Final Project

Instagram Data Analysis

Prepared by Ken
2022-04-30

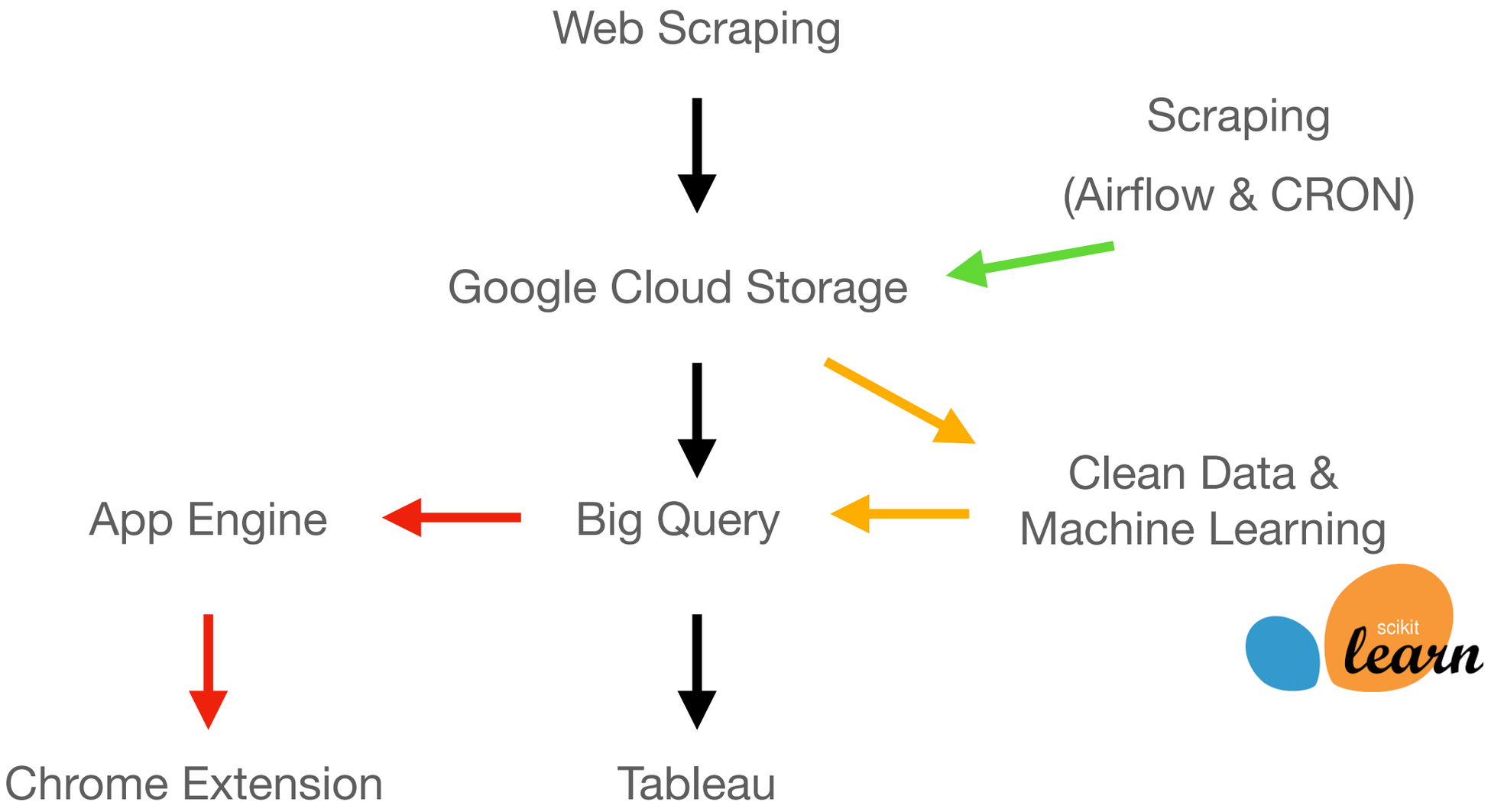
Problem

- The digital marketer difficultly to find a right IG KOL to spread their Ads.

Purpose

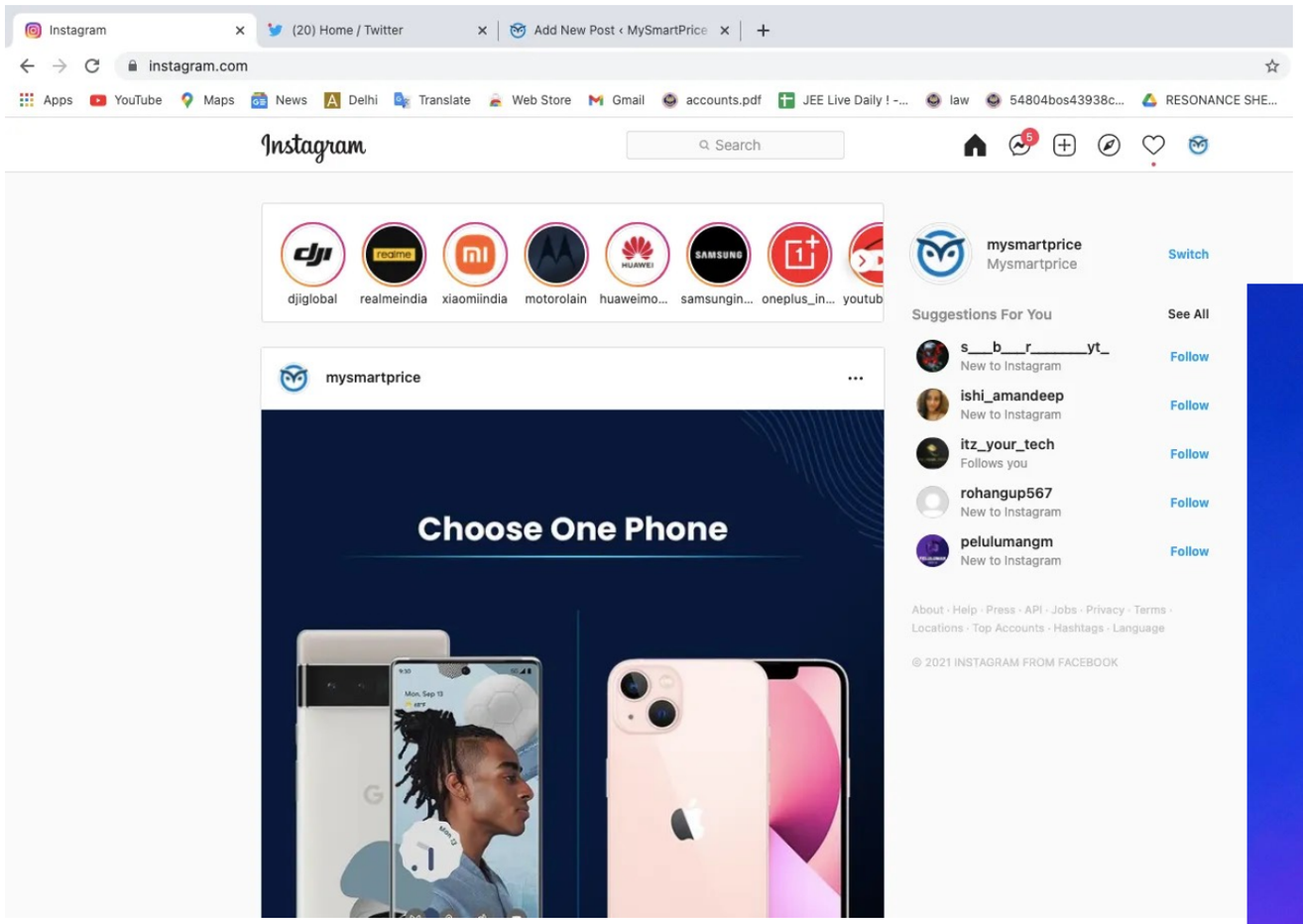
- To develop the tool to find out the KOL and it provides different features of IG account for analysis.
- This tool can continually update the IG KOL information.

Architecture (Brief)



What’s now

- First, we need data
- Data is from Instagram (App or Web)



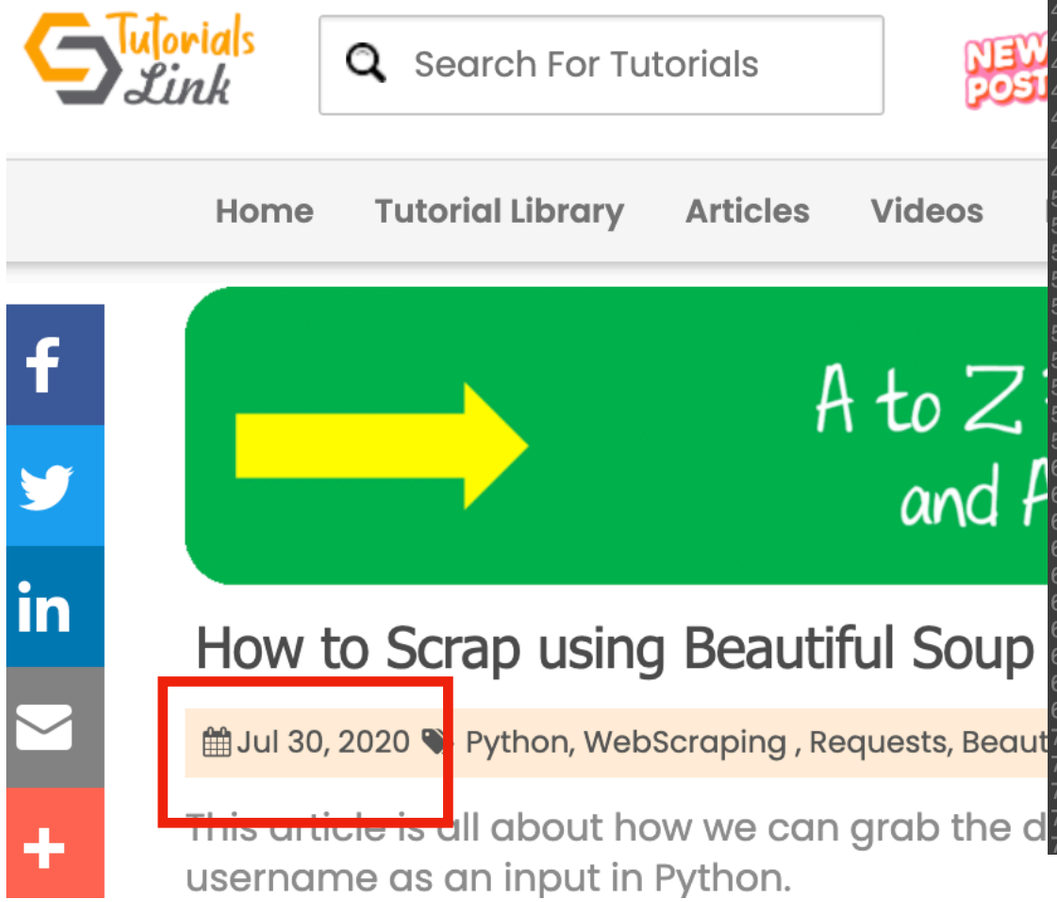
Thinking…

PART 1

WEB SCRAPING

Web Scraping

2021, request.get( ) is work



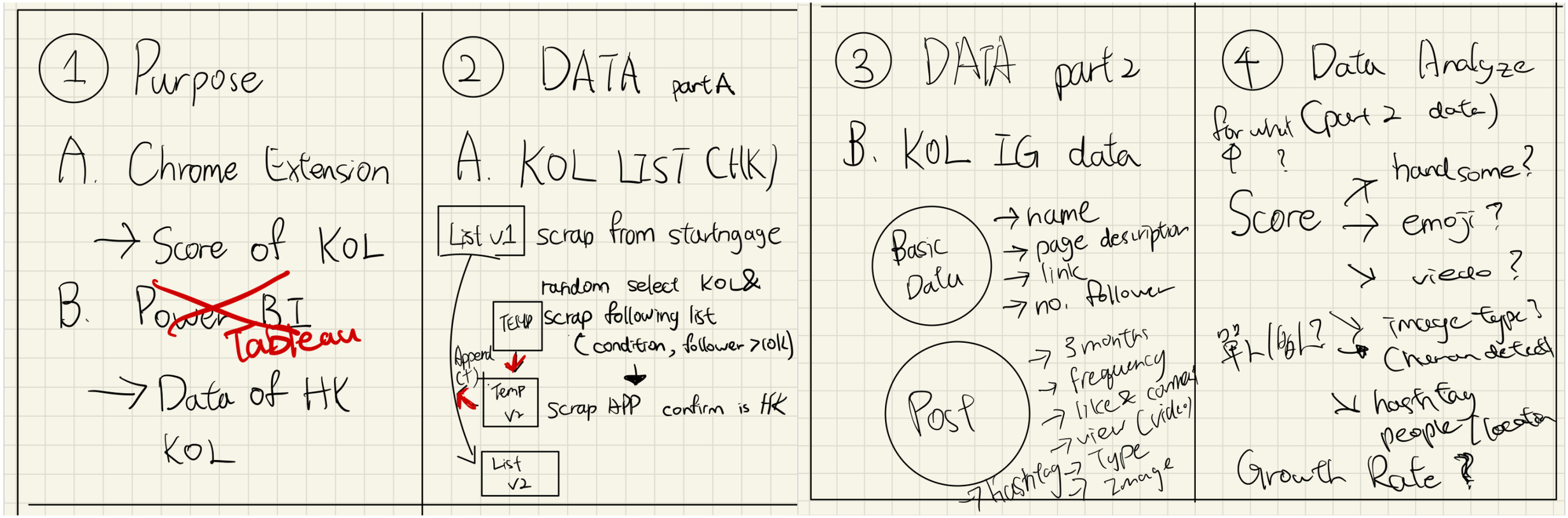
2022, request.get( ) not work anymore

```
        <script type="text/javascript">
        (function() {
  var docElement = document.documentElement;
  var classRE = new RegExp('(^|\\s)no-js(\\s|$)');
  var className = docElement.className;
  docElement.className = className.replace(classRE, '$1js$2');
})();
</script>
        <script type="text/javascript">
(function() {
  if ('PerformanceObserver' in window && 'PerformancePaintTiming' in window) {
    window.__bufferedPerformance = [];
    var ob = new PerformanceObserver(function(e) {
      window.__bufferedPerformance.push.apply(window.__bufferedPerformance,e.getEntries());
    });
    ob.observe({entryTypes:['paint']});
  }

  window.__bufferedErrors = [];
  window.onerror = function(message, url, line, column, error) {
    window.__bufferedErrors.push({
      message: message,
      url: url,
      line: line,
      column: column,
      error: error
    });
    return false;
  };
  window.__initialData = {
    pending: true,
    waiting: []
  };
  function asyncFetchSharedData(extra) {
    var sharedDataReq = new XMLHttpRequest();
    sharedDataReq.onreadystatechange = function() {
          if (sharedDataReq.readyState === 4) {
            if(sharedDataReq.status === 200){
              var sharedData = JSON.parse(sharedDataReq.responseText);
              window.__initialDataLoaded(sharedData, extra);
```

Selenium

- Planning:



What?

(List v2/v3)

(Top 100 HK IG id)

startngage.com

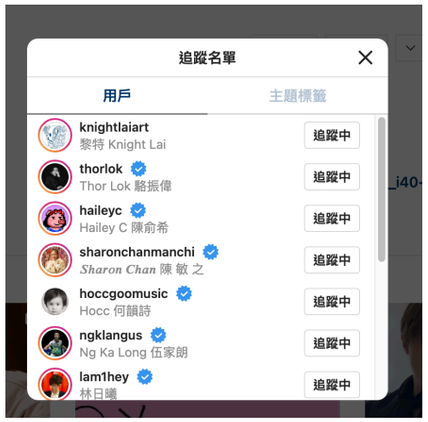
List v1

Details v1

(Condition: >1000 followers)

(Basic Information)

instagram.com



(following_list)

instagram.com

List v2

Details v2

(25 Posts in 2022)

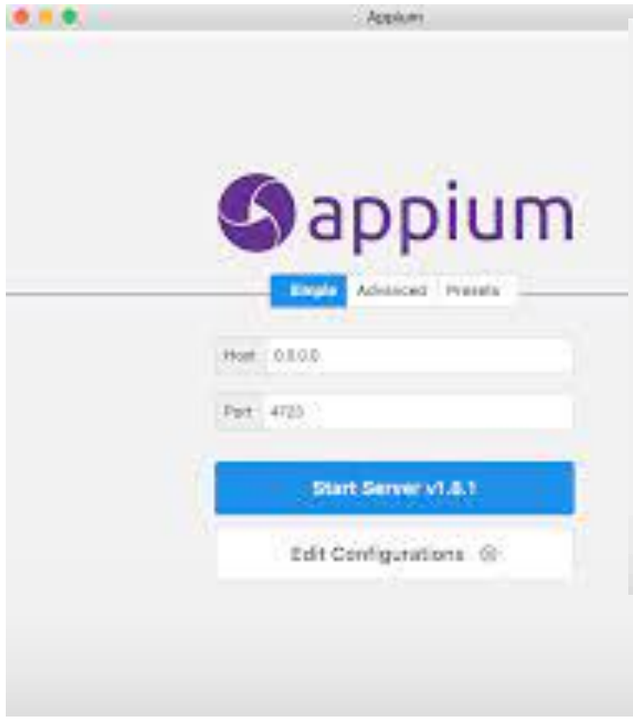
instagram.com

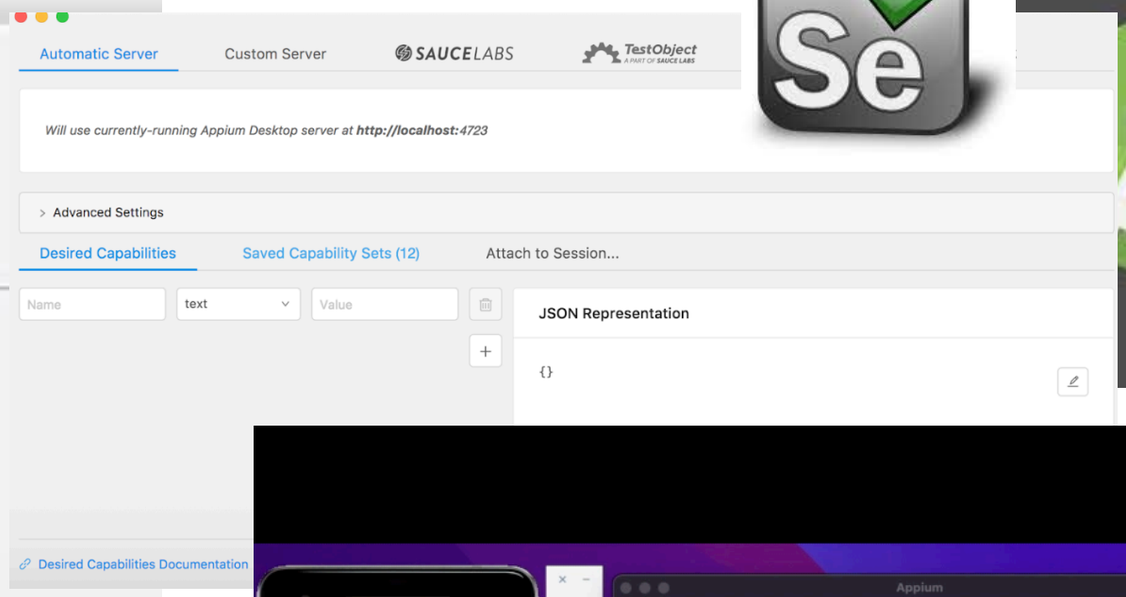
(Confirm country)

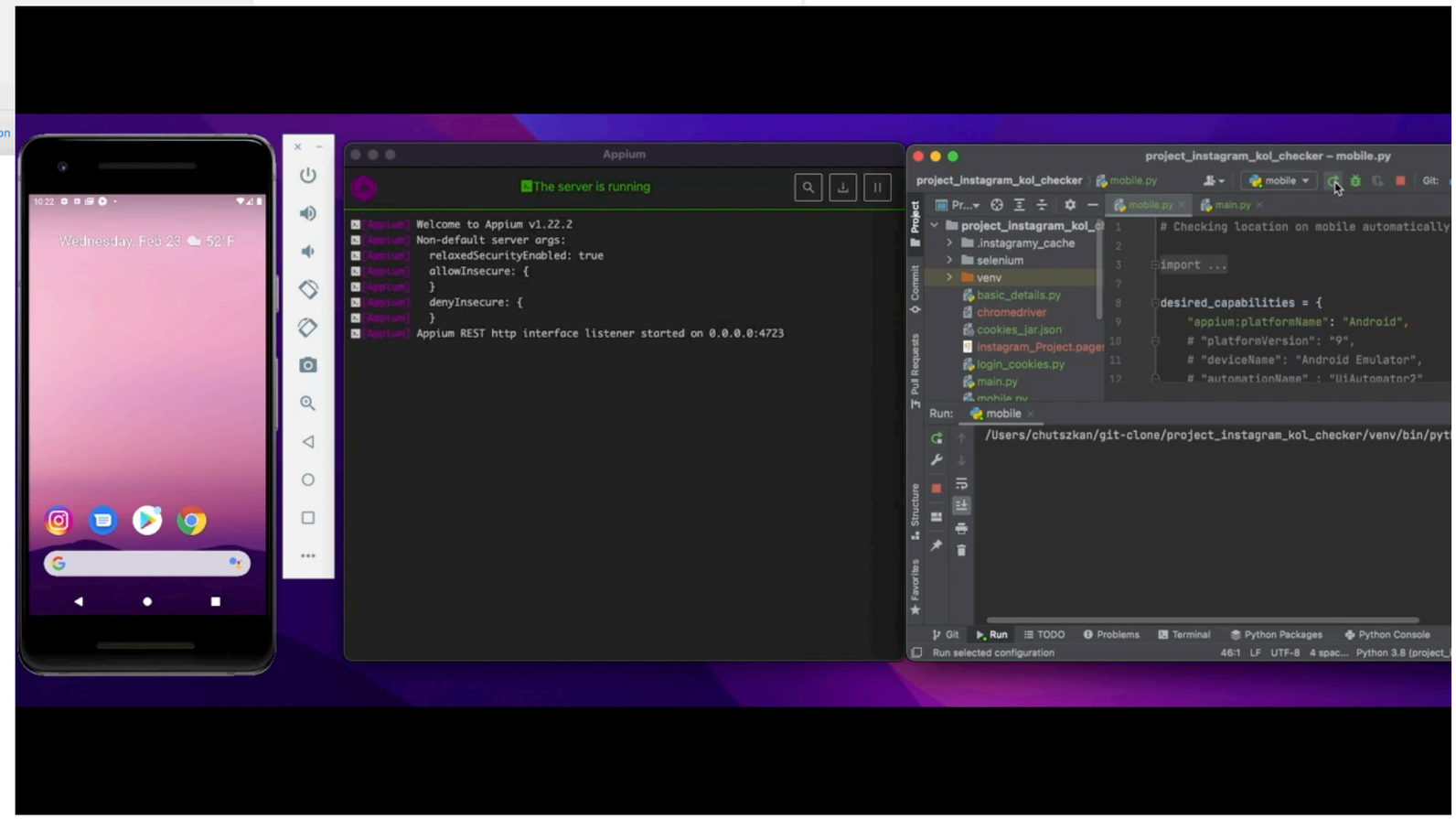
Instagram App

List v3

Mobile Scraping





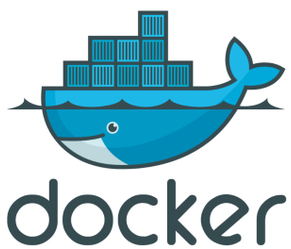


What’s data I get

Csv1	Csv2	Csv3
• ig_id	• ig_id	• ig_id
• country	• name	• post date
• updated following list time	• description	• post alt
• details updated time	• no of post	• post type
• following list	• no of follower	• like/view
	• no of following	• comment
	• profile image	• content

Secludedly scrap the data

kubernetes



Give Up at last !!

PART 2

DATA CLEANING

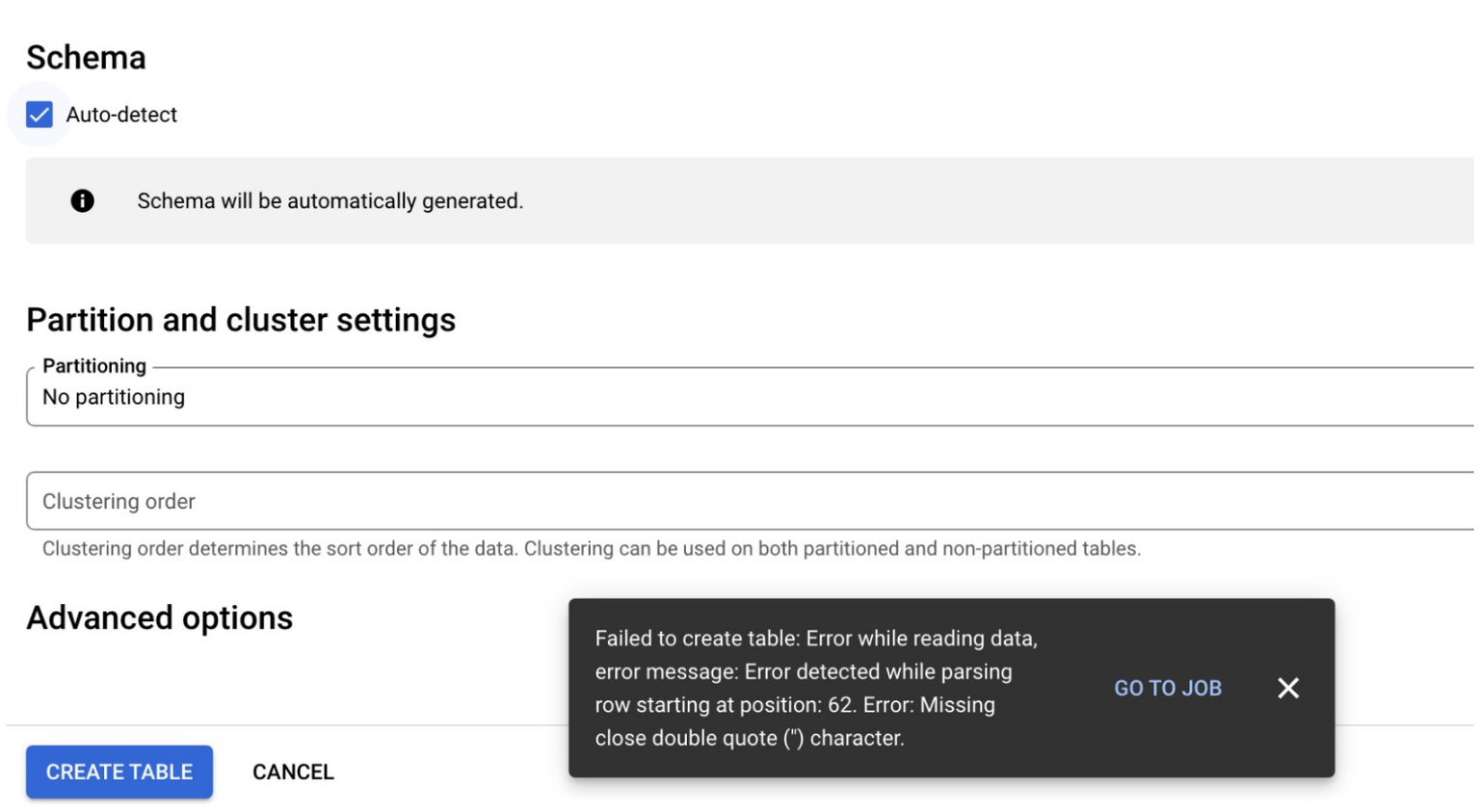
Get the features through these tables

- For example, prepare the Frequency of post, separate the image post and video post, also prepare the features for machine learning in the next step.

- Indexing

Before upload to BigQuery

- Jupiter-Lab
- Error return from BigQuery

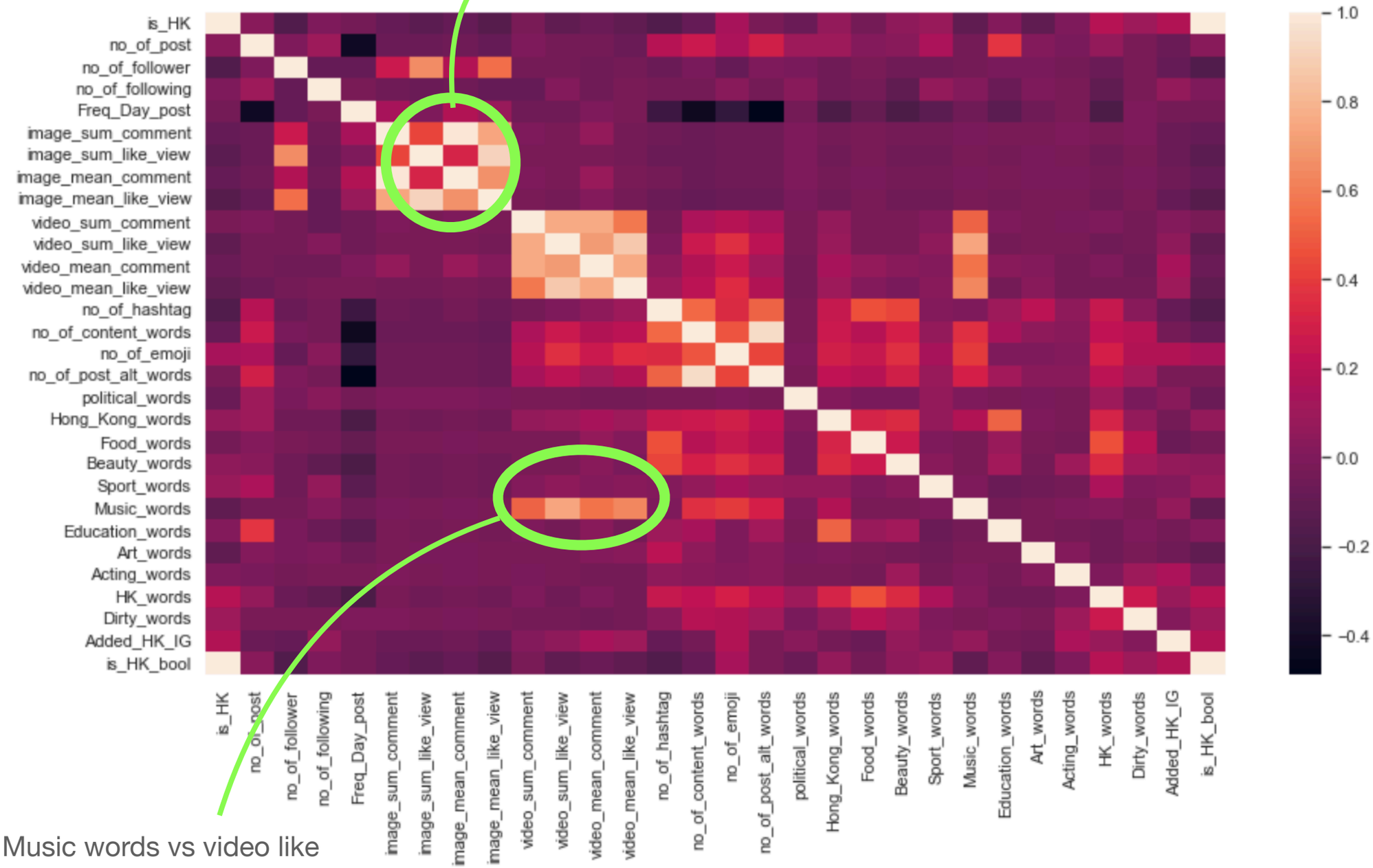


PART 3

MACHINE LEARNING

Correlation

Image post like vs comment



Music words vs video like

Classification

```
test_df2 = test_df[0:40]
data = {'food':    [0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,1,0,1,0,1,0,0,0,0,0,0,0,
        'beauty':  [0,0,1,1,0,0,0,1,0,1,1,1,0,1,0,0,0,0,0,0,0,0,0,1,0,1,0,1,0,0,0,0,0,0,0,0,
        'sport':   [0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,
        'music':   [0,0,0,0,0,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,1,1,0,0,0,1,0,
        'art':     [0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,1,1,1,1,1,0,1,0,1,0,1,0,0,1,0,0,0,0,0,0,0,
        'media':   [1,0,0,0,1,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,
        'hongkong':[1,0,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,0,1,1,1,1,
       }
```

train_size: 80%

test_size: 20%

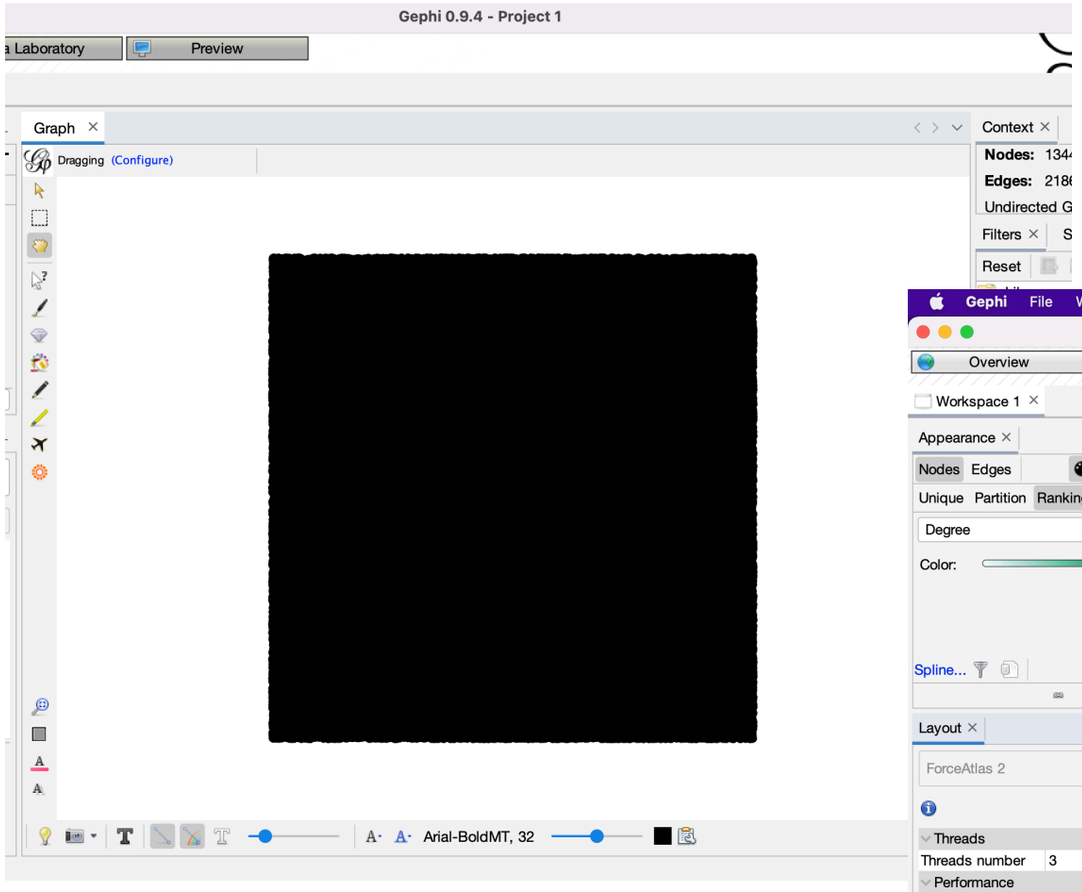
——>

Type	hongkong
RandomForest Model	
Actual	1101111
Predict	1101111
Result	100%
Gaussian Model	
Actual	11011111
Predict	11111111
Result	87.5%

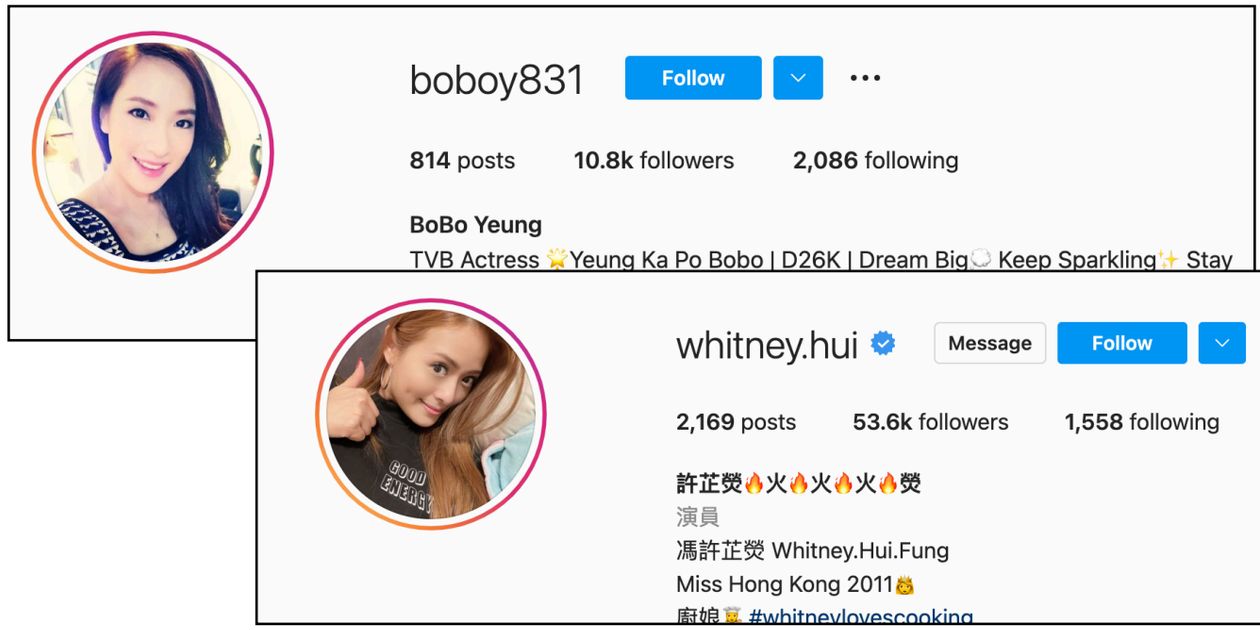
Graph Analysis

- Gephi

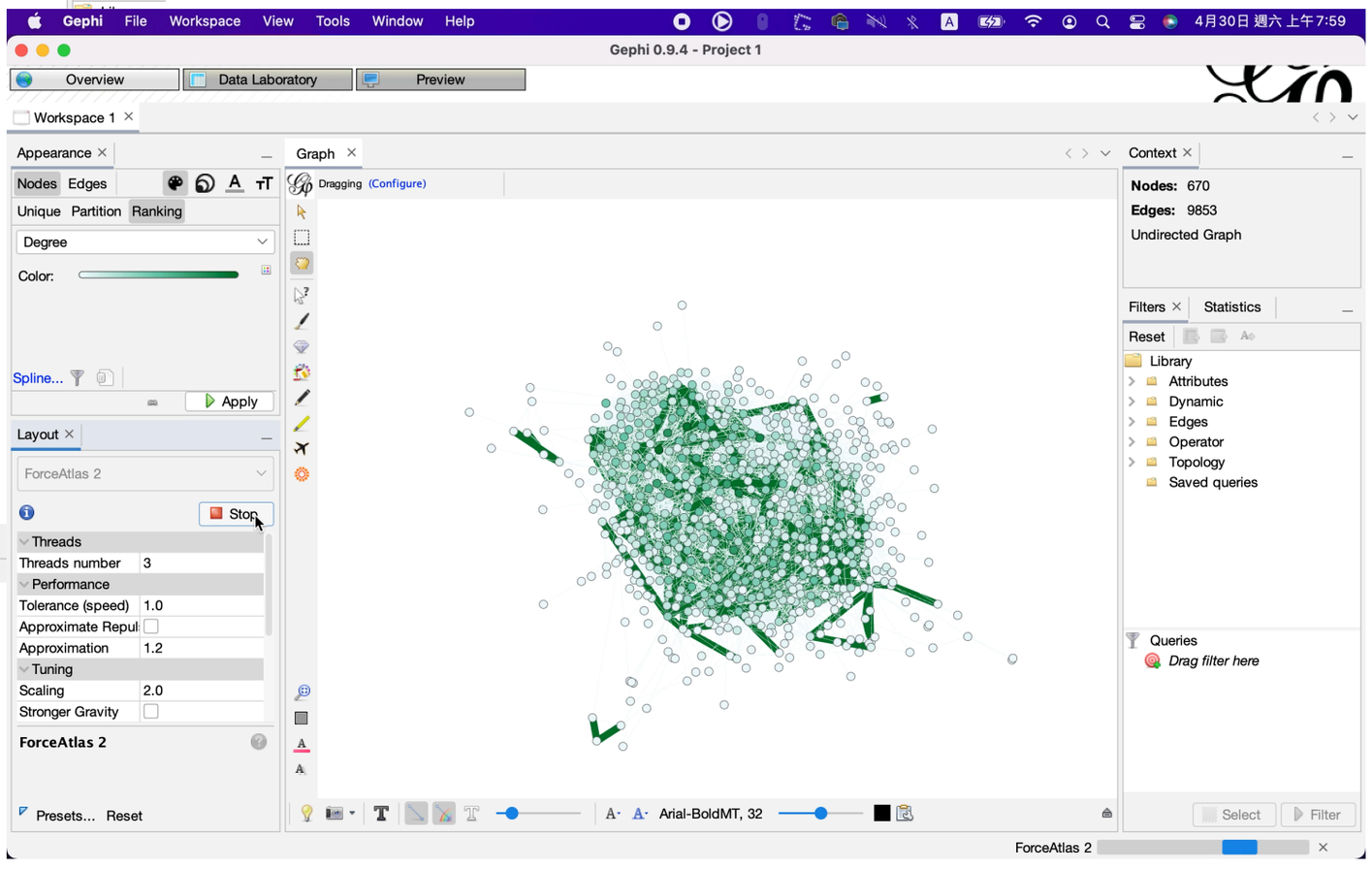
200,000 rows = dead



Center points:

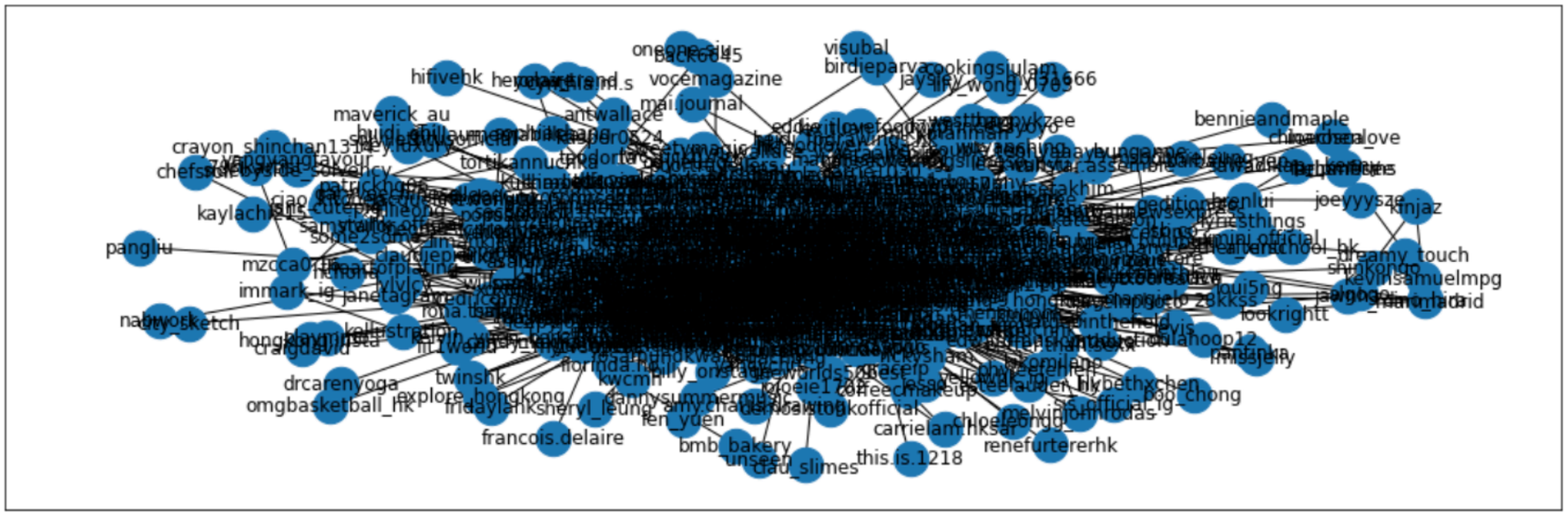


~10,000 rows



Graph Analysis

- networkx



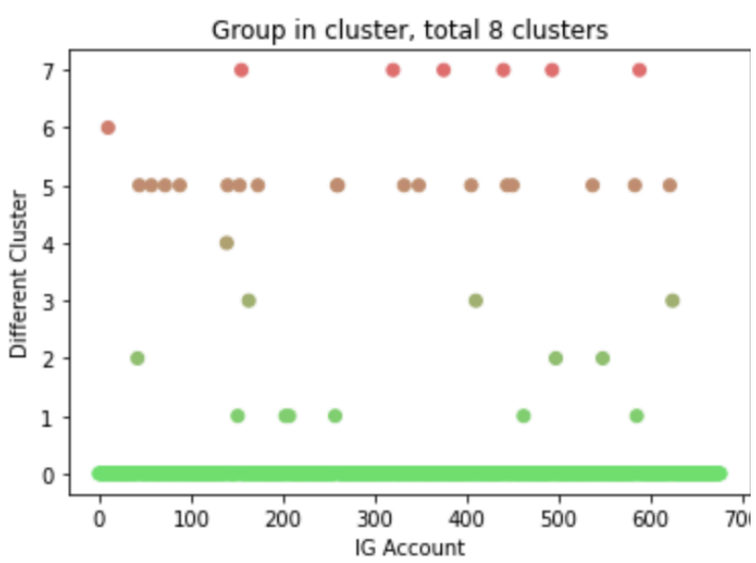

```
from networkx.algorithms.community import greedy_modularity_communities

c = list(greedy_modularity_communities(G))
```

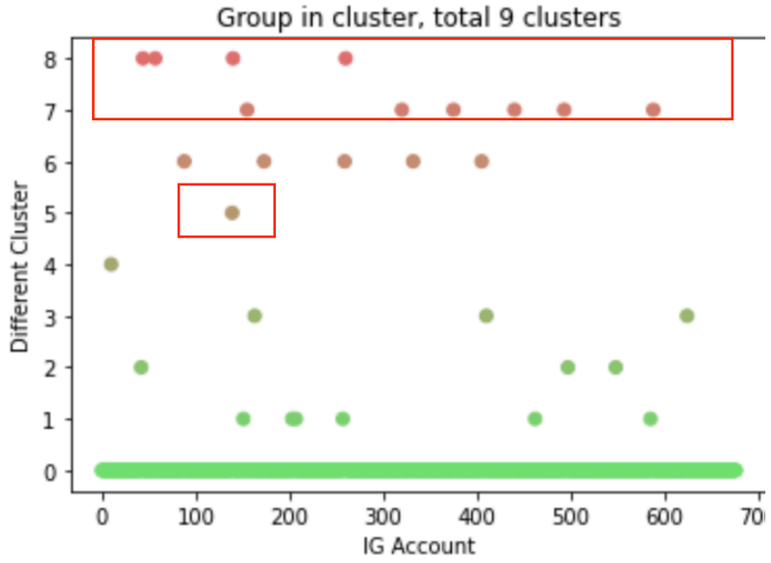
```
print("Number of groups:" , len(c))
print("Group sizes:" , list(map(len, c)))
```

```
Number of groups: 9
Group sizes: [339, 200, 89, 29, 5, 2, 2, 2, 2]
```

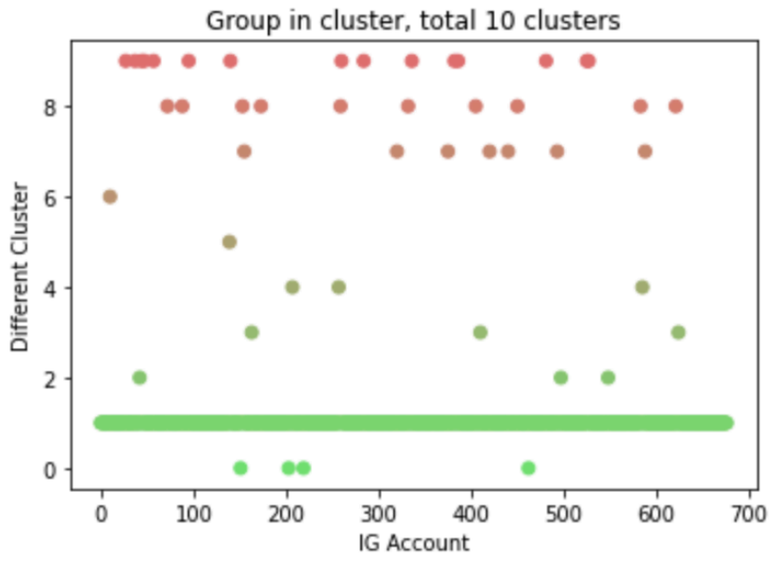
Clustering - KMeans



score: 0.8191871130983015499733 9



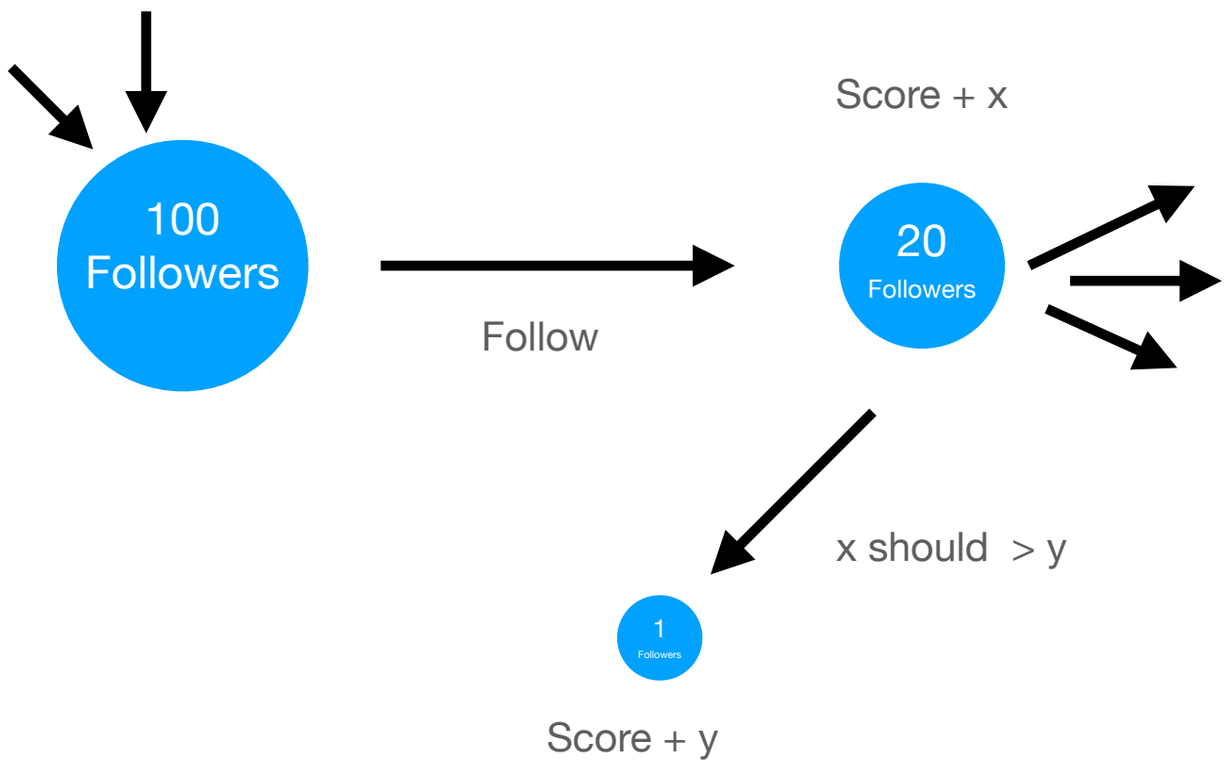
Score: 0.8512063034716708



Score: 0.79473254997339

- Total 8 different categories of words:
- Art, Food, Media, Sport, Education, Beauty, Music, Travel

Page Rank



In my case,
The original score is number of followers.

A follow B, B will gain A's original marks/mean

Finally, we will get the score index and then QUARTILE.

Divided into A, B, C, D

Recommender - Apriori

- Following list like a shopping busket
- When many people follow A and follow B at the meanwhile
- That mean there is great opportunity that people follow A, willing to follow B.

Example

{A} follow - > {B}, {C}

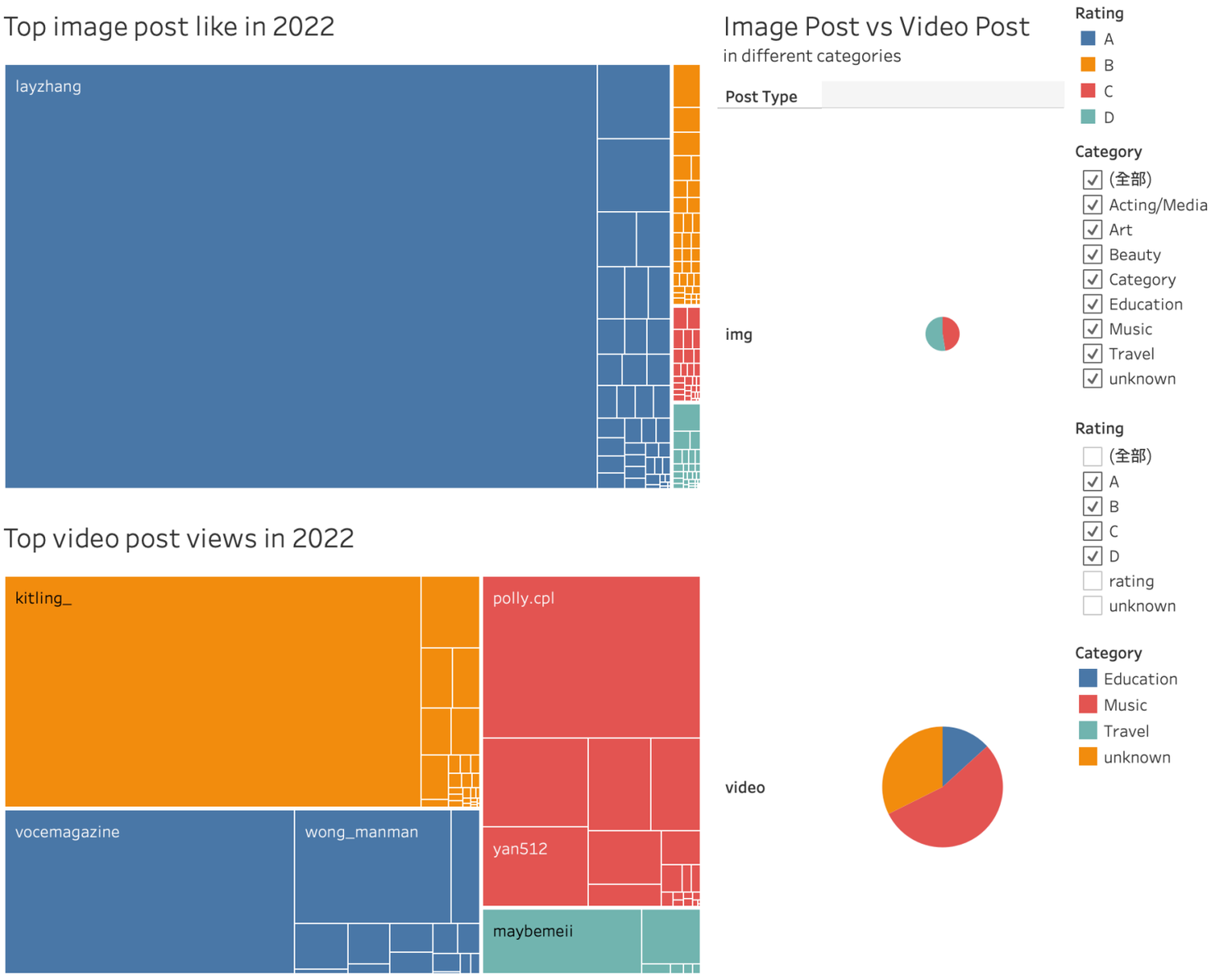
{D} follow - > {B, C, D}

When someone follow B, I will recommend C

PART 4
DEPLOYMENT

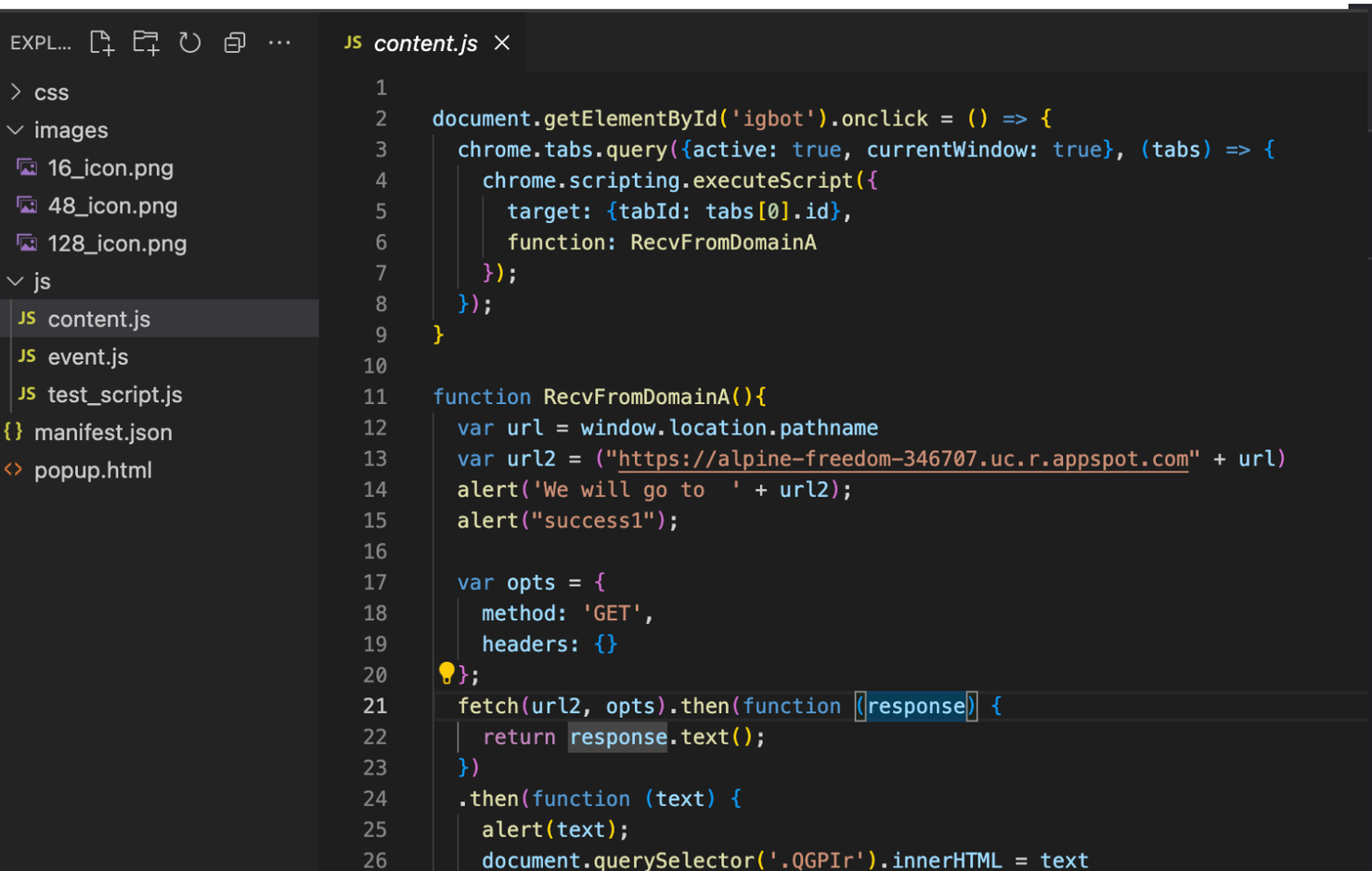
Tableau

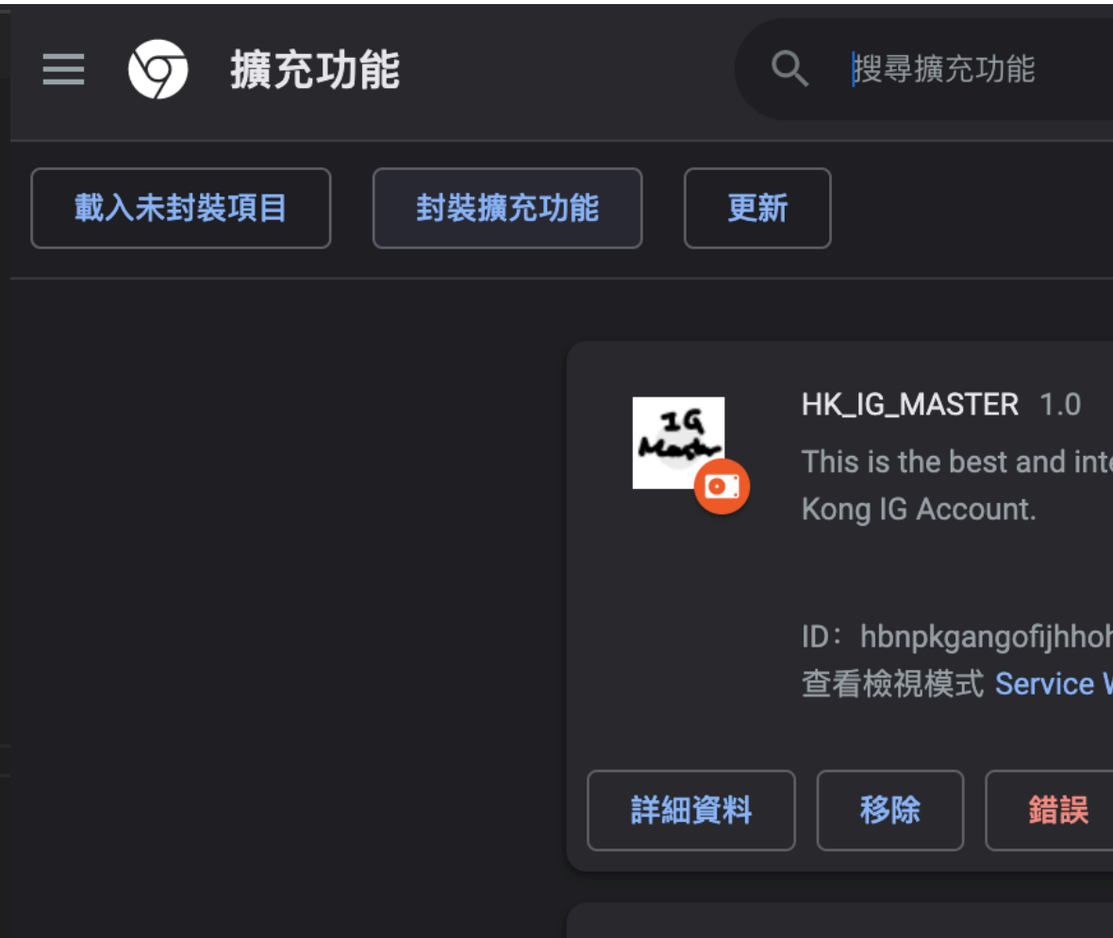
- Table for searching the IG KOL



Chrome Extension

- Easy to find the Machine learning result when browsing the Instagram website.
- App Engine > Chrome Extension





PART 5

FUTURE ENHANCEMENT

Think more …

- Scraping using real mobile should be better
- The best way is collecting all of the data
- Data set is too small
- Data structure (consider more)
- Use SQL/Spark
- Analysis the data deeply

PART 6
Q & A Section

THANKS FOR LISTENING