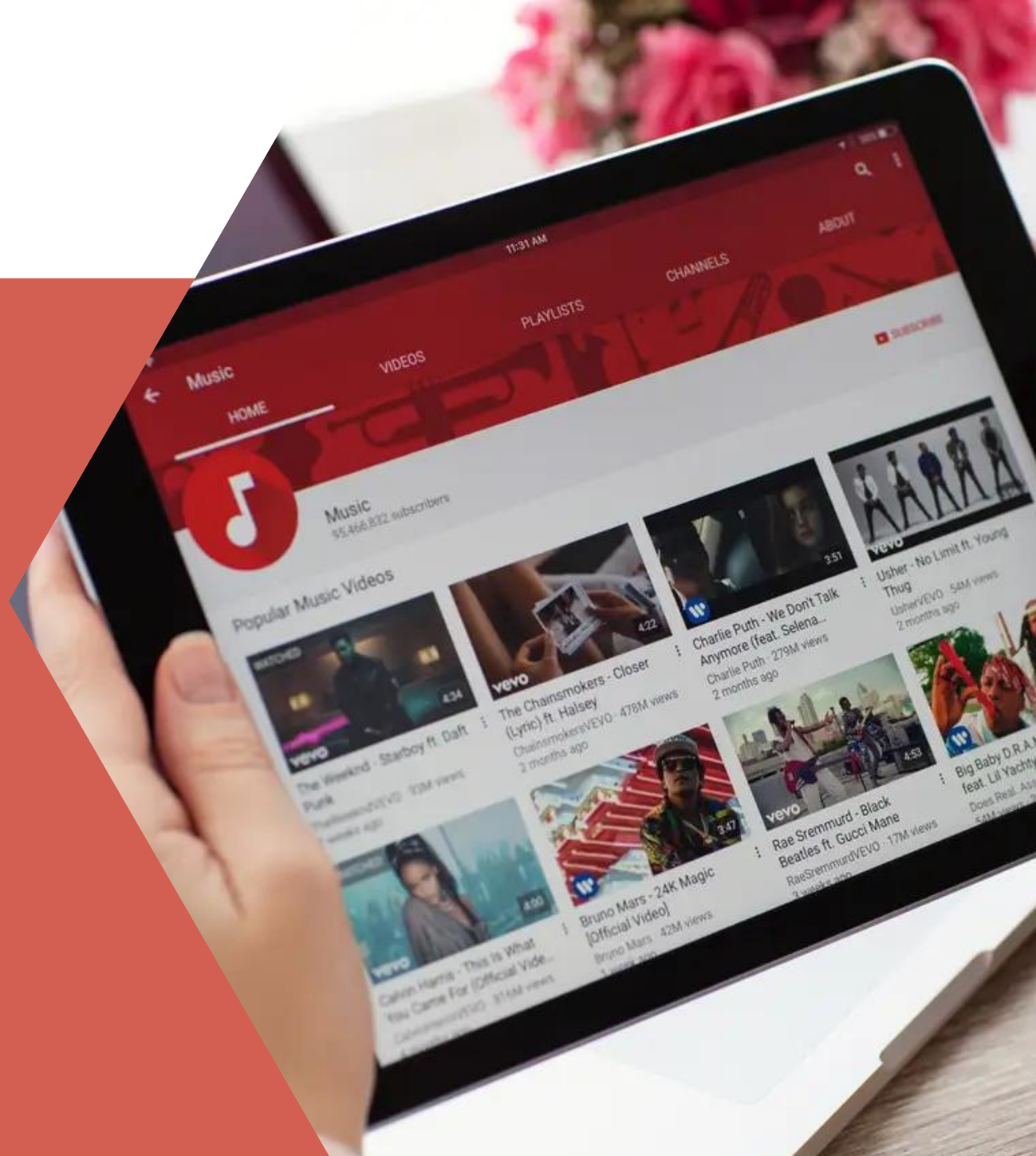
Team 7 | 工喜發財 Gong-Si-Fa-Tsai
Members | 彭家祐 任翊瑄

DEVELOPING YOUTUBE REVIEW CRAWLING & ANALYSIS API



Outline

Developing
YouTube Review
Crawling & Analysis API

01 Motivation

02 API Struture

03 YouTube Review Crawler

04 Text Mining

05 Conclusion and Future Work

1. Motivation

Users

Hope to know more opinions about the video.

YouTubers

Hope to know key words of viewers' comments.

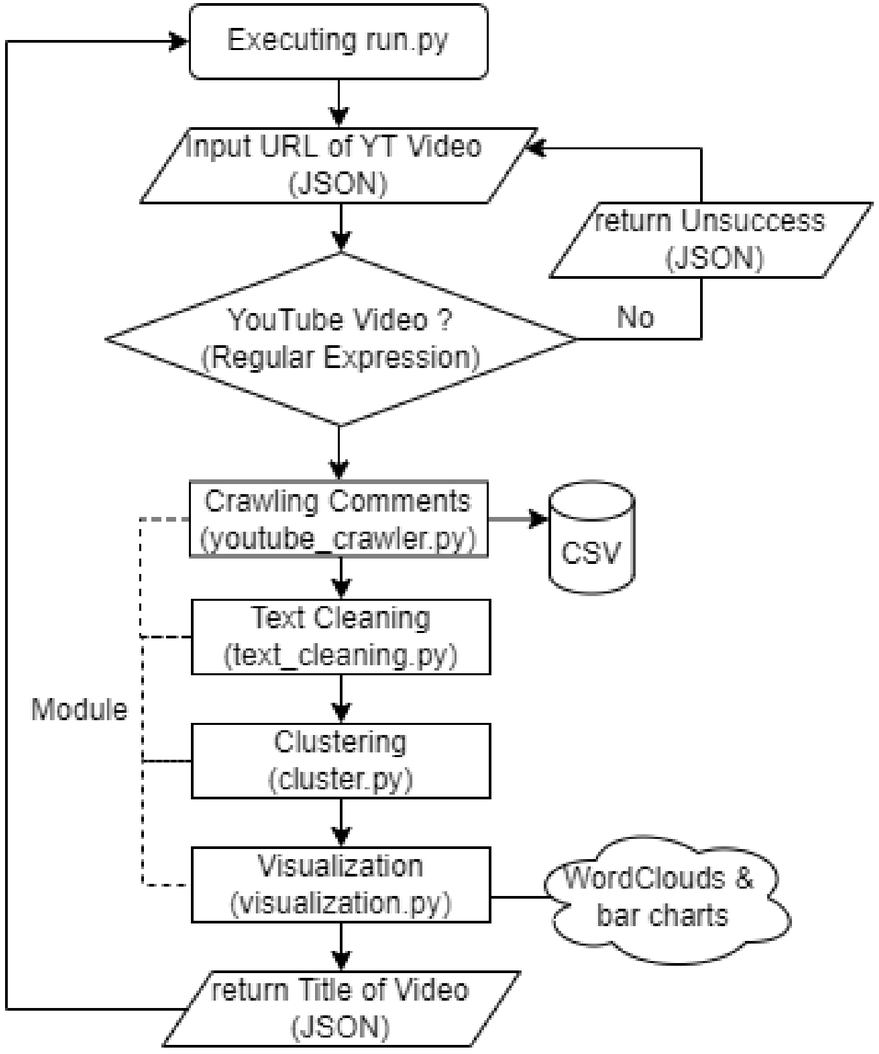
Biggest video platform

2 billions users

2. API Structure

Developing
YouTube Review
Crawling & Analysis API

API Structure



3. YouTube Review Crawler

Developing
YouTube Review
Crawling & Analysis API

YouTube Review Crawler

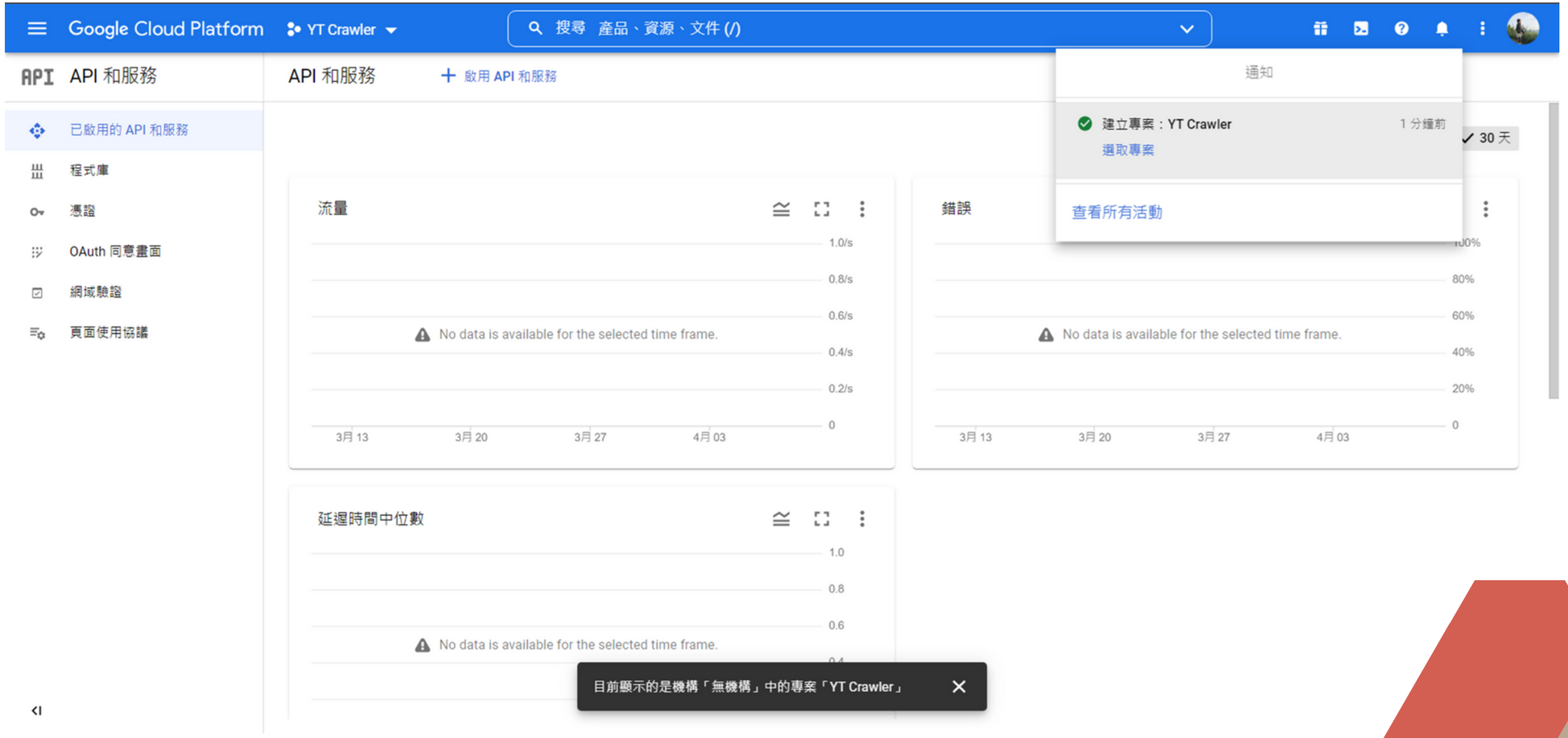
Google Cloud Platform : YouTube Data API Key

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

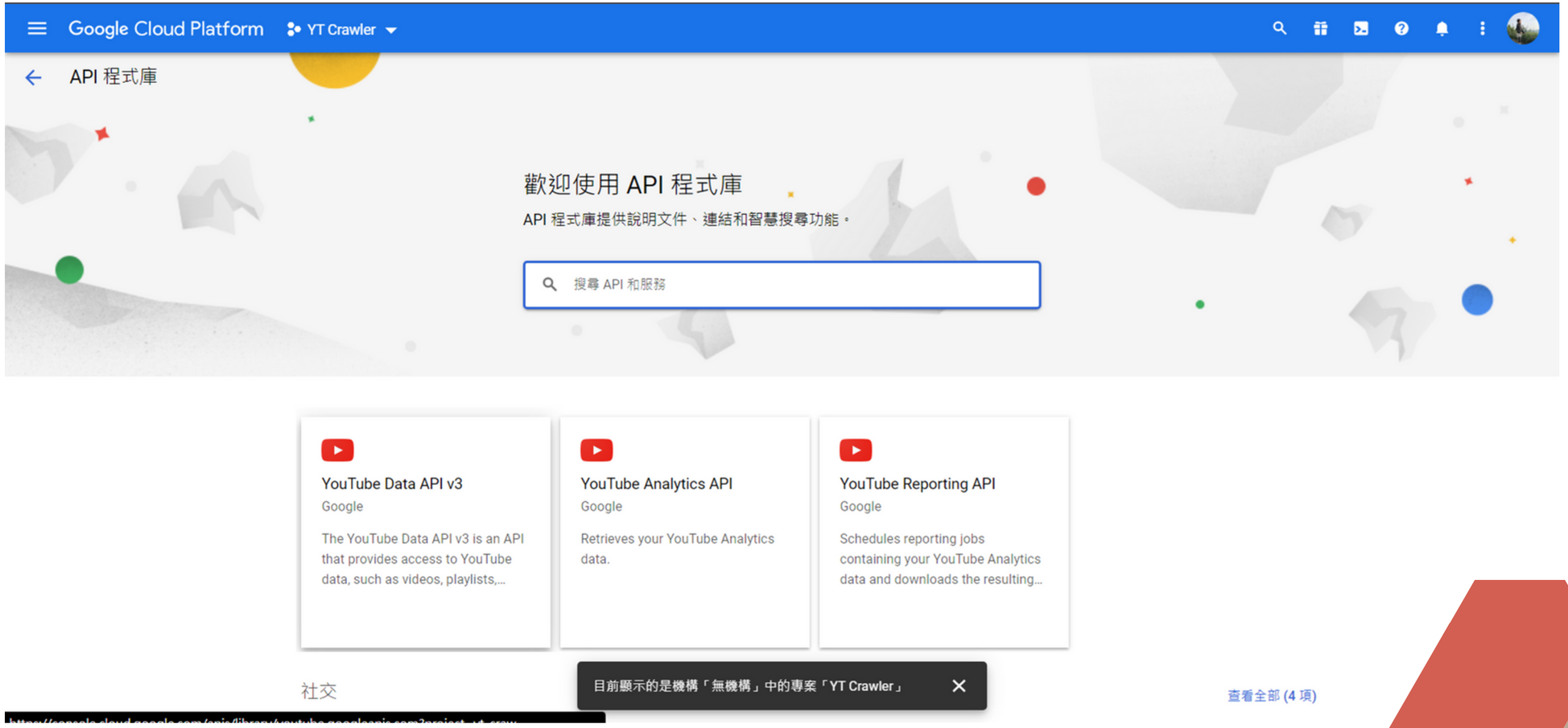
YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key



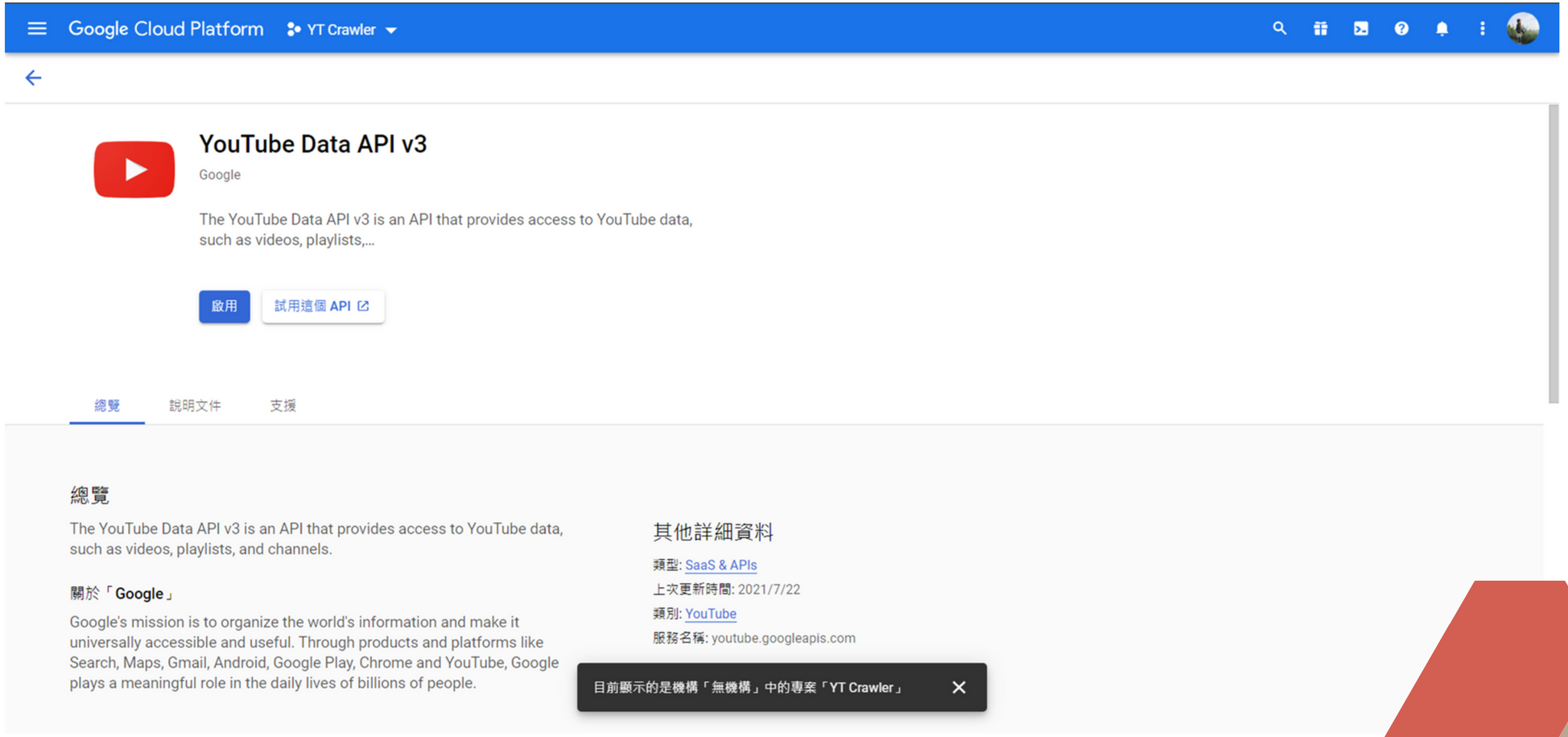
YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key



YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key



YouTube Review Crawler

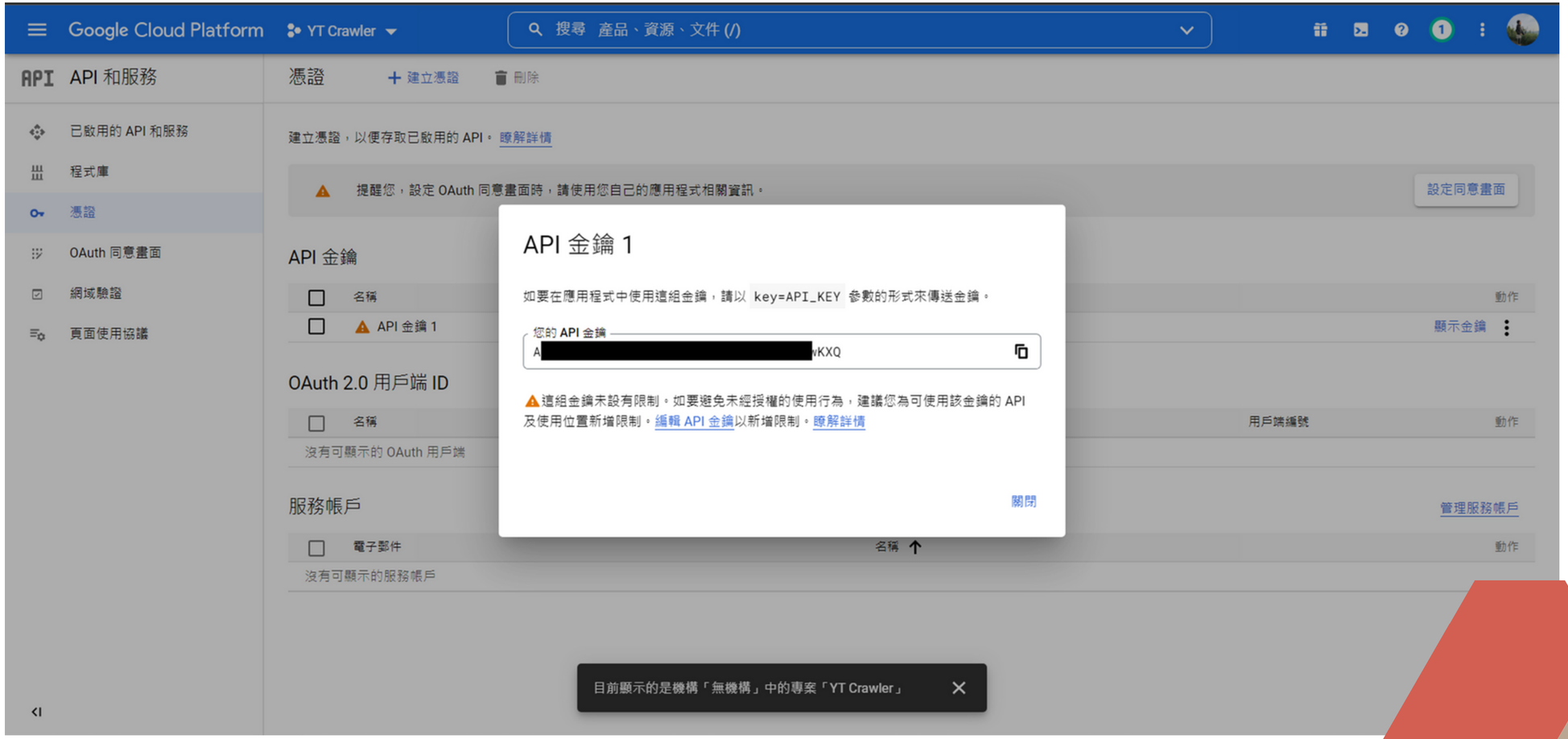
Google Cloud Platform : YouTube Data API Key

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

YouTube Review Crawler

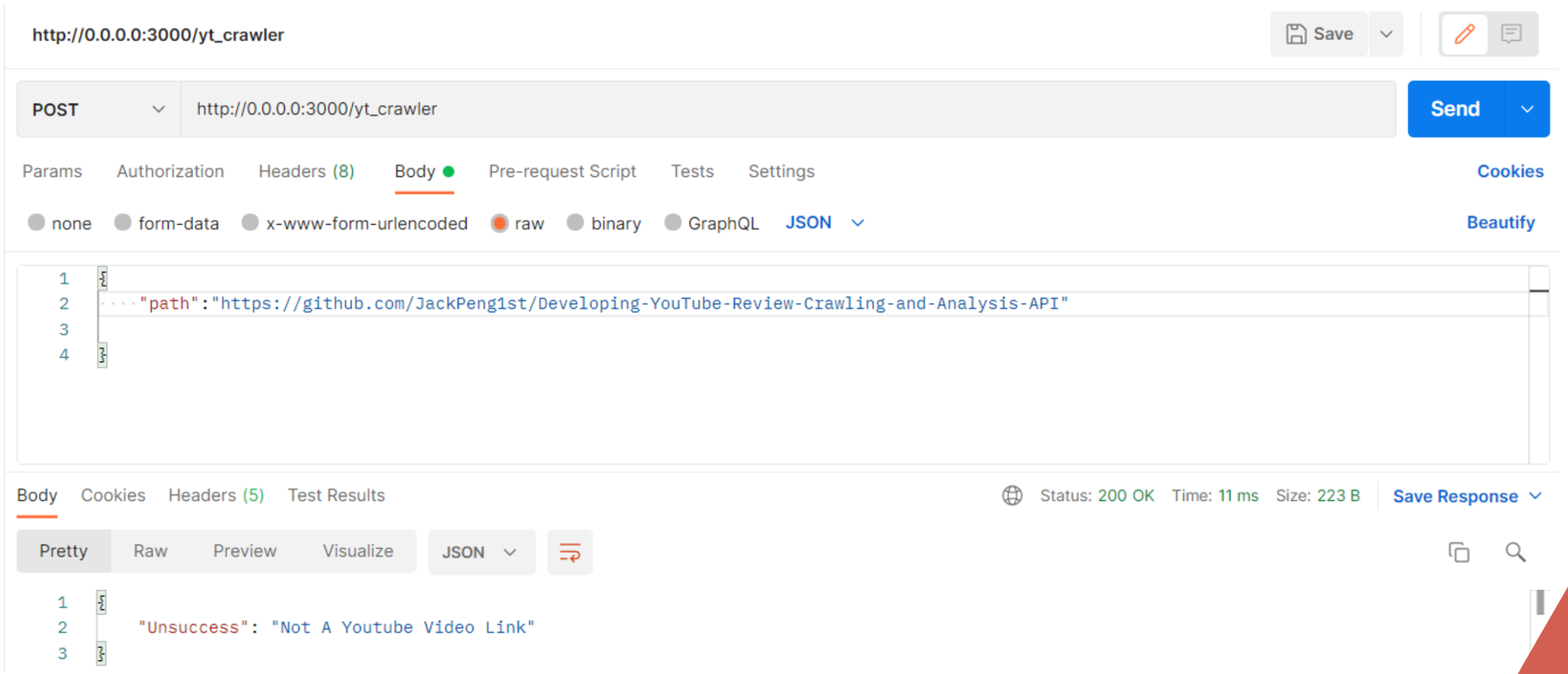
Google Cloud Platform : YouTube Data API Key



YouTube Review Crawler

YouTube Link Structure : https://www.youtube.com/watch?v=kbrTRDmwWs4&t=39s

Regular Expression : ^https\:\/\/www\.youtube\.com\/watch\?v\=.+$



YouTube Review Crawler

Video Information


```
{'id': 'kbrTRDmwWs4',
 'channelTitle': '愛莉莎莎 Alisasa',
 'publishedAt': datetime.datetime(2021, 2, 14, 15, 24, 39),
 'video_url': 'https://www.youtube.com/watch?v=kbrTRDmwWs4',
 'title': '對不起 | 愛莉莎莎Alisasa',
 'description': '嗨大家好  我是愛莉莎莎\n在這支影片上架之前，我有寄信和蒼藍鴿道歉\n也同步想在這邊澄清「肝膽排石法僅為個人體驗分享，並未有商業合作利益關係」\n\n再次跟大家說聲抱歉，也謝謝大家\n（這支影片沒有開廣告盈利）',
 'likeCount': '46780',
 'commentCount': '34161',
 'viewCount': '3064322'}
```


	name	comment	positive_num	re_comment_num
0	蒼藍鴿的醫學天地	那個...莎莎的信我已經收到了，內容也很有誠意，請大家就不要持續攻擊了。\n老實說12月底拍...	39215	488
1	Koleeeyz	有人點連結進來也以為是火鍋嗎 睡醒一堆讚是殺小	14110	403
2	Austin Yen	大家一定是：\n1.在床上\n2.沒開全螢幕\n3.滑留言	11294	330
3	愛莉莎莎 Alisasa	嗨大家好 我是愛莉莎莎\n在這支影片上架之前，我有寄信和蒼藍鴿道歉\n也同步想在這邊澄清「肝...	4932	474
4	黃郁華	希望你事後也有發現蒼藍鴿提到的那句「相信愛莉莎莎也是受害者」以及補充的「只是拿到一本不是那麼...	4616	151
...	...	...	...	...
14042	泰山寶	統神呢？ 只能按不喜歡了	0	0
14043	Liuhao	道歉就道歉還要偷誇自己是有影響力的網紅，今天錯是因為你沒知識還要嘴醫生關你是誰屁事，會不會太...	0	0
14044	Exe Mercany	下架不也是繼續賺廣告收益，反正整件事根本還沒有完，到現在沒有一個完整的道歉欸，超神奇的。你所...	0	
14045	CHENG HAO SONG	欠檢舉	0	
14046	神手周仙	端火鍋道歉可以嗎	0	

4. Text Mining

Developing
YouTube Review
Crawling & Analysis API

Text Mining

Preprocessing & tf-idf

1.Tokenizing words with NLTK.

2. Stemming and lemmatizing words to reduce inflectional or derivationally related forms of a word to a common base form.

3.Creating a custom stopwords.txt and removing the stopwords from the tokens list. (Available in English and Chinese respectively.)

4.Creating bag words and calculating TF-IDF with cosine normalization, then converting it to dataframe.

	方法	誠意	持續	引來	浪費	收到	真實	那部	攻擊	內容	...
doc0	0.107783	0.096401	0.153332	0.192405	0.155104	0.166577	0.161337	0.144147	0.099811	0.106716	...
doc1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
doc2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
doc3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
doc4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
...	...	...	...	...	...	...	...	...	...	...	...
doc14042	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
doc14043	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
doc14044	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
doc14045	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
doc14046	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...

14047 rows × 25503 columns

Text Mining

Selecting Words

Avoiding Sparseness by selecting words from positive and negative dictionaries

	方法	誠意	持續	引來
doc0	0.107783	0.096401	0.153332	0.192405
doc1	NaN	NaN	NaN	NaN
doc2	NaN	NaN	NaN	NaN
doc3	NaN	NaN	NaN	NaN
doc4	NaN	NaN	NaN	NaN
...	...	...	...	...
doc14042	NaN	NaN	NaN	NaN
doc14043	NaN	NaN	NaN	NaN
doc14044	NaN	NaN	NaN	NaN
doc14045	NaN	NaN	NaN	NaN
doc14046	NaN	NaN	NaN	NaN

14047 rows × 25503 columns

Dictionary	# of words
positive for chinese	2812
negative for chinese	8276
positive for English	2005
negative for English	4782

Total :
637 positive words
1204 negative words

top 10 of each

	知道	很棒	希望	喜歡
doc0	NaN	NaN	NaN	NaN
doc1	NaN	NaN	NaN	NaN
doc2	NaN	NaN	NaN	NaN
doc3	NaN	NaN	NaN	NaN
doc4	NaN	NaN	0.142134	NaN
...	...	...	...	...
doc14042	NaN	NaN	NaN	0.503252
doc14043	NaN	NaN	NaN	NaN
doc14044	NaN	NaN	NaN	NaN
doc14045	NaN	NaN	NaN	NaN
doc14046	NaN	NaN	NaN	NaN

14047 rows × 20 columns

Text Mining

Clustering

Unsupervised Learning

- Due to reviews of YouTube do not have the labels to identify which are positive or negative.

DBSCAN (Density-based spatial clustering of applications with noise)

- density-based clustering non-parametric algorithm
- Does not require to specify the number of clusters

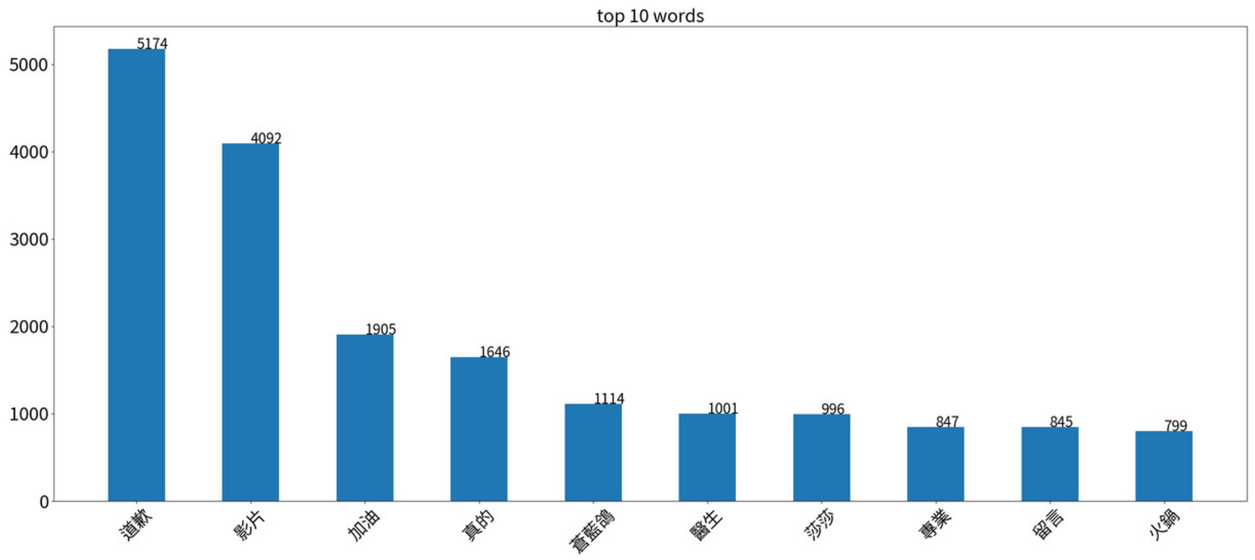
Kmeans

- partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid)
- Need to specify the number of clusters

Text Mining

Visualizing

Bar Chart
(top 10 words)



Word Cloud
(No clustering)

Text Mining Visualizing

DBSCAN

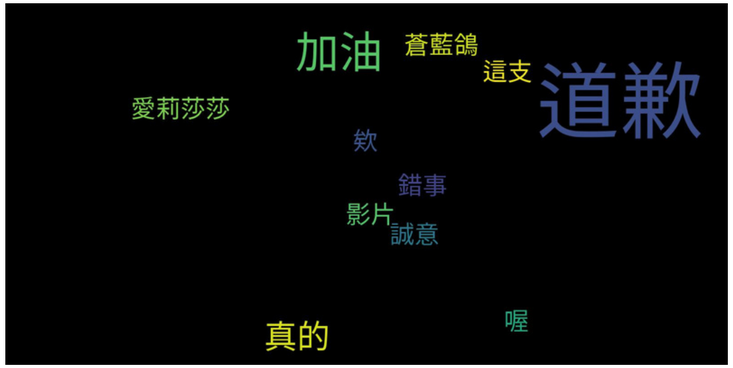
cluster 0 : # 63

cluster 1 : # 50

cluster 2 : # 13934

Kmeans

cluster 0 : # 72



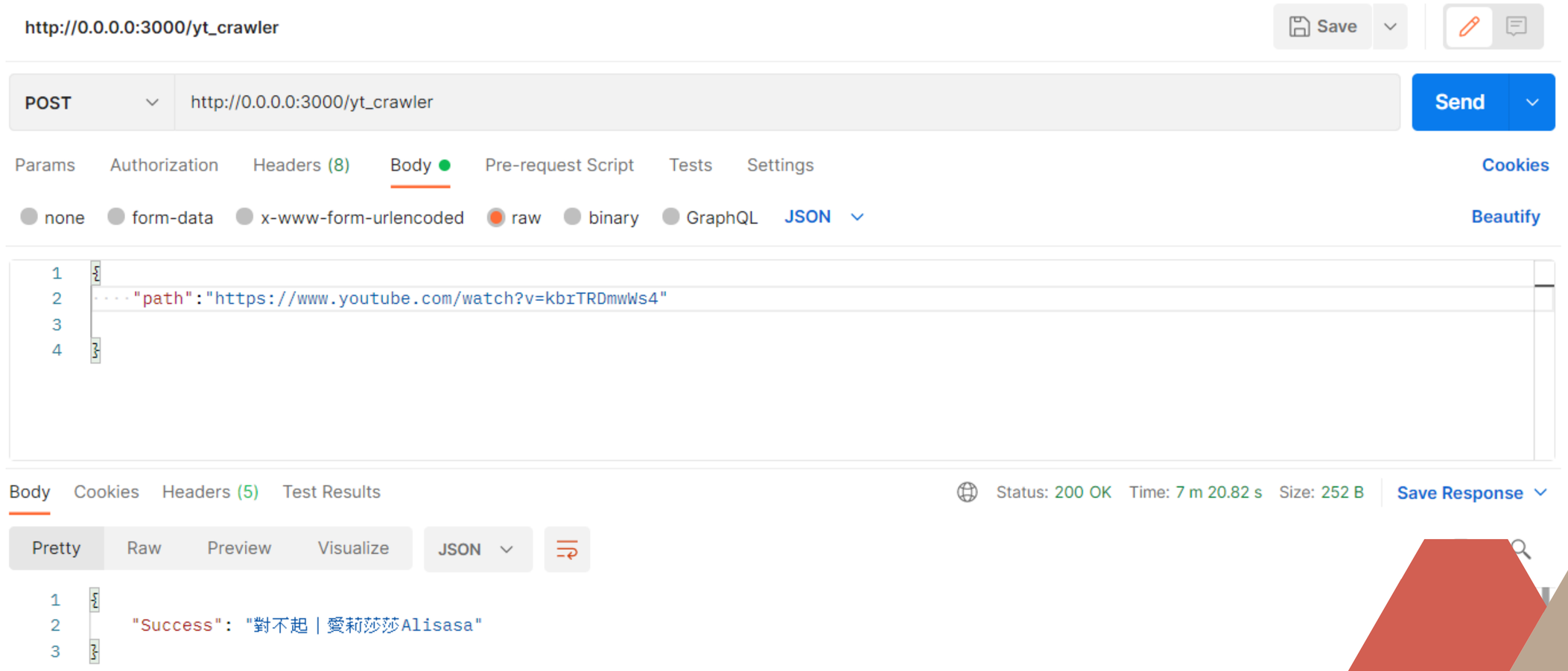
cluster 1 : # 117

cluster 2 : # 13858

API testing

Postman

JSON input & output : common data transmission format



5. Conclusion and Future Work

Developing
YouTube Review
Crawling & Analysis API

Conclusion

Using YouTube API Key, text Mining skill and development of YouTube Review Crawling & Analysis API, output the bar chart and word Cloud for Youtubers and Users

Users

Know more tags about the video, so that they will select the videos they want to watch more precisely. Besides, knowing the mainstream thoughts of others.

YouTubers

Improve the quality of the video and meet the thoughts of the viewers more

Future Work

Our objectives

Cluster

Optimizing the model to improve the differences between each clustering groups more significantly

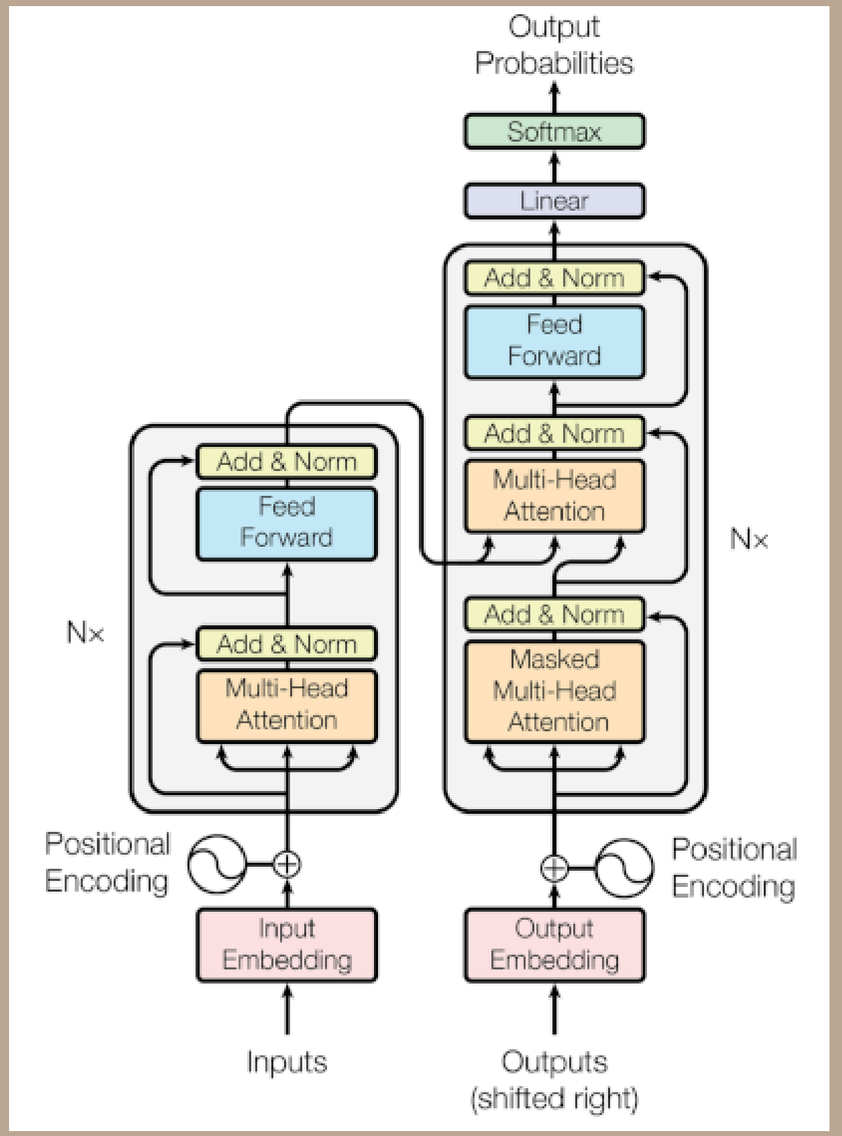
Sentiments

Building Sentiment Analysis model to identify comment is positive or not

Future Work

Tentative techniques

- Bert-based Sentiment Analysis Model
- Model Deployment on Web Pages
  (e.g. Heroku or github.io)



Thank you!

Feel free to approach us if you have any questions.