imgs: Number of images

- num_videos: Number of videos
- average_token_length: Average length of the words in the content
- num_keywords: Number of keywords in the metadata
- data_channel_is_lifestyle: Is data channel 'Lifestyle'?
- data_channel_is_entertainment: Is data channel 'Entertainment'?
- kw_min_min: Worst keyword (min. shares)
- kw_max_min: Worst keyword (max. shares)
- kw_avg_min: Worst keyword (avg. shares)

This dataset consists of 61 attributes in total. (58 predictive attributes, 2 non-predictive, 1 goal field)

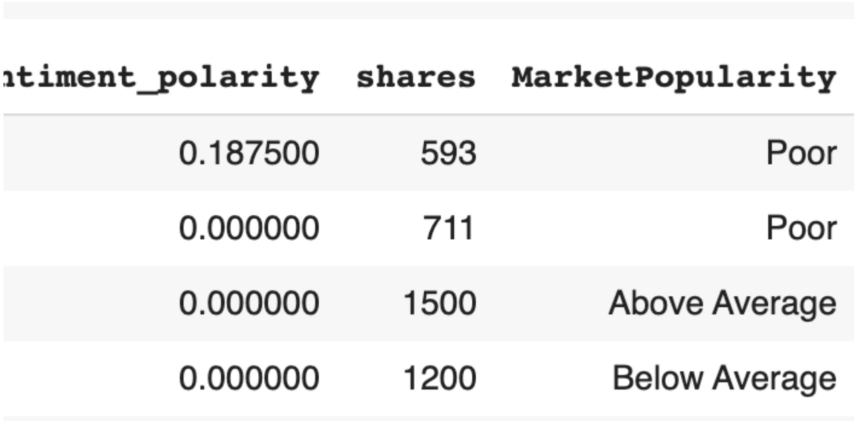
Data Processing

We removed ‘timedelta’ and ‘url’ columns because these two columns are useless for future prediction. Though there are no null values in our dataset, some zero values are meaningless. As a result, we also remove those zero values to keep the data clean and meaningful.

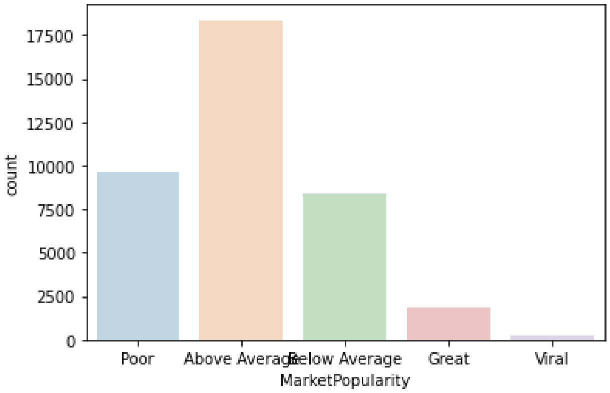


	n_tokens_title	n_tokens_content	n_unique_tokens
0	12.0	219.0	0.663594
1	9.0	255.0	0.604743
2	9.0	211.0	0.575130

We add a ‘MarketPopularity’ column to the dataset to label the articles with the ‘shares’ variable. A rank is created for article popularity. We labeled articles that have share less than 946 as “poor performance”, between 946 and 1400 (median) as “below average”, between 1400 and 10000 as “above average”, between 10000 and 40000 as “great performance”, else are labeled as “viral spreading performance”.



ntiment_polarity	shares	MarketPopularity
0.187500	593	Poor
0.000000	711	Poor
0.000000	1500	Above Average
0.000000	1200	Below Average

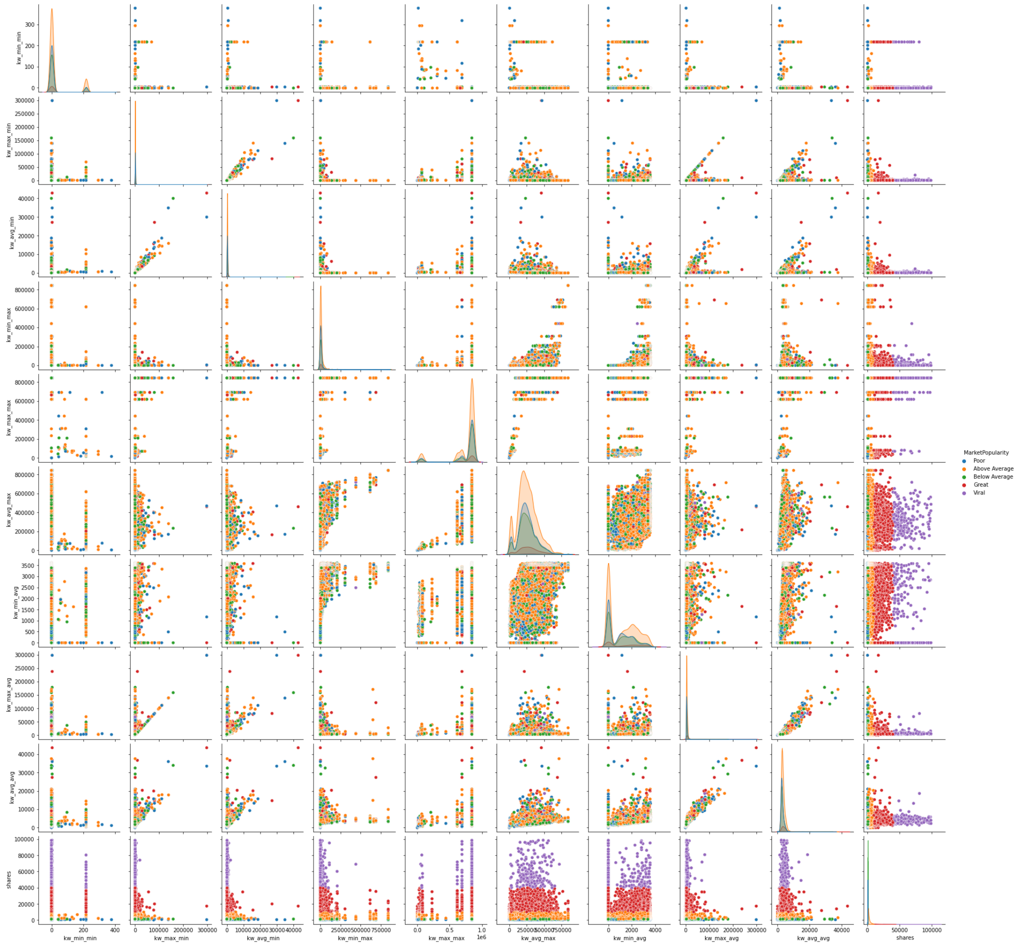


We check the class balance with the count plot. This plot indicates that about a quarter of the articles have poor performance, and less than a quarter of those have performance below average. Nearly one half of the articles have performance above average. About 2000 (5% of the articles) have great performance. Finally, very few articles have viral spreading performance.

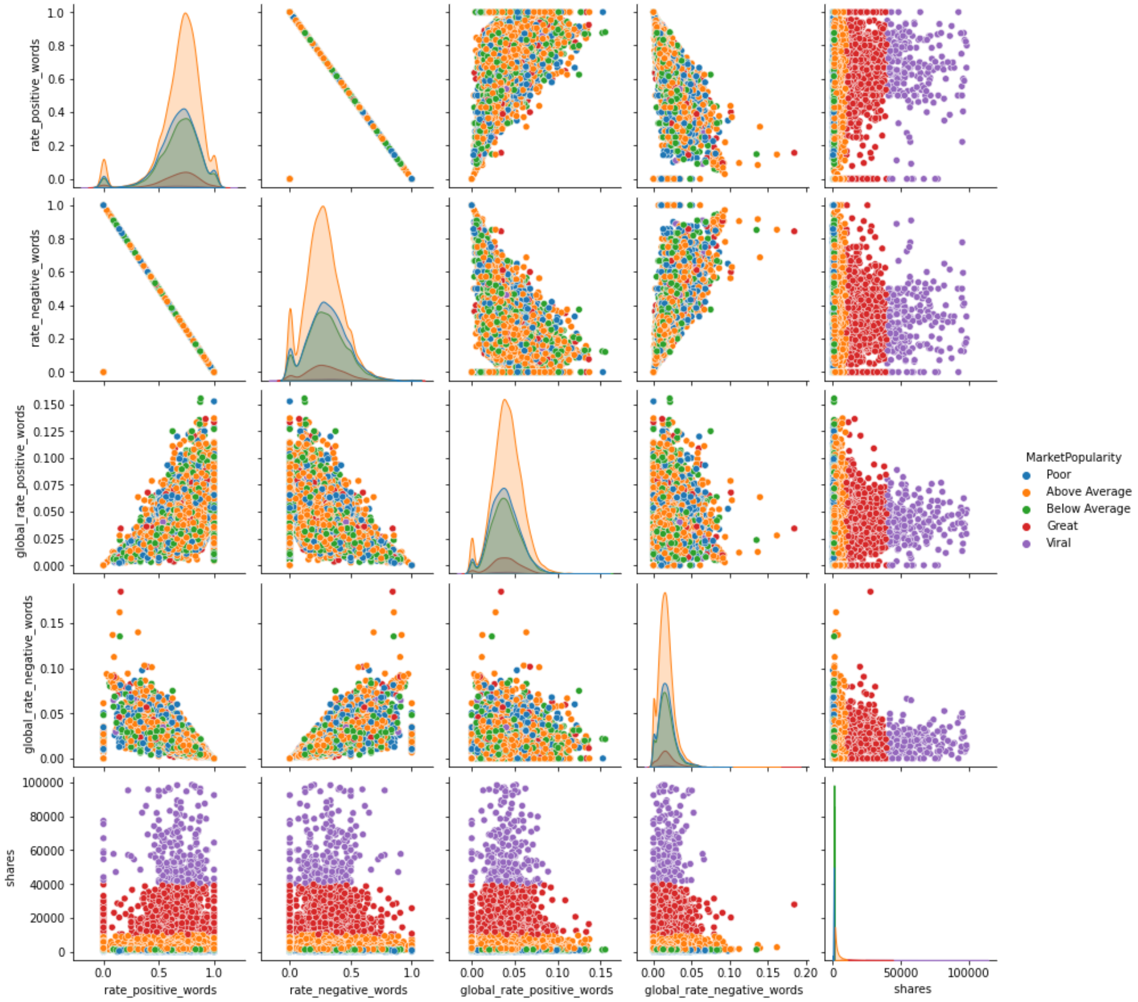
Exploratory Analysis

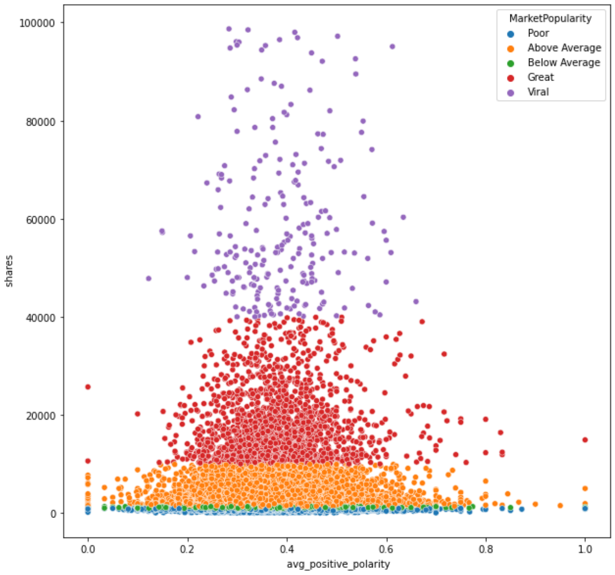
After ranking up the MarketPopularity, we draw a pairplot for the data to find the relationships among the keywords and shares.

By looking at the pairplot, we can estimate from the plot that for the popular and unpopular resources online, there always exist some words. Therefore, providers may need to consider to use some words frequently which can increase the popularity evaluated from the graph while avoiding to use some other words that exist in unpopular resources.



By looking at this pairplot, we found that there is a linear relationship between rate_negative_words and rate_positive_words which means there is no special relationship between this variable. While we looked at share VS positive/negative words and popularity we found that there is a slight relationship with shares. Positive words are slightly more with shares.

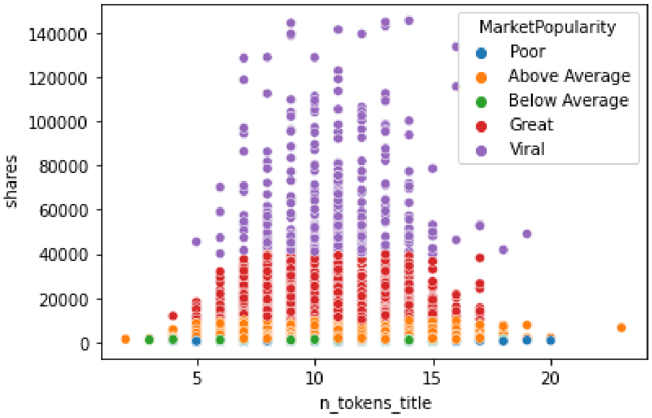


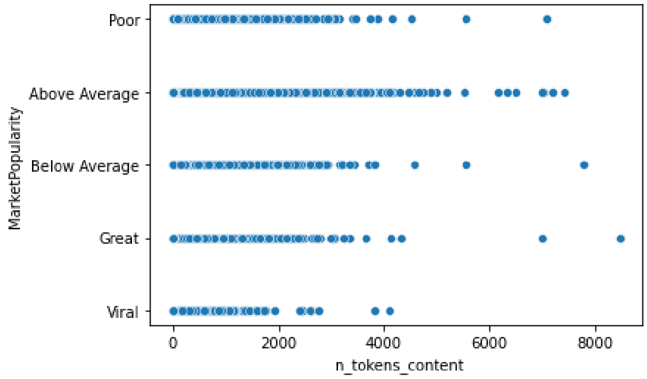


We can estimate from this plot that most online polulatries are centralized between 0.2 to 0.7 and nearly all of them’s market popularity are above average. What’s more, with the higher popularity of the market, the shares will become more considerable. It is essential to online service providers to evaluate their own business market and stimulate them to create more funny and creative things on the internet.

From the scatterplot above, it can be seen that articles with well performances generally have 6 to 17 words in their titles.

Our inference is that a moderate title length is more convenient for people to remember and recite.



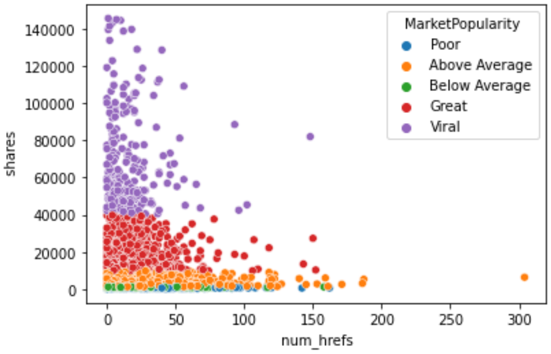


From the plot we can find that better performed articles tend to have less number of words in the content. Most articles with great performance have word length less than 3000, and most viral spreading articles have word length less than 1500.

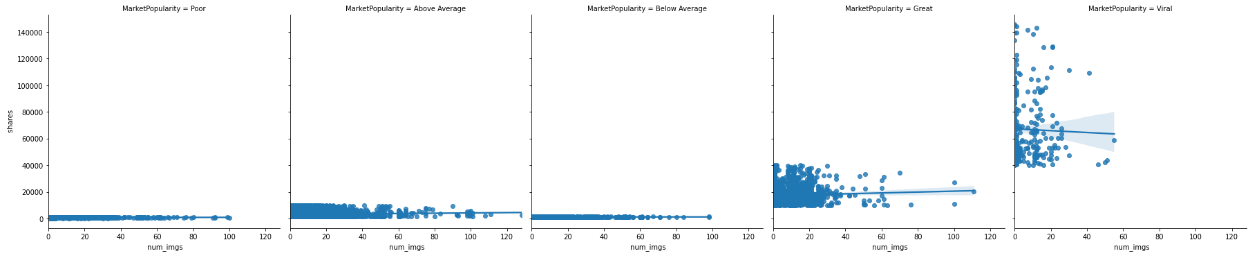
Our speculation is that readers are unwilling to spend too much time reading one single article, so fewer words is beneficial for the spread of an article.

This plot indicates that better performed articles tend to have less number of links in the content.

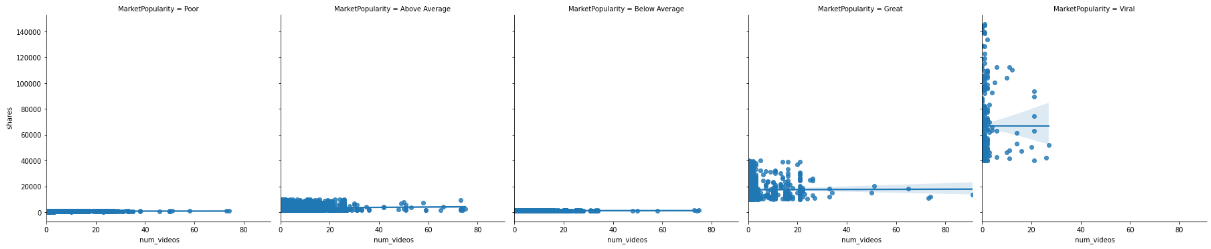
We would recommend reducing the number of referenced links in the article, and try to keep the number below 40.



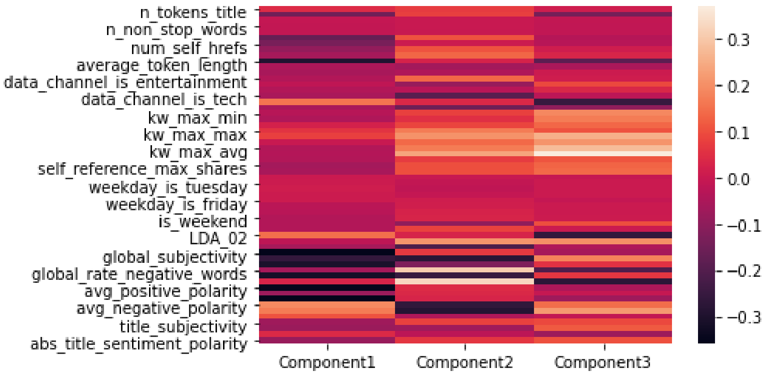
We have guessed that more images attached may make an article more vivid and more appealing to readers. However, beyond our expectation, too many pictures don't seem to make an article more popular. We would suggest that articles should keep numbers of images below 30 in order to have a better performance.



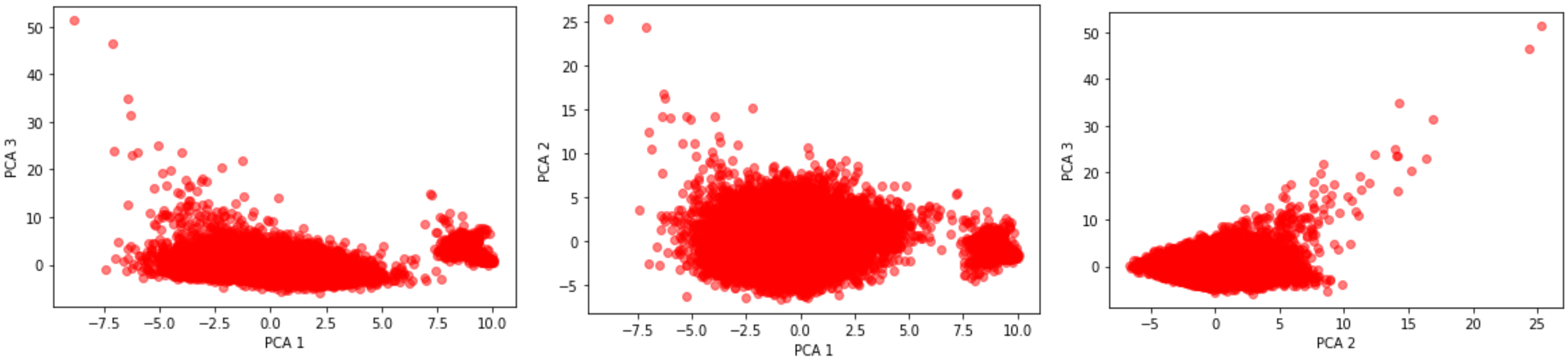
Similar to the number of images in an article, attaching a large number of videos to an article will also have a negative effect on the popularity. We would suggest keeping the number of videos under 20.



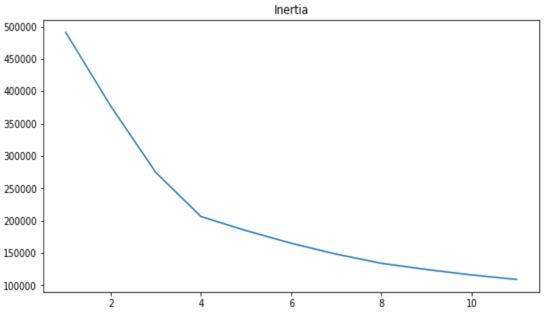
Build Different Processes to Test Below Models and Their Performances

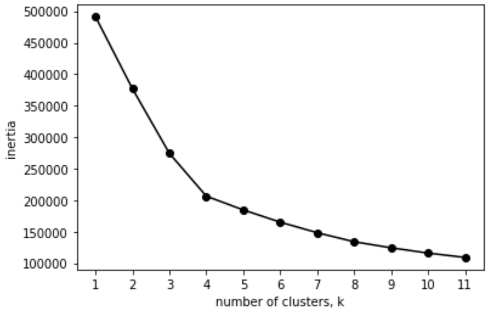


After normalizing and scaling data, we apply PCA for reduction to the components. We set n_component to 3 for better performance. We made a headmap to see how the features mixed up. In component 1 we found avg_negative_polarity contribute more, in component 2 we found global_rage_negative_words contribute more, and in component 3 we found kw_max_avg contribute more.

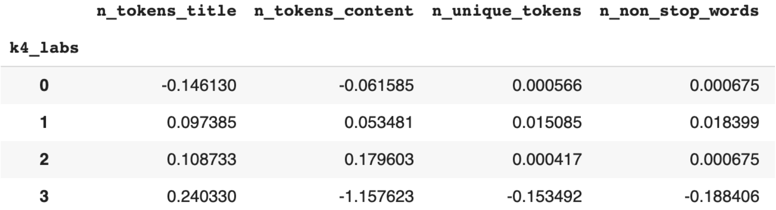


For PCA 1 and 2, with PCA 1 increase, PCA 2 decreases. For PCA 1 and PCA 3, it shows a similar relation with PCA 1 and 2, which is when PCA 1 decreases and PCA 3 increases. For PCA 2 VS PCA 3, with PCA 2 increase, PCA 3 increases.

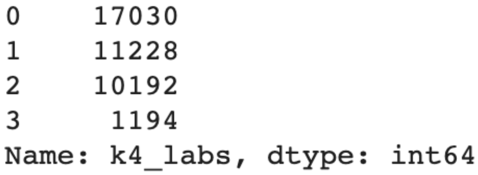




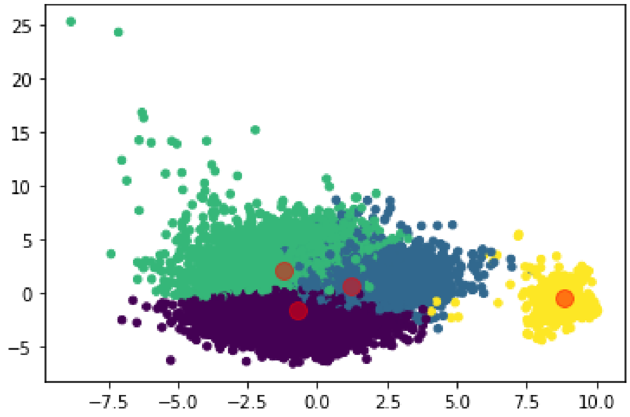
After that, we use Kmean iterations to find the best Kmean model we can train. Considering the elbow method and the inertia graph we think that we select 4 clusters to do the Kmean will have a better performance because the slope changes the fastest when the clusters equal to 4, which means that if we choose clusters more than 4, the performance of our model will not have significantly improvement but the computation will become lager.



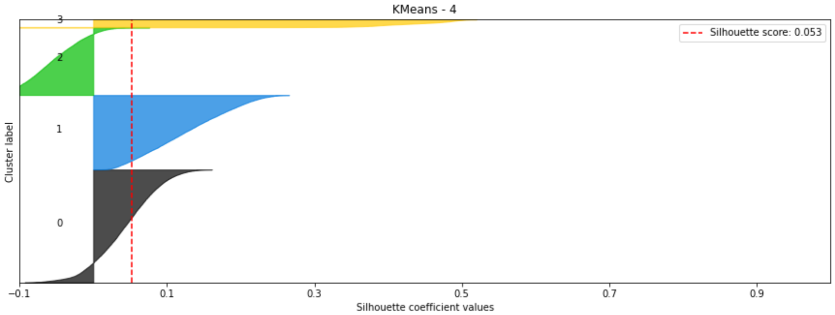
	n_tokens_title	n_tokens_content	n_unique_tokens	n_non_stop_words
k4_labs				
0	-0.146130	-0.061585	0.000566	0.000675
1	0.097385	0.053481	0.015085	0.018399
2	0.108733	0.179603	0.000417	0.000675
3	0.240330	-1.157623	-0.153492	-0.188406



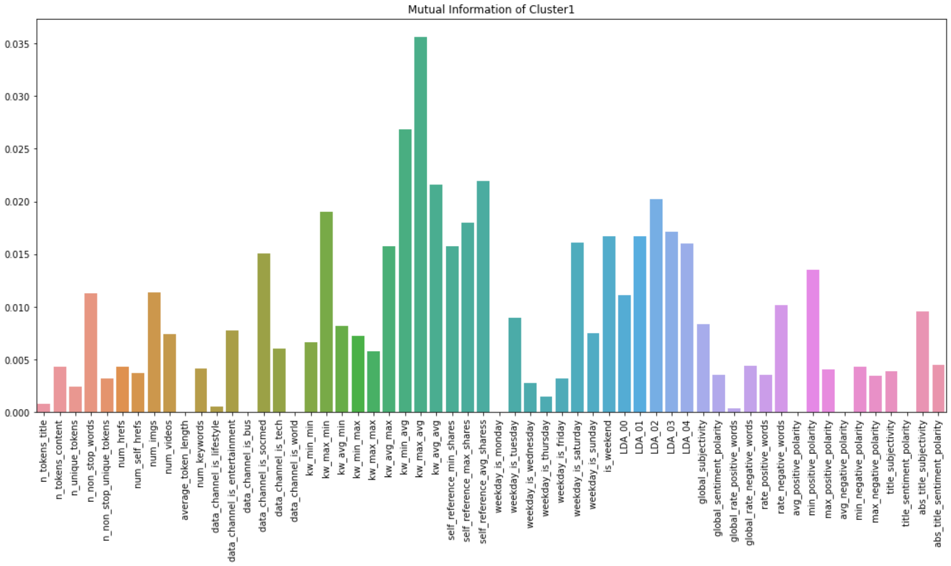
```
0     17030
1     11228
2     10192
3      1194
Name: k4_labs, dtype: int64
```

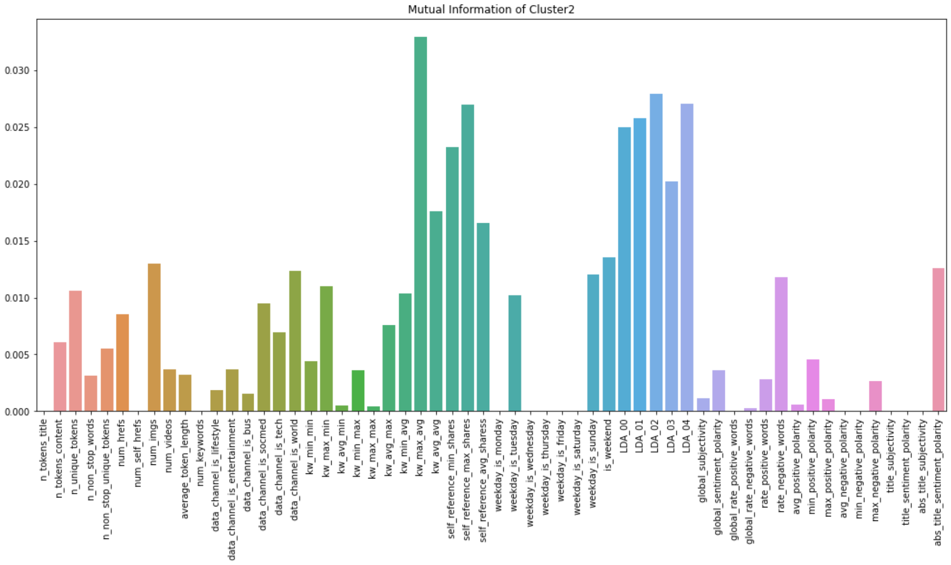


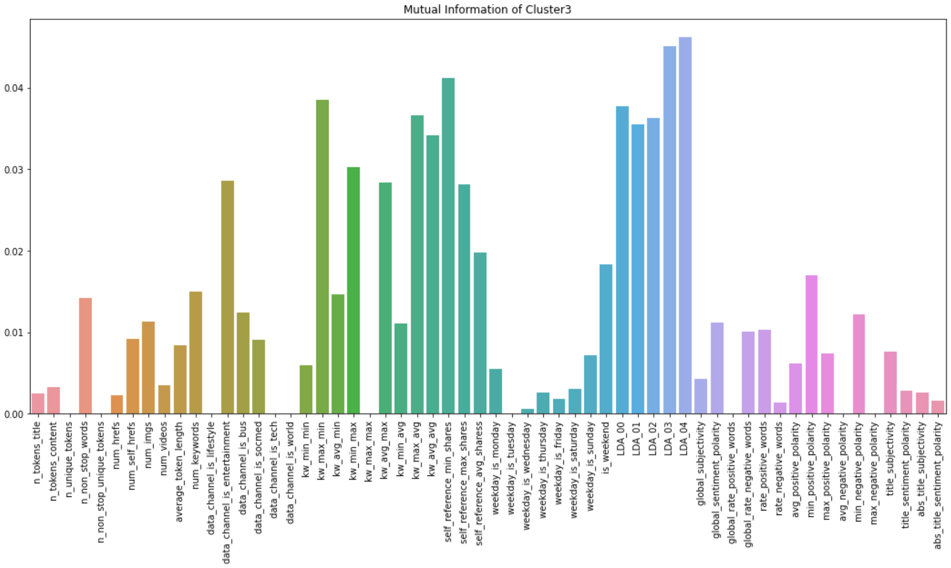
By using cluster centers, we found our dataset is grouped into 4 different clusters. Cluster 1 contains 17030 data, cluster 2 contains 11228 data, cluster 3 contains 10192 data, cluster 4 contains 1194 data.

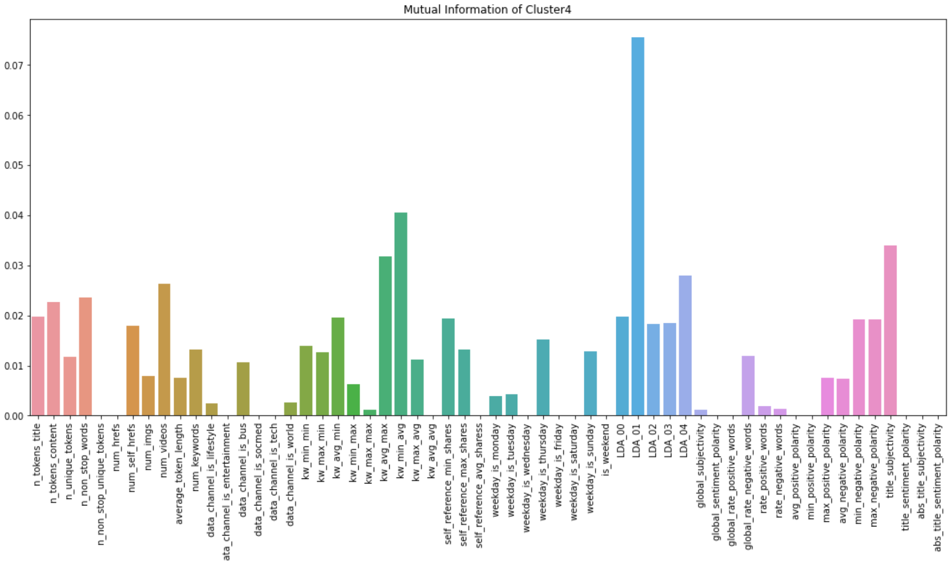


This is the Kmean plot for our model to choose 4 clusters, the Silhouette score of the model is 0.053 and there still have some points less than 0, which means that some points in our models do not fit well for the 4 clusters we choose especially for the level 3 cluster. It may become our future work to choose a better model to reduce the number of negative values and increase the Silhouette score.









The four plots are the mutual information analysis, it can be estimated from the graphs that there are several attributes such as n_non_stop_words, date_channel_is_entertainment, kw_max_avg, self_reference_min_shares, LDA_01, min_negative_polarity, etc that influence the popularity of the online streaming resources in common among 4 clusters. Combined all graphs we have we can evaluate that for good resources, we may need less stop words and focus on some hot topics such as entertainment, business and social media. The online service providers also need to consider which kind of words to use. The mutual information shows that popular and unpopular words also play important roles to influence the popularity of the news. Therefore, the providers are expected to add more popular words that publics love to read and reduce some words which publics will feel uncomfortable and unwill to read. For the LDA evaluation, our dataset has 5 topics and the 1-3 clusters fit all topics related to the LDA evaluation. For cluster 1, it fits the topic 2 better. For cluster 2, it fits topic 2 and topic 4 better. For cluster 3, it fits the topic 3 and 4 better. For cluster 4, it fits the topic 1 the best while it has less relationship to the topic 0, 2 and 3.

Conclusions & recommendations:

For our project, we are using the unsupervised machine learning method to find the popularity prediction. In the very beginning, we cleaned the data and removed useless prediction variables. By intuition, we thought that articles with larger market shares should be the articles with high popularity. Though the model cannot tell us what kind of articles are popular on the market, we can understand what components are important for popular articles with large market shares. Based on Kmean analysis, we draw the mutual information plots on the four clusters. The articles with the following recommendations are more popular on the market:

- Less kw_min_max, More market shares for the articles.
- More rate_positive_words & Less rate_negative_words, More market shares for the articles.
- The articles which focus on hot social issues gain more market shares.
- Have titles between 6 and 17 words. (n_tokens_title)
- Have word lengths less than 1500. (n_tokens_content)
- Have less than 40 references. (num_hrefs)
- Have less than 30 images. (num_imgs)
- Have less than 20 videos. (num_videos)