

Final Project

Instagram Data Analysis



Prepared by Ken
2022-04-30

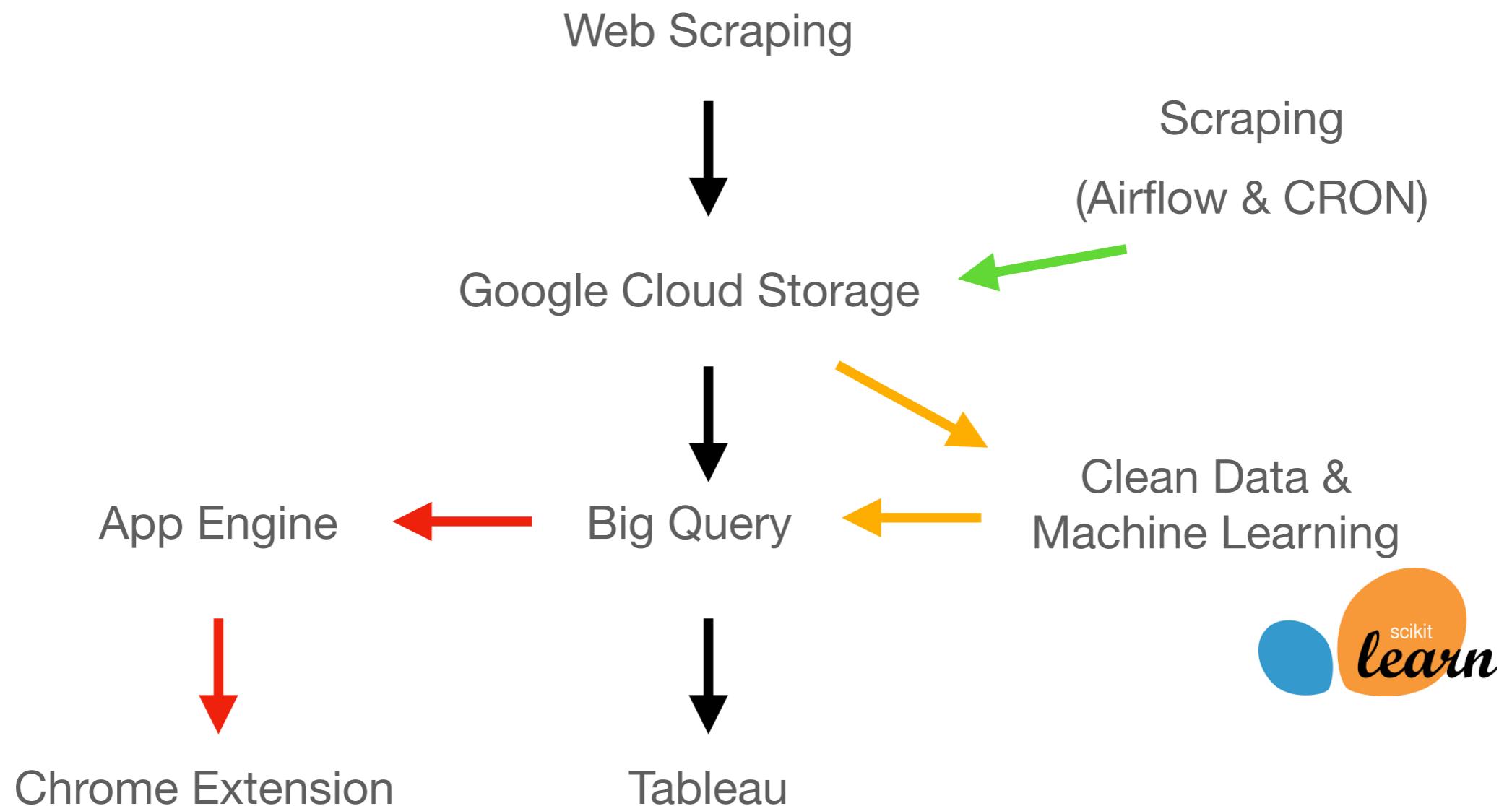
Problem

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

Purpose

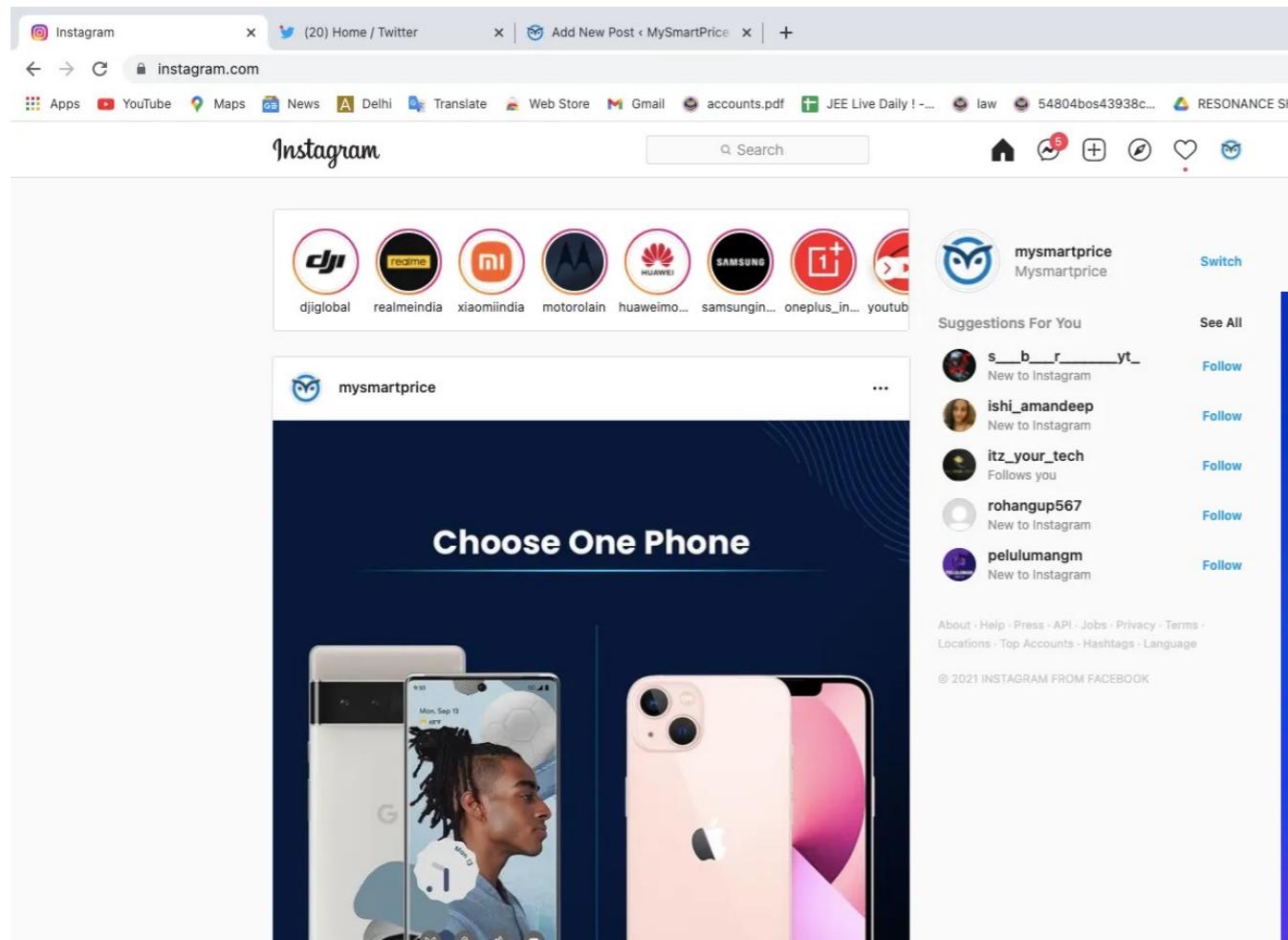
- To develop the tool to find out the KOL and it provides different features of IG account for analysis.
- This tool can continually update the IG KOL information.

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

The screenshot shows a website with a header containing a logo, a search bar, and navigation links for Home, Tutorial Library, Articles, and Videos. Below the header is a social sharing sidebar with icons for Facebook, Twitter, LinkedIn, Email, and a Plus sign. The main content area features a large green button with a yellow arrow pointing right. To the right of the button, the text "A to Z . . . and A" is visible. Below the button, the title "How to Scrap using BeautifulSoup" is displayed, with the word "Scrap" highlighted in red. A red box highlights the date "Jul 30, 2020". The post content begins with "This article is all about how we can grab the d..." followed by "username as an input in Python."

2022, request.get() not work anymore

```
35      <script type="text/javascript">
36      (function() {
37          var docElement = document.documentElement;
38          var classRE = new RegExp('(^|\\s)no-js(\\s|$)');
39          var className = docElement.className;
40          docElement.className = className.replace(classRE, '$1js$2');
41      })();
42  </script>
43  <script type="text/javascript">
44  (function() {
45      if ('PerformanceObserver' in window && 'PerformancePaintTiming' in window) {
46          window.__bufferedPerformance = [];
47          var ob = new PerformanceObserver(function(e) {
48              window.__bufferedPerformance.push.apply(window.__bufferedPerformance, e.getEntries());
49          });
50          ob.observe({entryTypes:['paint']});
51      }
52
53      window.__bufferedErrors = [];
54      window.onerror = function(message, url, line, column, error) {
55          window.__bufferedErrors.push({
56              message: message,
57              url: url,
58              line: line,
59              column: column,
60              error: error
61          });
62          return false;
63      };
64      window.__initialData = {
65          pending: true,
66          waiting: []
67      };
68      function asyncFetchSharedData(extra) {
69          var sharedDataReq = new XMLHttpRequest();
70          sharedDataReq.onreadystatechange = function() {
71              if (sharedDataReq.readyState === 4) {
72                  if(sharedDataReq.status === 200){
73                      var sharedData = JSON.parse(sharedDataReq.responseText);
74                      window.initialDataLoaded(sharedData, extra);
75                  }
76              }
77          };
78      }
79  })();
80  
```

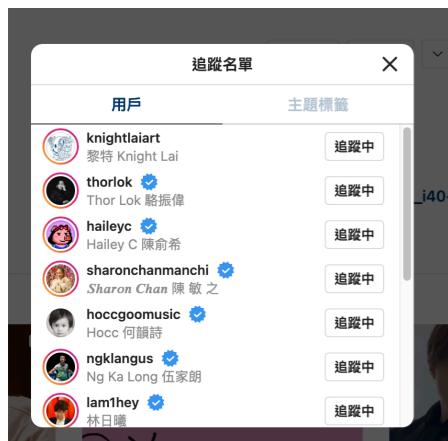
Selenium



- Planning:

<p>① Purpose</p> <p>A. Chrome Extension → Score of KOL</p> <p>B. Power BI Tableau → Data of HK KOL</p>	<p>② DATA partA</p> <p>A. KOL LIST CH(K)</p> <pre>graph TD; L1[List v1] --> T1[Temp]; T1 --> L2[Temp VR]; L2 --> L3[Append CT]; L3 --> T2[Temp APP]; T2 --> L4[Scrap APP]; L4 --> C[confirm is FK]; C --> L5[Scrap from startingage]; L5 --> R1[random select KOL & condition, follower > 10k]; R1 --> L6[Scrap following list]; L6 --> L7[Scrap APP]; L7 --> C;</pre>	<p>③ DATA part2</p> <p>B. KOL IG data</p> <ul style="list-style-type: none">Basic Data<ul style="list-style-type: none">→ name→ page description→ link→ no. followerPost<ul style="list-style-type: none">→ 3 months frequency→ like & comment→ view (video)→ hashtag → Type→ image	<p>④ Data Analyze</p> <p>for what (part 2 data)</p> <ul style="list-style-type: none">Profile ? → handsome?Score → emoji?→ video?Instagram? → image type?Churn detected?Hashtag people locationGrowth Rate?
---	--	---	---

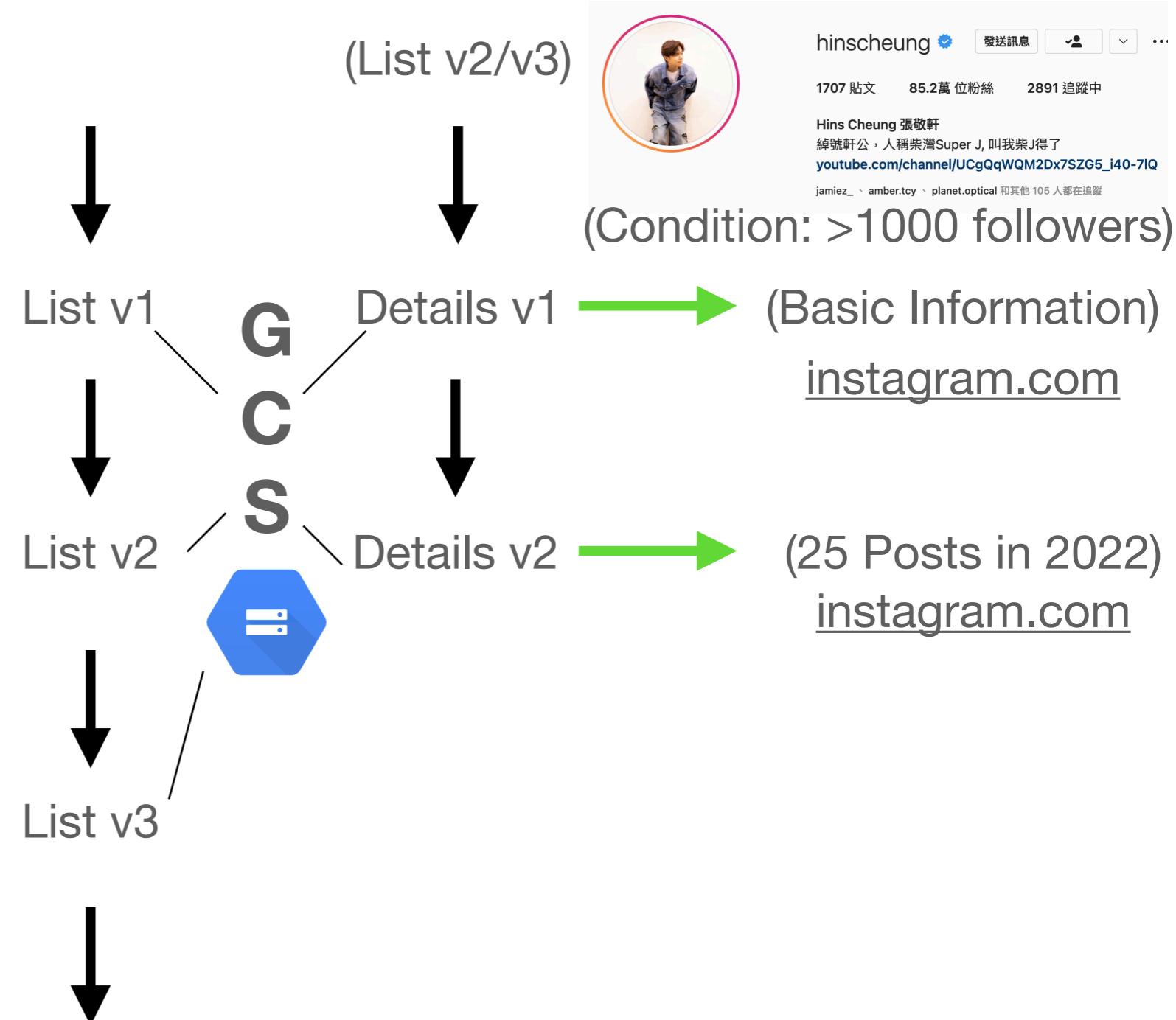
What?



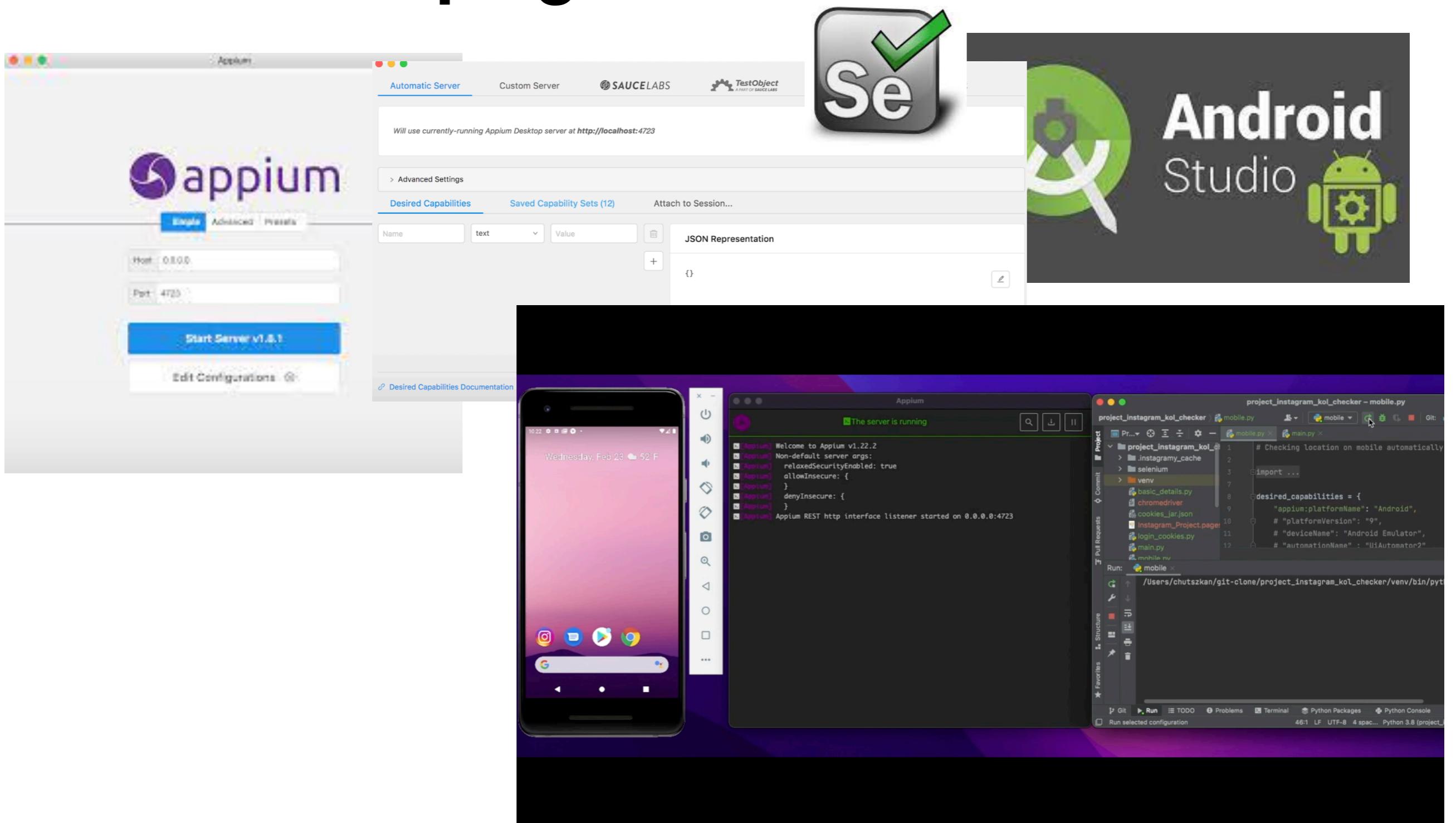
(Top 100 HK IG id)
startngage.com



(Confirm country)
Instagram App



Mobile Scraping



What's data I get

Csv1

- ig_id
- country
- updated following list time
- details updated time
- following list

Csv2

- ig_id
- name
- description
- no of post
- no of follower
- no of following
- profile image

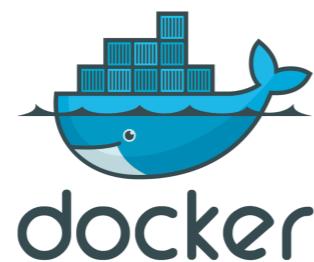
Csv3

- ig_id
- post date
- post alt
- post type
- like/view
- comment
- content

Secludedly scrap the data



kubernetes



docker



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

Schema

Auto-detect

ⓘ Schema will be automatically generated.

Partition and cluster settings

Partitioning —
No partitioning

Clustering order

Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.

Advanced options

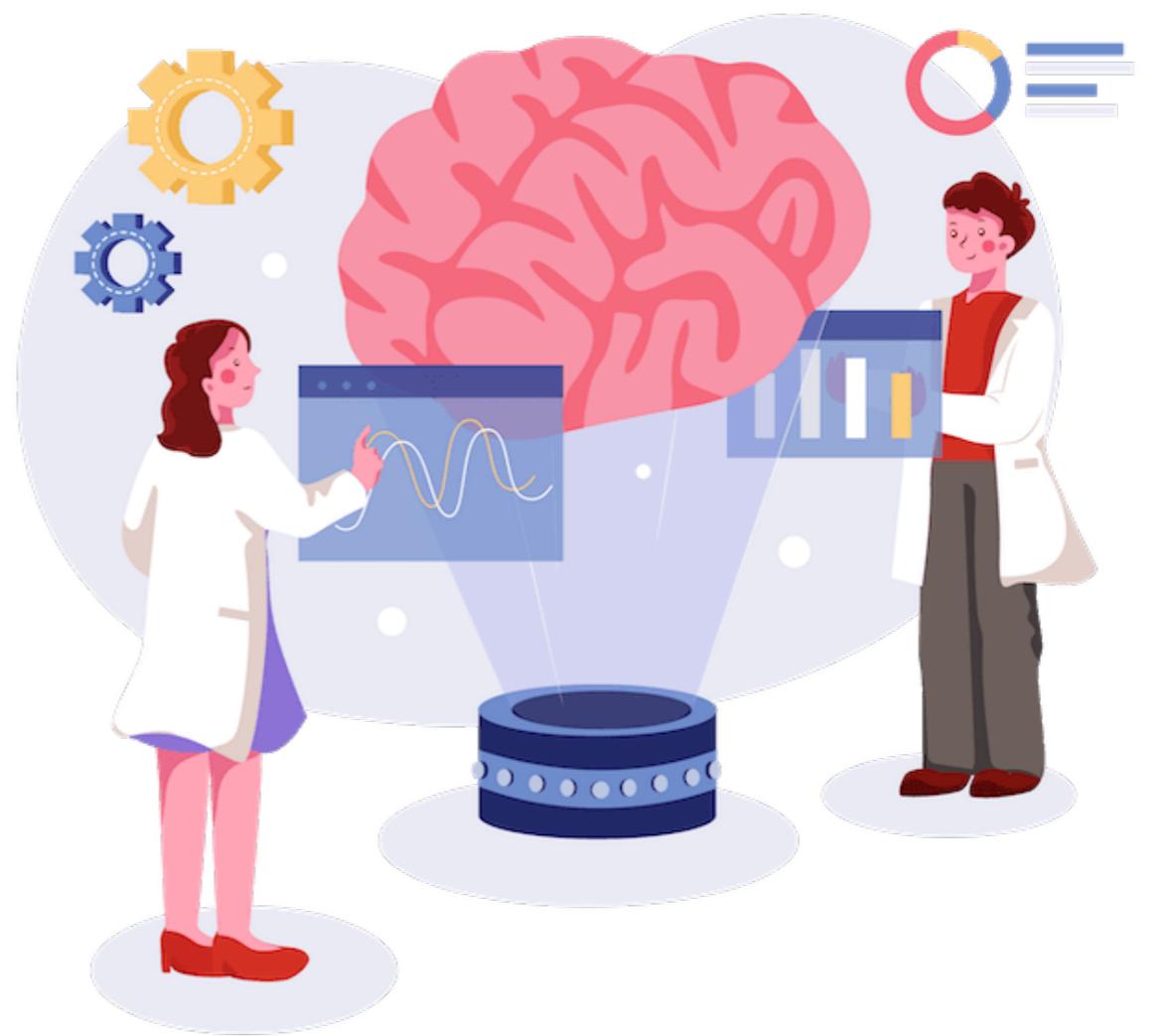
Failed to create table: Error while reading data,
error message: Error detected while parsing
row starting at position: 62. Error: Missing
close double quote ("") character.

GO TO JOB X

CREATE TABLE CANCEL

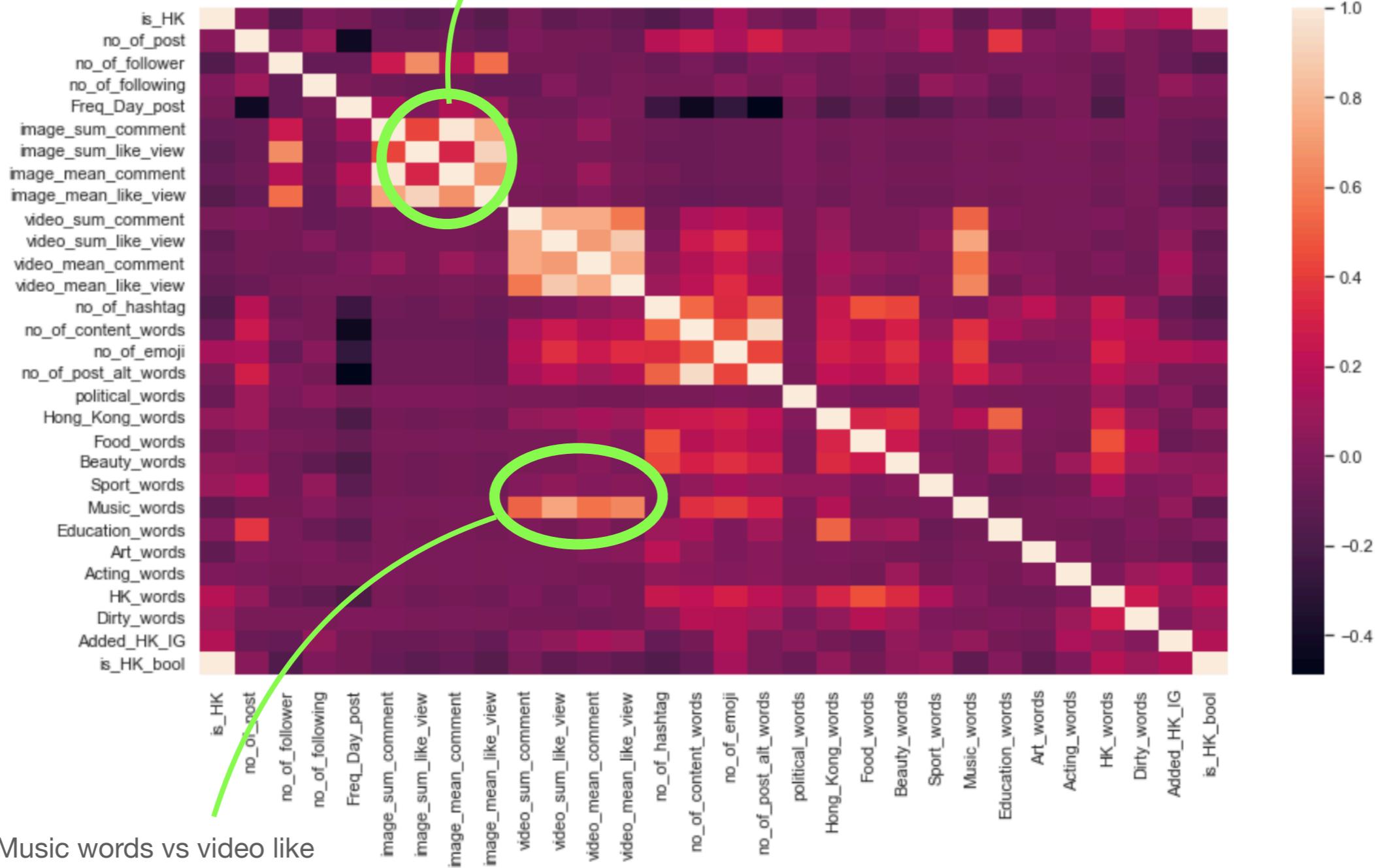
PART 3

MACHINE LEARNING



Correlation

Image post like vs comment



Classification

train_size: 80%

→

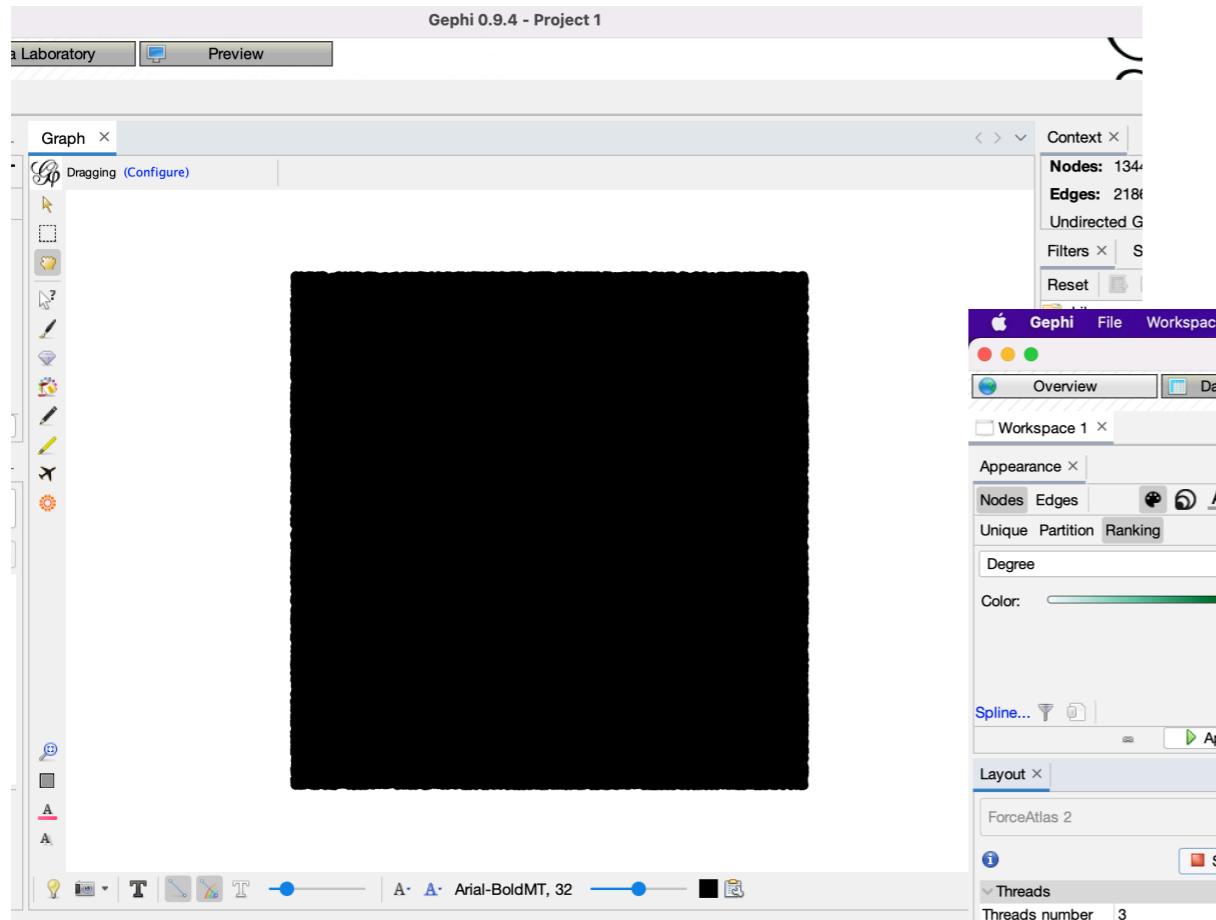
test_size: 20%

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:

boboy831 [Follow](#) [...](#)

814 posts 10.8k followers 2,086 following

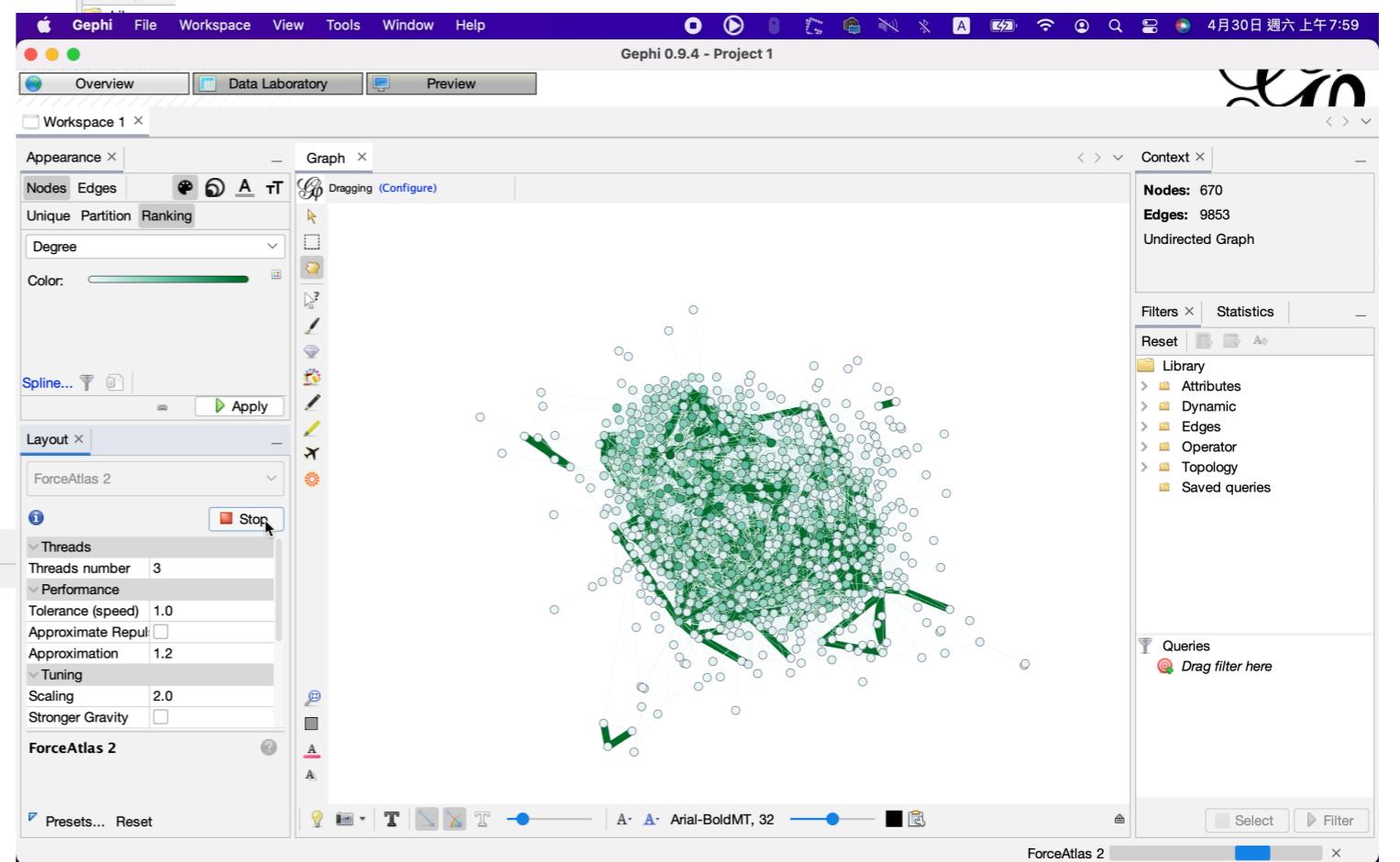
BoBo Yeung
TVB Actress ★ Yeung Ka Po Bobo | D26K | Dream Big Keep Sparkling Stay

whitney.hui [Message](#) [Follow](#) [...](#)

2,169 posts 53.6k followers 1,558 following

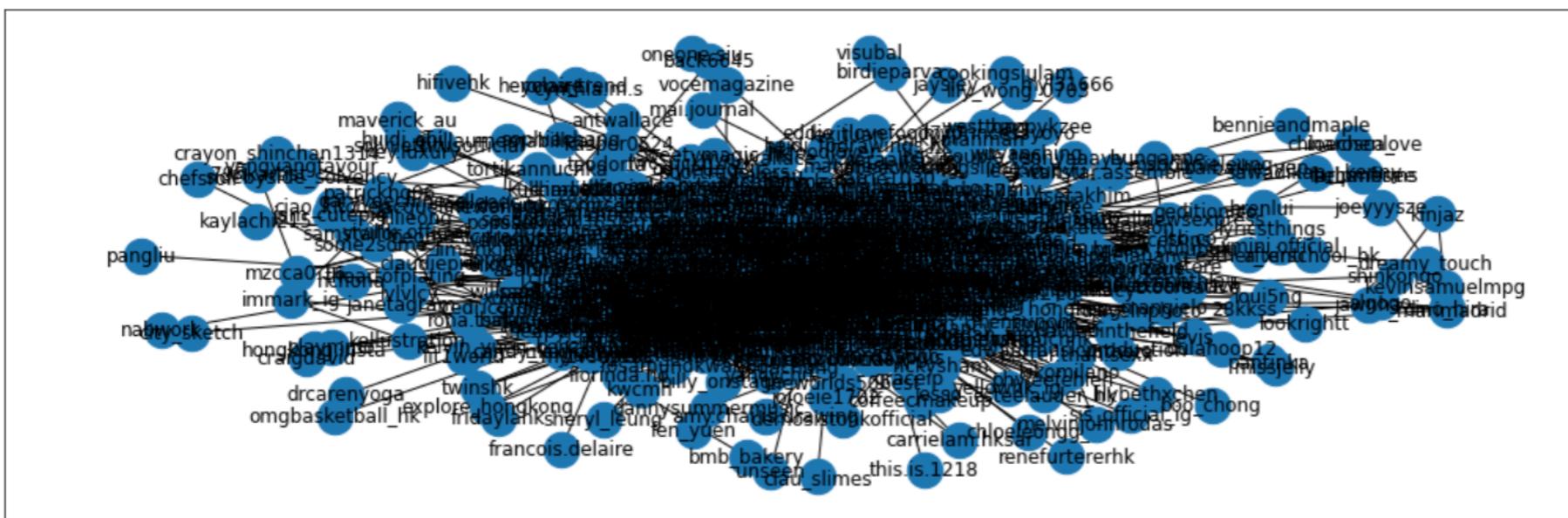
許芷熚🔥火🔥火🔥火🔥熚
演員
馮許芷熚 Whitney.Hui.Fung
Miss Hong Kong 2011👑
廚娘🍳 #whitnevlovescooking

~10,000 rows



Graph Analysis

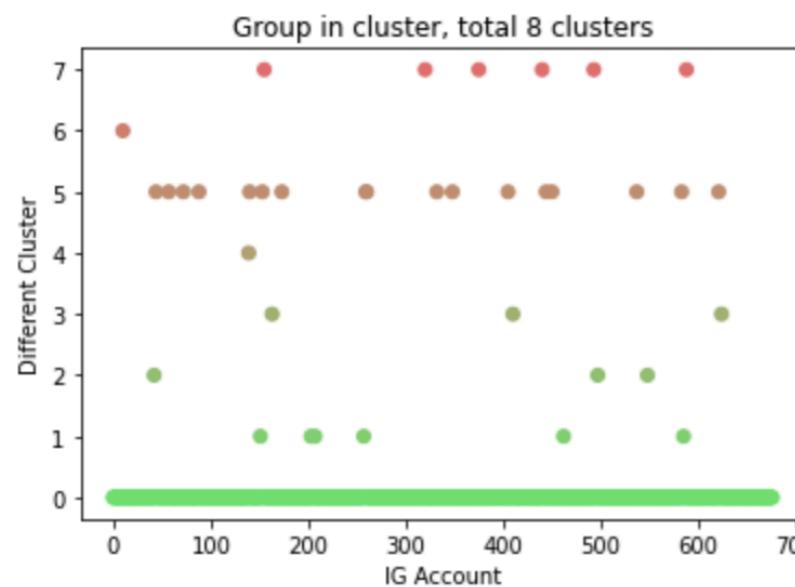
- networkx



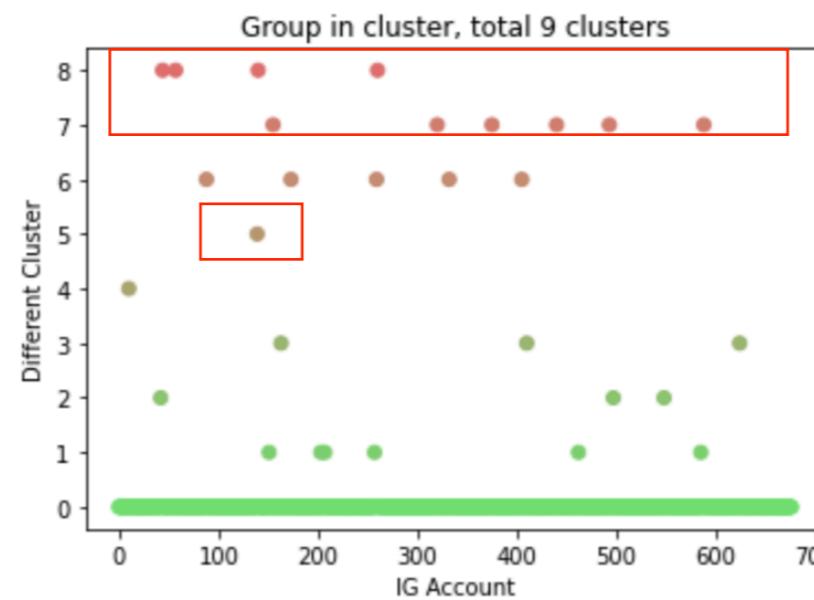
```
: from networkx.algorithms.community import greedy_modularity_communities  
  
c = list(greedy_modularity_communities(G))  
  
: print("Number of groups:" , len(c))  
print("Group sizes:" , list(map(len, c)))
```

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

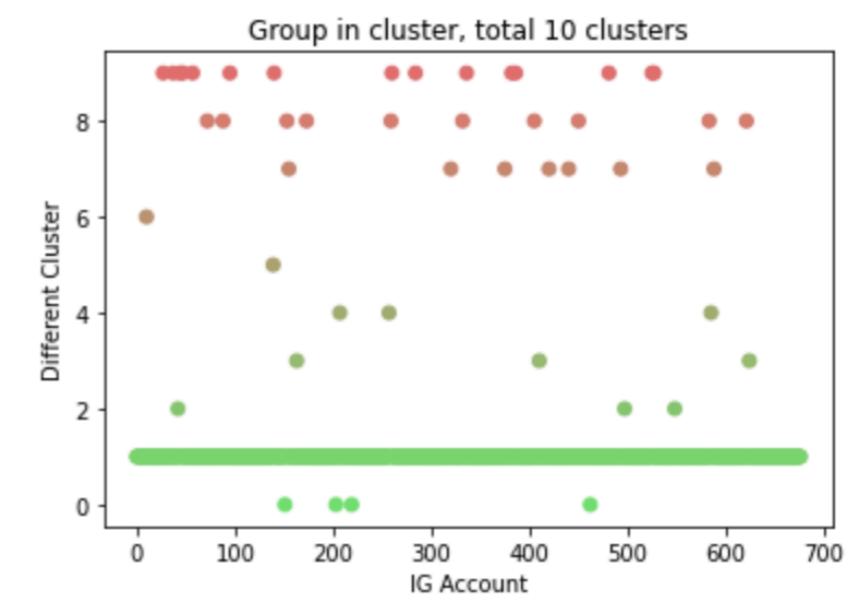
Clustering - KMeans



score: 0.81918711309830154997339



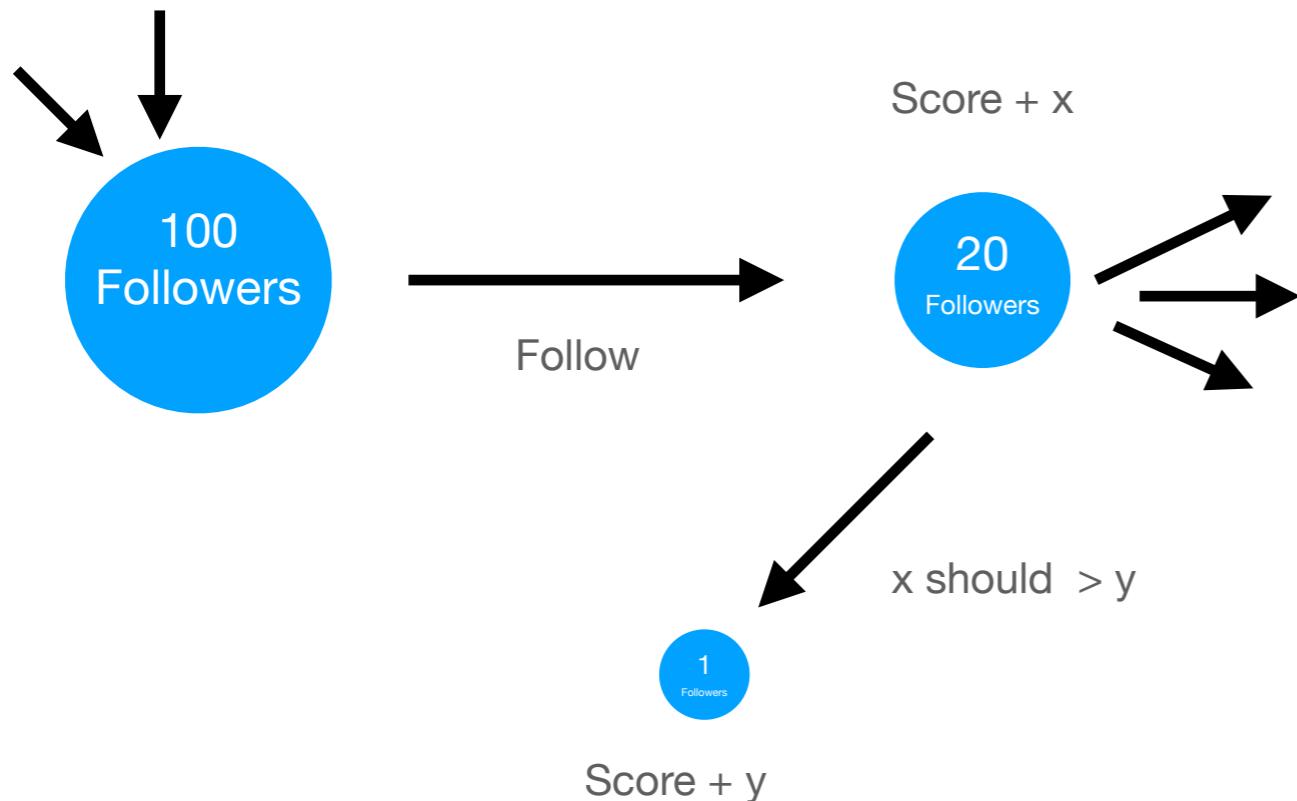
Score: 0.8512063034716708



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

Top image post like in 2022



Image Post vs Video Post
in different categories

Post Type

Rating

- ✓ A
- ✓ B
- ✓ C
- ✓ D

Category

- ✓ (全部)
- ✓ Acting/Media
- ✓ Art
- ✓ Beauty
- ✓ Category
- ✓ Education
- ✓ Music
- ✓ Travel
- ✓ unknown

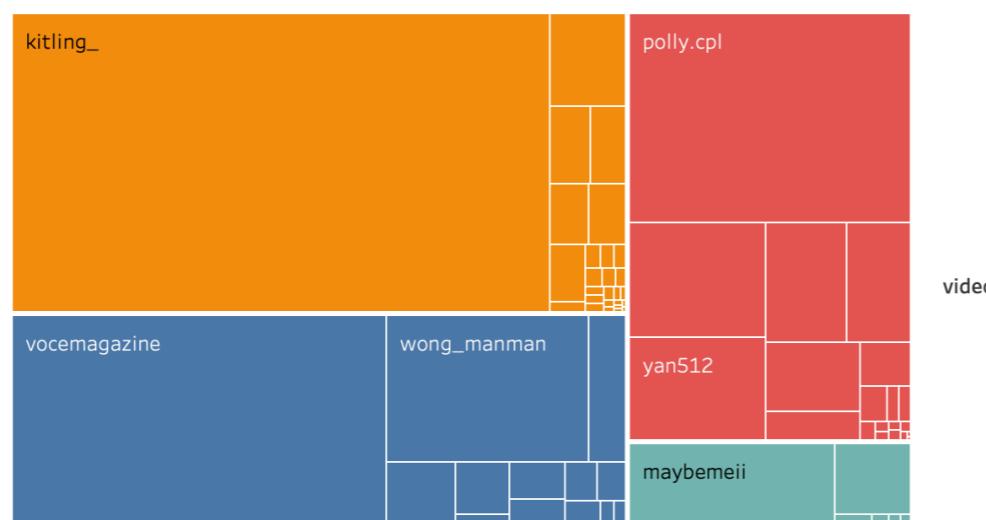
Rating

- (全部)
- ✓ A
- ✓ B
- ✓ C
- ✓ D
- rating
- unknown

Category

- Education
- Music
- Travel
- unknown

Top video post views in 2022



video



Chrome Extension

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

The screenshot shows a developer's environment with two main windows. On the left is a code editor with a dark theme, displaying the file `content.js`. The code is written in JavaScript and includes functions for handling button clicks and performing fetch requests to a specific URL. On the right is a screenshot of the Chrome extensions management page, showing the extension "HK_IG_MASTER" version 1.0. The extension icon features the letters "IG" and "Master". The page displays the extension's name, ID, and a brief description: "This is the best and int... Kong IG Account.". It also includes buttons for "詳細資料" (Details), "移除" (Remove), and "錯誤" (Error).

```
JS content.js ×
1
2  document.getElementById('igbot').onclick = () => {
3    chrome.tabs.query({active: true, currentWindow: true}, (tabs) => {
4      chrome.scripting.executeScript({
5        target: {tabId: tabs[0].id},
6        function: RecvFromDomainA
7      });
8    });
9  }
10
11 function RecvFromDomainA(){
12   var url = window.location.pathname
13   var url2 = ("https://alpine-freedom-346707.uc.r.appspot.com" + url)
14   alert('We will go to ' + url2);
15   alert("success1");
16
17   var opts = {
18     method: 'GET',
19     headers: {}
20   };
21   fetch(url2, opts).then(function (response) {
22     return response.text();
23   })
24   .then(function (text) {
25     alert(text);
26     document.querySelector('.QGPIr').innerHTML = text
}
```

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