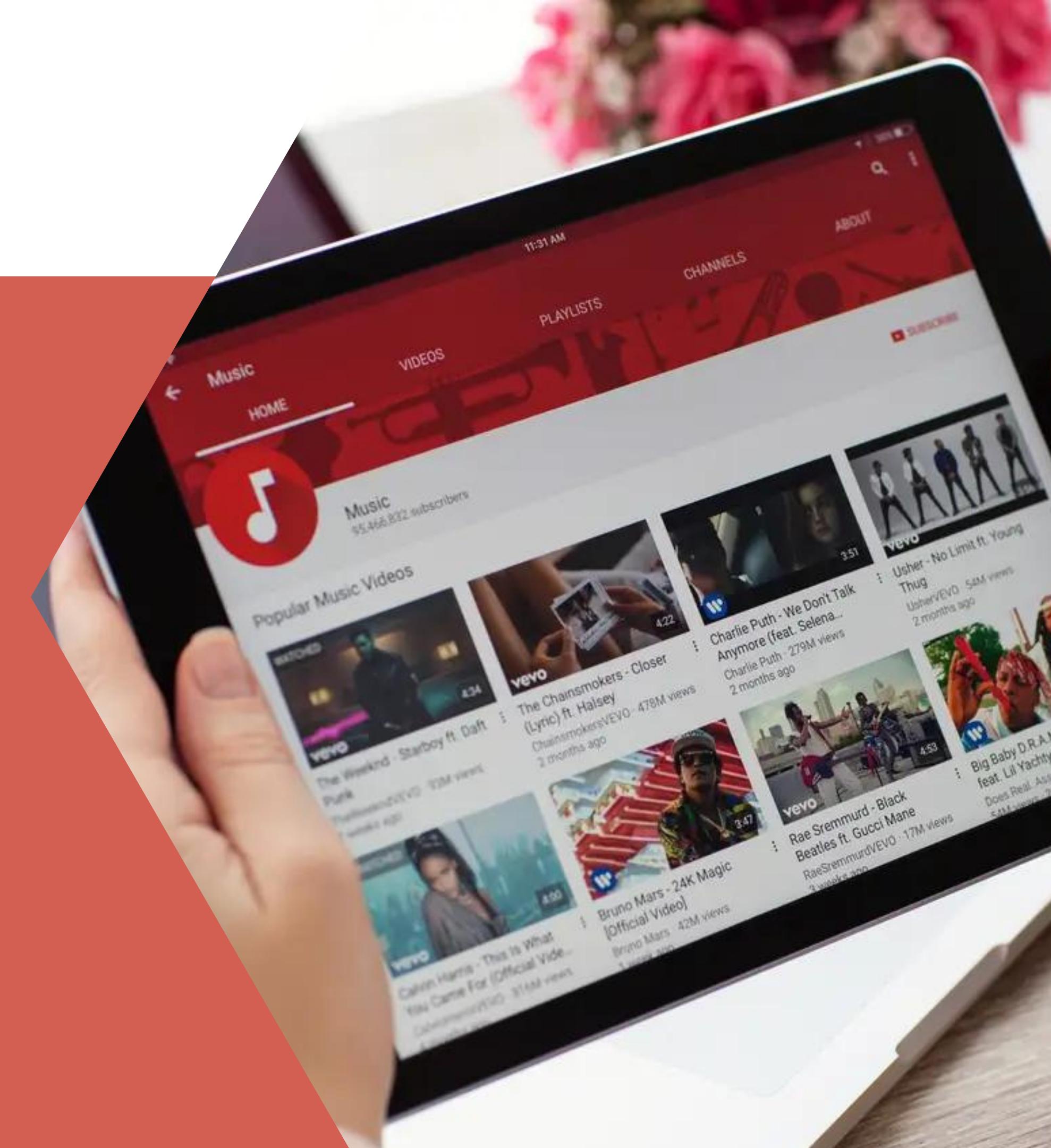




Team 7 | 工喜發財 Gong-Si-Fa-Tsai
Members | 彭家祐 任翊瑄

DEVELOPING YOUTUBE REVIEW CRAWLING & ANALYSIS API



Outline

Developing
YouTube Review
Crawling & Analysis API

01

Motivation

02

API Structure

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

01

Motivation

02

API Structure

03

YouTube Review Crawler

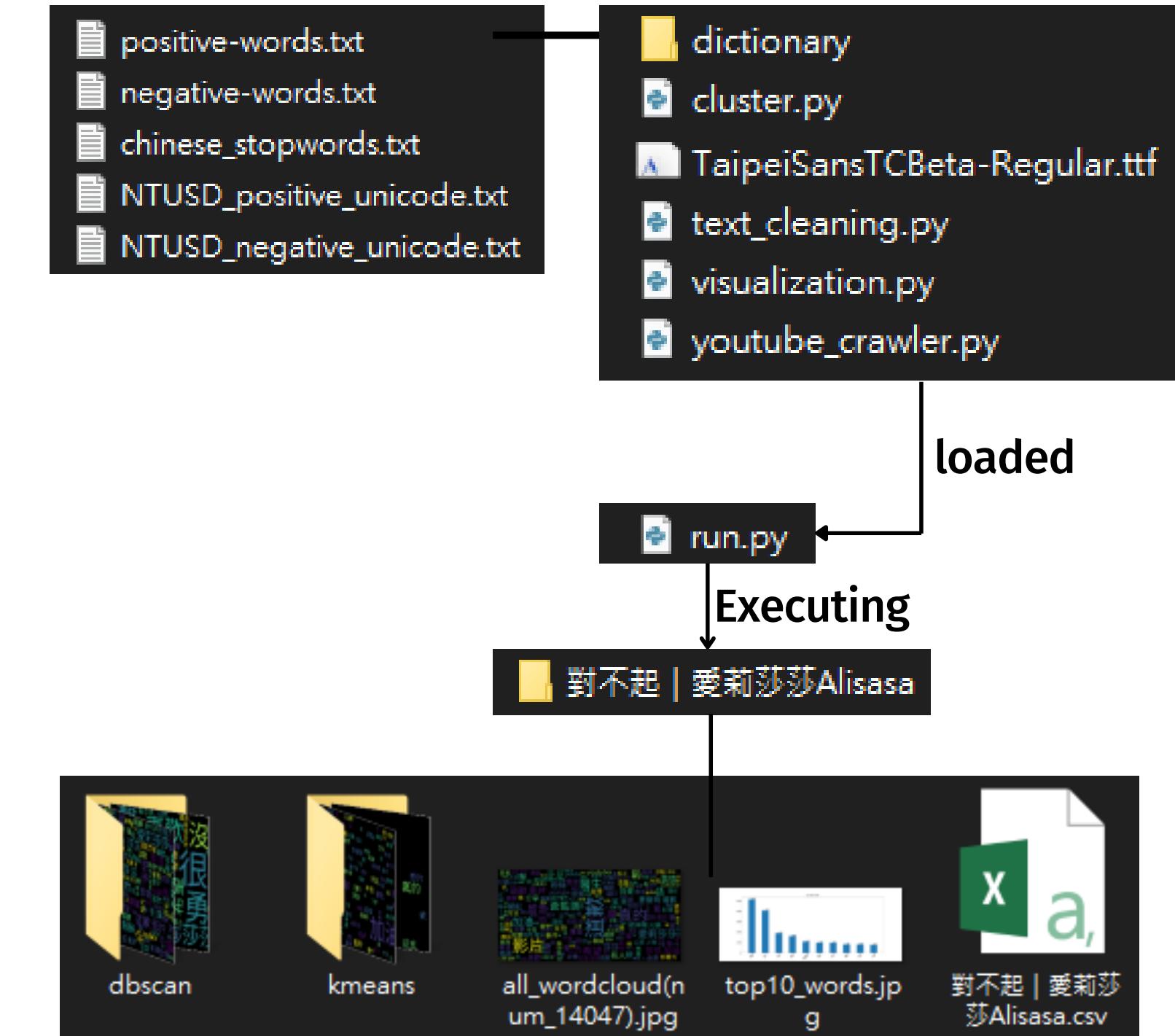
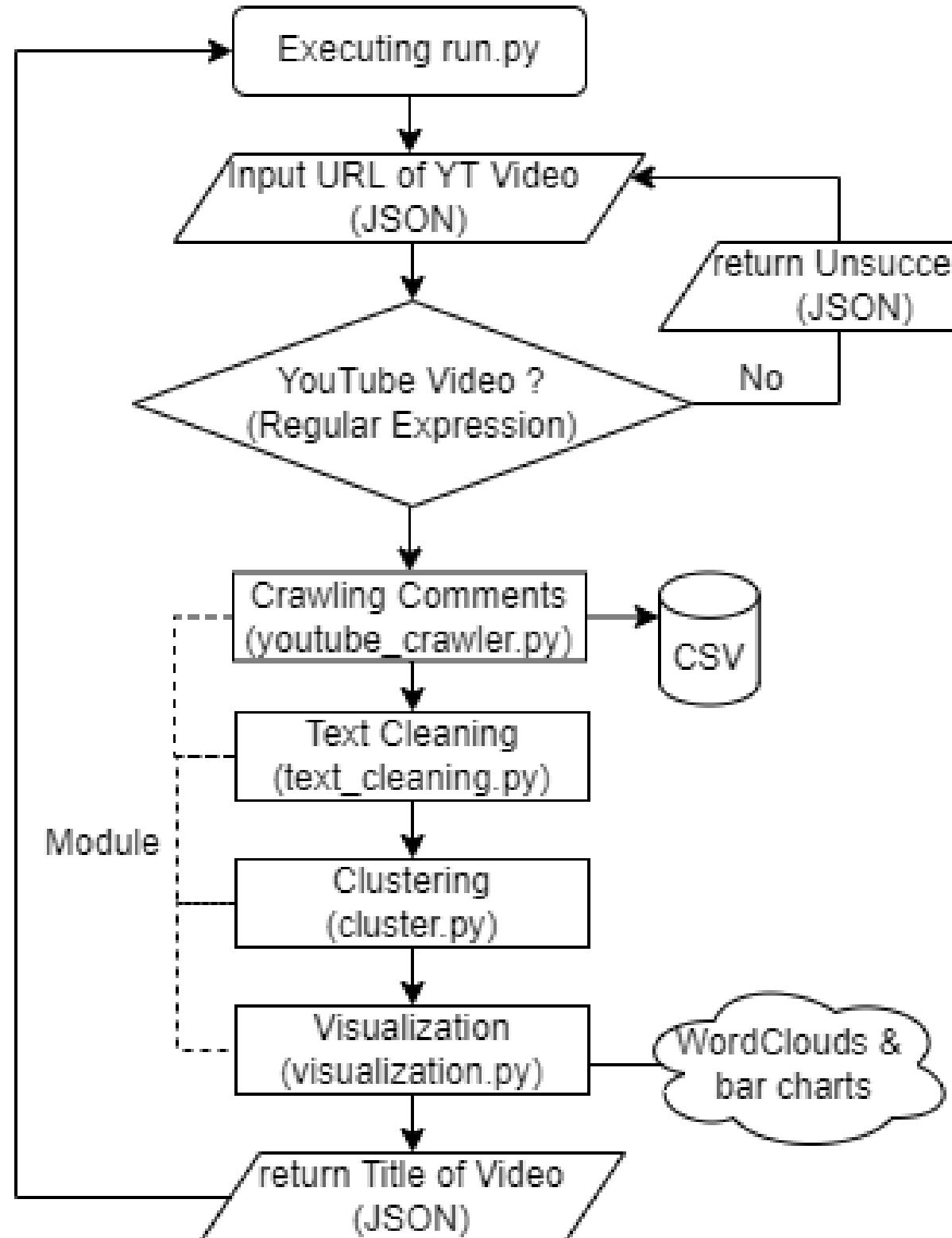
04

Text Mining

05

Conclusion and Future Work

API Structure



3. YouTube Review Crawler

Developing
YouTube Review
Crawling & Analysis API

01

Motivation

02

API Structure

03

YouTube Review Crawler

04

Text Mining

05

Conclusion and Future Work

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

The screenshot shows the Google Cloud Platform interface for managing APIs. The top navigation bar includes the Google Cloud logo, a search bar, and various navigation icons. The main content area is titled "API 和服務" (API Services) and "已啟用的 API 和服務" (Enabled APIs and Services). A sidebar on the left lists several options under "已啟用的 API 和服務" (Enabled APIs and Services), including "程式庫" (Libraries), "憑證" (Certificates), "OAuth 同意畫面" (OAuth Consent Screen), "網域驗證" (Domain Verification), and "頁面使用協議" (Page Use Agreement). A message in the center states "如要查看這個頁面，請選取專案。" (To view this page, please select a project.) with a "建立專案" (Create Project) link.

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

The screenshot shows the Google Cloud Platform web interface. At the top, there's a blue header bar with the "Google Cloud Platform" logo, a search bar containing "搜尋 產品、資源、文件 (/)", and various navigation icons on the right. Below the header, a large orange arrow-shaped graphic points to the right.

The main content area displays a form for creating a new project:

- Alert:** A yellow warning icon with the text: "您的配額還可供建立 12 projects。建議您要求增加配額或刪除專案。" followed by a "瞭解" link and a "詳情" link.
- Project Name:** A text input field labeled "專案名稱 *". It contains the value "YT Crawler" and has a question mark icon in the top right corner.
- Project ID:** A text input field labeled "專案 ID : yt-crawler-346706". It includes the note "專案 ID 設定完成後即無法變更。" and a "編輯" link.
- Location:** A dropdown menu labeled "位置 *". It shows "無機構" selected and a "瀏覽" button.
- Buttons:** At the bottom left are two buttons: a blue "建立" button and a white "取消" button.

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

The screenshot shows the Google Cloud Platform API Services page for the project 'YT Crawler'. The left sidebar is titled 'API 和服務' and has a sub-section '已啟用的 API 和服務' which is selected. It lists several services: 程式庫, 憑證, OAuth 同意畫面, 網域驗證, and 頁面使用協議. The main area displays two charts: '流量' and '延遲時間中位數', both of which show 'No data is available for the selected time frame.' A notification bubble is visible in the top right corner, indicating a new notification about the project creation.

Google Cloud Platform YT Crawler

API 和服務

已啟用的 API 和服務

程式庫

憑證

OAuth 同意畫面

網域驗證

頁面使用協議

流量

No data is available for the selected time frame.

延遲時間中位數

No data is available for the selected time frame.

通知

建立專案 : YT Crawler 1 分鐘前 30 天

選取專案

查看所有活動

錯誤

No data is available for the selected time frame.

目前顯示的是機構「無機構」中的專案「YT Crawler」

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

The screenshot shows the Google Cloud Platform API Library interface. The top navigation bar includes the 'Google Cloud Platform' logo, a dropdown for 'YT Crawler', and various icons for search, refresh, help, and notifications. Below the navigation is a left sidebar with a back arrow and the text 'API 程式庫'. The main content area features a decorative background of abstract shapes and text in Chinese: '歡迎使用 API 程式庫' and 'API 程式庫提供說明文件、連結和智慧搜尋功能。'. A search bar at the bottom of this section contains the placeholder '搜尋 API 和服務'. Below this are three service cards:

- YouTube Data API v3** (Google) - Describes it as an API for accessing YouTube data like videos and playlists.
- YouTube Analytics API** (Google) - Describes it as retrieving YouTube Analytics data.
- YouTube Reporting API** (Google) - Describes it as scheduling reporting jobs to download YouTube Analytics data.

社交

目前顯示的是機構「無機構」中的專案「YT Crawler」



查看全部 (4 項)

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

The screenshot shows the Google Cloud Platform interface with a blue header bar. The header includes the 'Google Cloud Platform' logo, a dropdown menu for 'YT Crawler', and various navigation icons. Below the header, there's a back arrow icon.

The main content area displays the 'YouTube Data API v3' page. It features a red YouTube play button icon and the text 'YouTube Data API v3' followed by 'Google'. A brief description states: 'The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists,...'. Below this, there are two buttons: a blue '啟用' (Enable) button and a white '試用這個 API' (Try this API) button with a blue outline.

At the bottom of the page, there are three tabs: '總覽' (Overview), '說明文件' (Documentation), and '支援' (Support). The '總覽' tab is currently selected.

In the bottom right corner of the main content area, there's a dark overlay with white text that reads '目前顯示的是機構「無機構」中的專案「YT Crawler」' and a small 'X' icon.

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

The screenshot shows the Google Cloud Platform interface for managing APIs. The left sidebar has 'API' selected under 'API 和服務'. The main area shows the '憑證' (Credentials) tab selected. A message says '建立憑證，以便存取已啟用的 API。' (Create a credential to access enabled APIs.) with a link to '瞭解詳情' (Learn more). A warning message states: '提醒您，設定 OAuth 同意畫面時，請使用您自己的應用程式相關資訊。' (Please note, when setting up the OAuth consent screen, use your own application's related information.) with a '設定同意畫面' (Set up consent screen) button. Below this are sections for 'API 金鑰' (API Keys), 'OAuth 2.0 用戶端 ID' (OAuth 2.0 Client IDs), and '服務帳戶' (Service Accounts). A message at the bottom says '目前顯示的是機構「無機構」中的專案「YT Crawler」' (The current view is for the project 'YT Crawler' in the organization 'None').

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

The screenshot shows the Google Cloud Platform interface for managing APIs and services. The left sidebar is titled 'API 和服務' (API & Services) and lists several options: '已啟用的 API 和服務' (Enabled APIs & Services), '程式庫' (Libraries), '憑證' (Credentials, highlighted in blue), 'OAuth 同意畫面' (OAuth Consent Screen), '網域驗證' (Domain Verification), and '頁面使用協議' (Page Use Agreement). The main content area is titled '憑證' (Credentials) and includes a '建立憑證, 以便透過一組簡單的 API 金鑰識別專案, 以便查看配額與存取權限' (Create a credential to identify your project using a simple API key, so you can view usage limits and access permissions) button. A dropdown menu is open over this button, showing three options: 'API 金鑰' (API Key), 'OAuth 用戶端 ID' (OAuth Client ID), and '服務帳戶' (Service Account). The 'API 金鑰' option is selected. Below the dropdown, there is a '請幫我選擇' (Help me choose) section with the note '精靈會詢問幾個問題, 協助您決定要使用何種類型的憑證' (The wizard will ask a few questions to help you decide which type of credential to use). At the bottom of the credentials section, there is a table for 'OAuth 2.0 用戶端 ID' (OAuth 2.0 Client ID) with one row: '沒有可顯示的 OAuth 用戶端' (No displayable OAuth clients). Below this is a section for '服務帳戶' (Service Accounts) with one row: '電子郵件' (Email) and '沒有可顯示的服務帳戶' (No displayable service accounts). A small black banner at the bottom center of the page says '目前顯示的是機構「無機構」中的專案『YT Crawler』' (The current display is the project 'YT Crawler' in the organization 'No organization').

YouTube Review Crawler

Google Cloud Platform : YouTube Data API Key

The screenshot shows the Google Cloud Platform interface for managing APIs. The left sidebar is titled 'API 和服務' and includes sections for '已啟用的 API 和服務', '程式庫', '憑證' (selected), 'OAuth 同意畫面', '網域驗證', and '頁面使用協議'. The main content area is titled '憑證' and has a sub-section 'API 金鑰'. A modal window titled 'API 金鑰 1' displays the API key value: 'A [REDACTED] vKXQ'. Below the modal, a note states: '這組金鑰未設有限制。如要避免未經授權的使用行為，建議您為可使用該金鑰的 API 及使用位置新增限制。編輯 API 金鑰以新增限制。' At the bottom of the modal, there is a '關閉' button. A footer message at the bottom of the page says: '目前顯示的是機構「無機構」中的專案「YT Crawler」'.

YouTube Review Crawler

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

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

The screenshot shows a POST request to `http://0.0.0.0:3000/yt_crawler`. The Body tab is selected, showing a JSON payload:

```
1 {"path": "https://github.com/JackPeng1st/Developing-YouTube-Review-Crawling-and-Analysis-API"}
```

The response status is 200 OK with a size of 223 B.

Body (Pretty) Response:

```
1 {"Unsuccess": "Not A Youtube Video Link"}
```

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，我有寄信和蒼藍鵠道歉\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	蒼藍鵠的醫學天地	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，我有寄信和蒼藍鵠道歉\\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鵠提到的那句「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	Liuhalo	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	0
14045	CHENG HAO SONG	欠檢舉	0	0
14046	神手周仙	端火鍋道歉可以嗎	0	0

4. Text Mining

Developing
YouTube Review
Crawling & Analysis API

01

Motivation

02

API Structure

03

YouTube Review Crawler

04

Text Mining

05

Conclusion and Future Work

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	...									
doc2	NaN	...									
doc3	NaN	...									
doc4	NaN	...									
...
doc14042	NaN	...									
doc14043	NaN	...									
doc14044	NaN	...									
doc14045	NaN	...									
doc14046	NaN	...									

14047 rows × 25503 columns

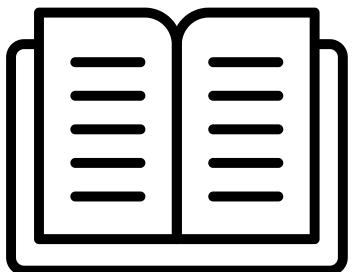
Text Mining

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



Total :
637 positive words
1204 negative words

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

	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

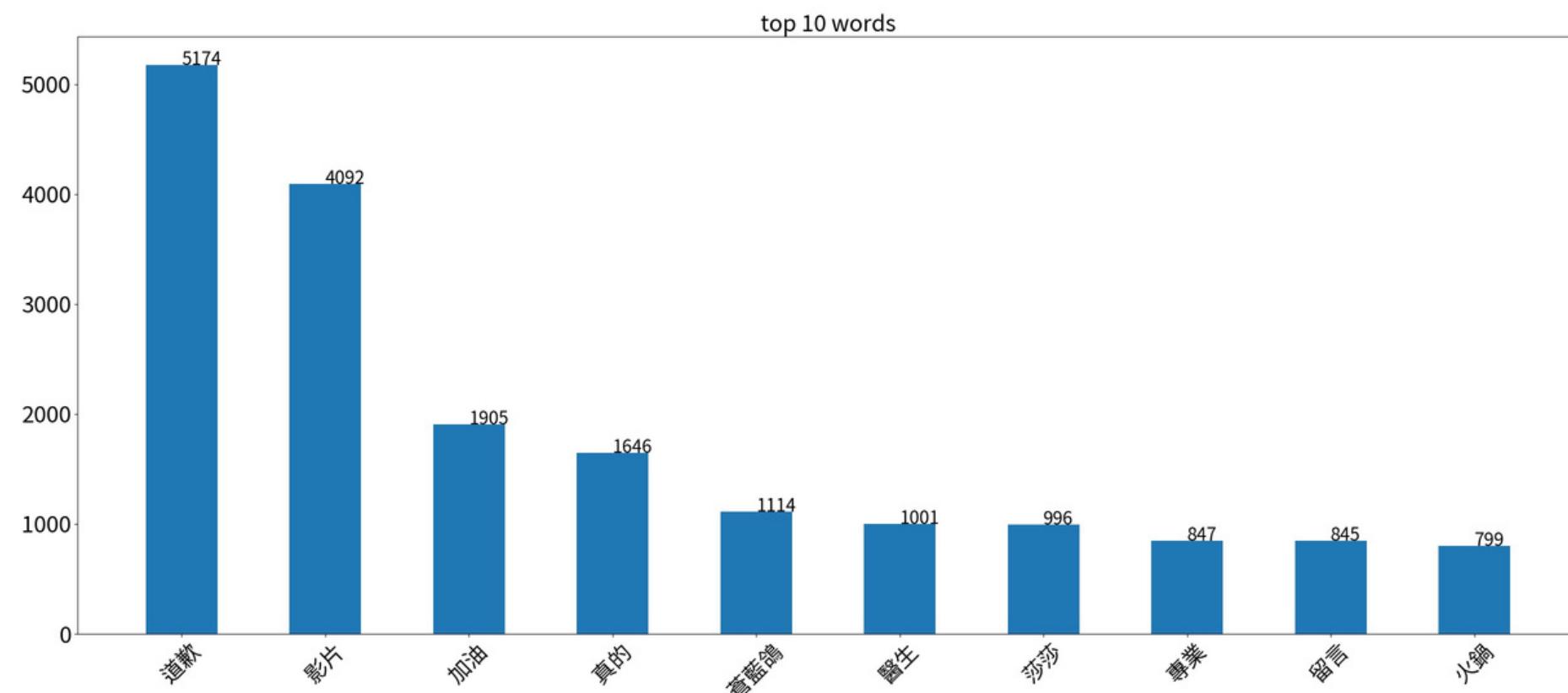
Kmeans

- 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

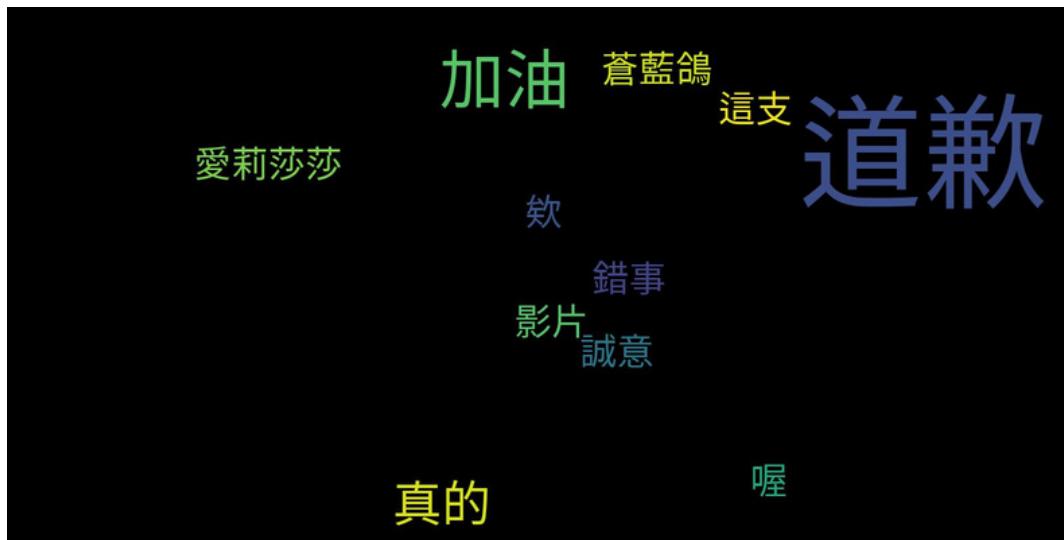


cluster 2 : # 13934



Kmeans

cluster 0 : # 72



cluster 1: # 117



cluster 2 : # 13858



API testing

Postman

JSON input & output : common data transmission format

The screenshot shows the Postman application interface. At the top, the URL `http://0.0.0.0:3000/yt_crawler` is entered. Below it, a POST request is selected, and the URL is again shown. The 'Body' tab is active, showing the JSON input:

```
1 {  
2   "path": "https://www.youtube.com/watch?v=kbrTRDmwls4"  
3 }  
4
```

The 'Body' tab also includes options for 'none', 'form-data', 'x-www-form-urlencoded', 'raw', 'binary', and 'GraphQL'. The 'raw' option is selected, and the JSON tab is highlighted. Below the input area, the response status is shown as 200 OK.

At the bottom, the response body is displayed in Pretty format:

```
1 {  
2   "Success": "對不起 | 賈莉莎莎 Alisasa"  
3 }
```

5. Conclusion and Future Work

Developing YouTube Review Crawling & Analysis API

- 01 Motivation
- 02 API Structure
- 03 YouTube Review Crawler
- 04 Text Mining
- 05 Conclusion and Future Work

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

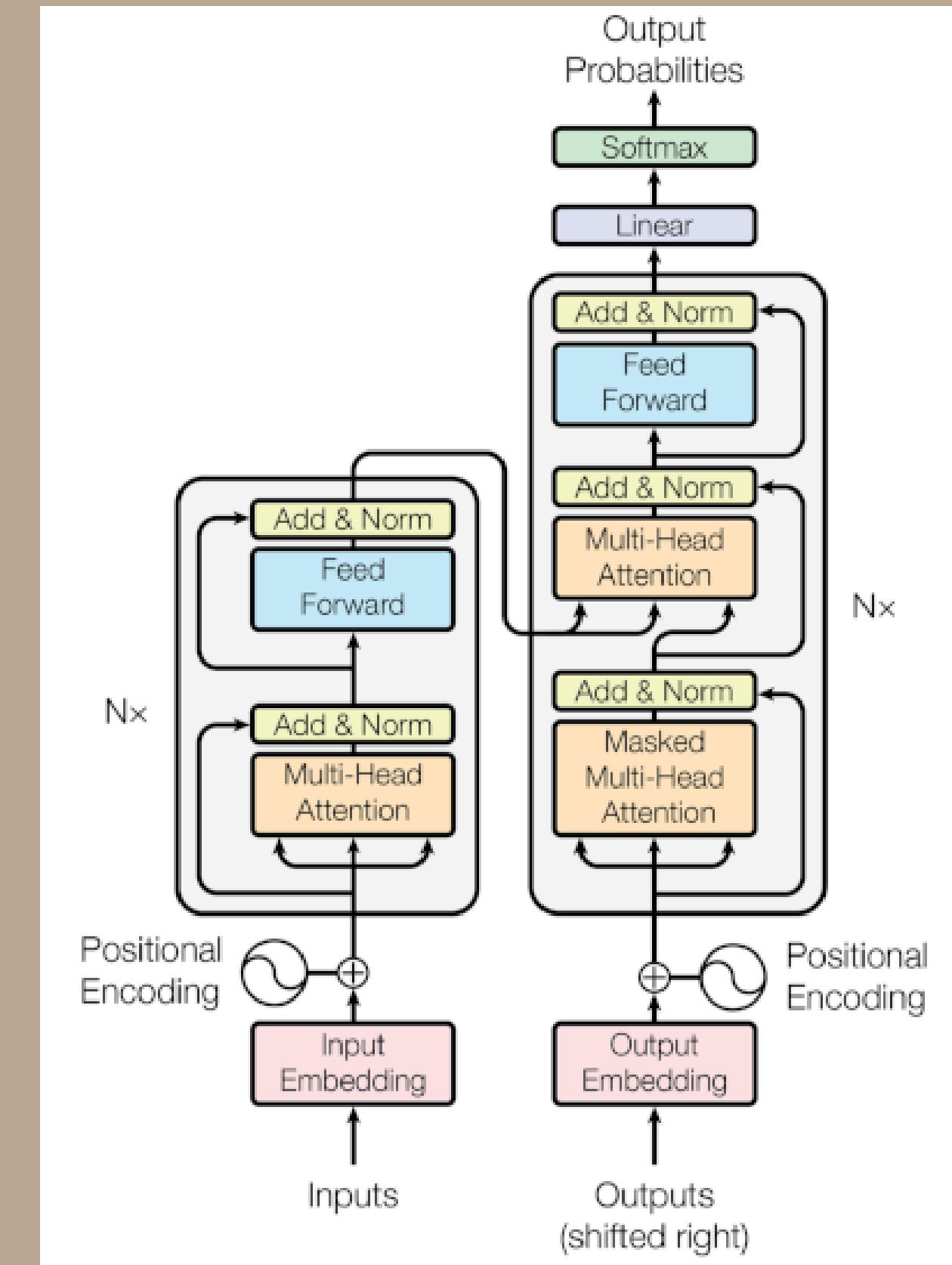
Building Sentiment Analysis model to identify comment is positive or not

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.