




Emoji Analysis on Instagram

Data Engineering and Analytics project
presented by Stephanie Wong



Emojis have become
integral part of our digital
conversations

你覺得係乜？



鄙視？無奈？



陰濕？蠱惑？

WORLD EMOJI AWARDS Most 2023 Emoji



港聞 特價酒店 颱風 娛樂 國際 生活 即時 最Hit 體育 中國 科技 經濟

國際 / 環球趣聞

澳洲員工發短訊沒加emoji 上司一睇即嬲 怒擲手機繼而炒魷

撰文：中天新聞網

11:57

出版：2022-07-17 14:46

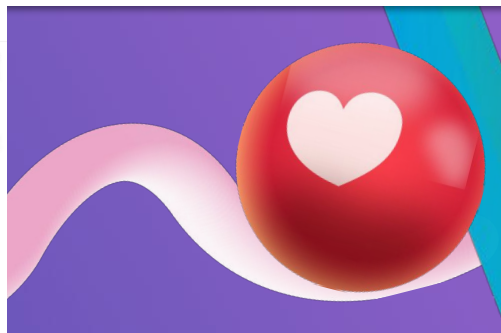


表情符號

optical.online.hk

追蹤中

2023熱門常用emoji



Purposes

- To explore the usage of emojis
- To provide directions for future emoji developments





Data Collection

Scope

- Instagram Post captions and comments
- Top Hong Kong Influencers of All categories



Tools

- **Web Scraper:** Instagram Post Scraper & Instagram Comment Scraper from Apify
- **Data Warehouse:** Databricks
- **Job Scheduler:** Databricks Scheduler
- **Version Control:** Git, Github



databricks



Workflow



01 Data Mining

- Collect IG accounts, Posts and Comments

04 Set up Scheduler

02 Data Cleaning

05 Data Analytics

- Sentiment Analysis
- Basket Analysis

03 Feature Engineering

06 Data Visualisation



01

Data Mining








Top Instagram Influencers in Hong Kong

Find top Instagram influencers in Hong Kong. Find out who is the #1 Instagram creator in 2023 and get a list of the most popular Instagram accounts.

[Get Started Now!](#)

All Categories Hong Kong ⓘ How we calculate 🔊 Share all

| Rank ⓘ | Influencer | Category | Followers | Country | Eng. (Auth.) | Eng. (Avg.) | | |
|--------|--|---|-----------|-----------|--------------|-------------|---|-----------------------------|
| 1 |  momo 모모 (MOMO) | Lifestyle | 13.4M | Turkey | 11M | 1.7M | 🔊 | Free Report |
| 2 |  gem0816 G.E.M. 鄧紫棋 | Music | 7.3M | China | 87.3K | 96.7K | 🔊 | Free Report |
| 3 |  stephenchow Stephen Chow | | 1.4M | Hong Kong | 105.2K | 116.4K | 🔊 | Free Report |
| 4 |  hamstagramss HK 🇭🇰 Lam's Fam (Lam's Fam...) | Animals | 111.3K | Hong Kong | 637.9K | 1.1M | 🔊 | Free Report |
| 5 |  ansonht Anson Lo 盧瀚霆 | Family Music Cinema & Actors/a... | 1.3M | Hong Kong | 39.3K | 64.9K | 🔊 | Free Report |
| 6 |  keung_show KEUNG TO 龔浩志 | | 1M | Hong Kong | 42.1K | 48.8K | 🔊 | Free Report |
| 7 |  mingi_choi 최민기 REN | | 1.2M | Hong Kong | 28.5K | 36.8K | 🔊 | Free Report |
| 8 |  edanlui Edan Lui 呂爵安 | Cinema & Actors/a... Music Shows | 700.1K | Hong Kong | 31.8K | 39.6K | 🔊 | Free Report |
| 9 |  wongsuiyu Priscilla Wong 黃翠如 | Lifestyle Cinema & Actors/a... | 1.2M | Hong Kong | 17.9K | 21.4K | 🔊 | Free Report |
| 10 |  lancychan Ian Chan 陳卓賢 | Lifestyle Music Cinema & Actors/a... | 597.5K | Hong Kong | 29K | 34.8K | 🔊 | Free Report |

1a. Collect IG Accounts

- Third-party Data from HypeAuditor (Marketing Platform)
- Web Scraping using *Requests*, *BeautifulSoup*



Table <ig_account_rank>

step_proj >

step_proj.ig_account_rank



Create

Owner: Not set Size: 110.8KiB, 3 files Last Updated: last week

Comment:

Columns Sample Data Details Permissions History

| Category | Rank | Influencer | Account_ID | Followers | Country | Followers_num |
|----------------|------|-----------------------|-----------------------|-----------|-----------|---------------|
| All Categories | 1 | 모모 (MOMO) | momo | 13.3M | Turkey | 133000000 |
| All Categories | 2 | G.E.M.鄧紫棋 | gem0816 | 7.3M | China | 73000000 |
| All Categories | 3 | Stephen Chow | stephenchow | 1.4M | Hong Kong | 14000000 |
| All Categories | 4 | HK 🇭🇰 Lam's Family 🇭🇰 | hamstagramsss | 111.5K | Hong Kong | 111500 |
| All Categories | 5 | Anson Lo • 盧瀚霆 | ansonlht | 1.3M | Hong Kong | 13000000 |
| All Categories | 6 | KEUNG TO 姜濤 | keung_show | 1M | Hong Kong | 10000000 |
| All Categories | 7 | 최민기 REN | mingi_choi | 1.2M | Hong Kong | 12000000 |
| All Categories | 8 | Yurick T cmm | yurick.17 | 153.4K | Hong Kong | 153400 |
| All Categories | 9 | Edan Lui 呂爵安 | edanlui | 700.1K | Hong Kong | 700100 |
| All Categories | 10 | Priscilla Wong 黃翠如 | wongtsuiyu | 1.2M | Hong Kong | 12000000 |
| All Categories | 11 | | on_9life | 403.1K | Hong Kong | 403100 |
| All Categories | 12 | 張柏芝 | cecilia_pakchi_cheung | 790.9K | Hong Kong | 790900 |
| All Categories | 13 | Ian Chan 陳卓賢 | ianychan | 597.5K | Hong Kong | 597500 |

[← View all Actors](#)

Instagram Post Scraper

`apify/instagram-post-scraper`**Try for free**

No credit card required

Use this no-code Instagram tool for scraping data from Instagram posts. Just add one or more Instagram usernames and get your data in seconds including hashtags, mentions, comments, images, likes, locations, and metadata. Download Instagram data in JSON, Excel, HTML and other formats.

[README](#) [Input](#) [API](#) [Changelog](#) [Related Actors](#)

API Client

API Endpoints

To run the code examples, you need to have an Apify account. Replace `<YOUR_API_TOKEN>` in the code with your [API token](#). For a more detailed explanation, please read about [running Actors via the API](#) in Apify Docs.

Node.js **Python** curl

```
from apify_client import ApifyClient

# Initialize the ApifyClient with your API token
client = ApifyClient("<YOUR_API_TOKEN>")

# Prepare the Actor input
run_input = {
    "username": ["zelenskiy_official"],
    "resultsLimit": 30,
}

# Run the Actor and wait for it to finish
run = client.actor("apify/instagram-post-scraper").call(run_input=run_input)

# Fetch and print Actor results from the run's dataset (if there are any)
for item in client.dataset(run["defaultDatasetId"]).iterate_items():
    print(item)
```

1b. Collect IG Posts

- *Instagram Post Scraper* from *Apify*
- **Inputs:**
 - **Username:**
Top 10 most followed in HK +
Top 3 most followed in each
category
 - **resultsLimit:**
50
- **Outputs:**
 - Id, type, caption, hashtags, mentions,
url, commentsCount, firstComment,
latestComments, likeCounts.....

Table <ig_account_rank>

Select most followed ig accounts

```
1 #Get ig accounts
2 account_df = spark.sql("select * from step_proj.ig_account_rank").toPandas()
3 all_df = account_df[(account_df["Category"] == "All Categories") & (account_df["Country"] == "Hong Kong")].sort_values(by=["Followers_num"], ascending=False).head(10)
4 ig_accounts = all_df["Account_ID"].values
5 for i in account_df["Category"].unique():
6     temp_df = account_df[(account_df["Category"] == i) & (account_df["Country"] == "Hong Kong")].sort_values(by=["Followers_num"], ascending=False).head(3)
7     ig_accounts = np.append(ig_accounts, temp_df["Account_ID"].values)
8 ig_accounts = list(set(ig_accounts))
```

Filtered ig accounts as Input

Web Scraper

Select key fields from
outputs and save into
Hive table

Table <ig_posts>

Table <ig_posts>

step_proj >

step_proj.ig_posts

Owner: Not set Size: 12.8MiB, 15 files Last Updated: 2 hours ago

Comment: Add comment

Columns Sample Data Details Permissions History

Q Filter columns...

| Column | Type |
|----------------|-----------|
| id | string |
| type | string |
| caption | string |
| hashtags | array |
| url | string |
| commentsCount | bigint |
| firstComment | string |
| latestComments | array |
| displayUrl | string |
| likesCount | bigint |
| timestamp | timestamp |
| ownerFullName | string |
| ownerUsername | string |
| ownerId | string |

| | id | type | shortCode | caption | hashtags | mentions | url | commentsCount | firstComment | latestComm |
|---|---------------------|-------|-------------|---|----------|--------------------------------|--|---------------|-------------------------|--------------------------------------|
| 0 | 3213153862000463855 | Image | CyXamnZNk_v | [Guest] 希望可以抽中我啦🥰 Thx Admin 話說小妹有一個由小學到而家中學... | | [戀人未滿, 朋友, 感情, 鍾意, hkgirl] | https://www.instagram.com/p/CyXamnZNk_v/ | 22 | 人生苦短, 去回顧今天的你, 將來只係雲淡風輕 | '1800356322192': 'text': '人生苦短' |
| 1 | 3212580102416717320 | Image | CyVYJUJCvYI | [Guest] 正所謂「少年情懷總是詩」。今朝心情直教心如鹿撞。我係一個中六既毒毒, 舊年成年... | | [女同學, 青春, 愛情, 感情煩惱, 拍拖, hkboy] | https://www.instagram.com/p/CyVYJUJCvYI/ | 9 | 專燃總是容易動情🥰 | '1796694821662': 'text': '專燃總是容易動情' |
| 2 | 3200980370669241444 | Image | CxsKrAGvQRk | 搬咗入我嘅My home已經4個月, 最後一件大傢俬, 梳化都買咗喇~ \\n4個月前條片, 成間... | | | https://www.instagram.com/p/CxsKrAGvQRk/ | 3 | 👍👍 | '1826504487409': 'text': '👍👍' |
| 3 | 3211704498552521586 | Video | CySRDmfV0Ny | [Guest]我鍾意咗我個friend...\\n\\n我係女仔 識咗個男仔兩年半\\n\\n咁咁識(巨無耐... | | [各校 secrets, 沉船, 暗恋, 朋友, 好感] | https://www.instagram.com/p/CySRDmfV0Ny/ | 28 | @cc_08.21 | '1804178731353': 'text': '@cc_08.21' |

Instagram Comment Scraper

API Client API Endpoints

To run the code examples, you need to have an Apify account. Replace `<YOUR_API_TOKEN>` in the code with your [API token](#). For a more detailed explanation, please read about [running Actors via the API](#) in Apify Docs.

Node.js Python curl



```
import { ApifyClient } from 'apify-client';

// Initialize the ApifyClient with API token
const client = new ApifyClient({
  token: '<YOUR_API_TOKEN>',
});

// Prepare Actor input
const input = {
  "directUrls": [
    "https://www.instagram.com/p/CH-MgQOn-7E/",
    "https://www.instagram.com/p/Bi-hISighYe/"
  ],
  "resultsLimit": 24
};

(async () => {
  // Run the Actor and wait for it to finish
  const run = await client.actor("apify/instagram-comment-scraper").call(input);

  // Fetch and print Actor results from the run's dataset (if any)
  console.log('Results from dataset');
  const { items } = await client.dataset(run.defaultDatasetId).listItems();
  items.forEach((item) => {
    console.dir(item);
  });
})();
```

1c. Collect IG Comments

- *Instagram Comment Scraper* from *Apify*
- **Inputs:**
 - **directUrls:**
Top 15 most commented post of each scraped ig account
 - **resultsLimit:**
100
- **Outputs:**
 - Id, type, caption, hashtags, mentions, url, commentsCount, firstComment, latestComments, likeCounts.....



Table <ig_posts>

Select most commented posts of each ig accounts

```
1 #Get ig posts
2 post_df = spark.sql("select * from step_proj.ig_post_init").toPandas()
3 temp_list=[]
4 for i in post_df["ownerId"].unique():
5     temp_df = post_df[(post_df["ownerId"] == i)].sort_values(by=["commentsCount"], ascending=False).head(15)
6     temp_list = np.append(temp_list, temp_df["url"].values)
7 ig_posts = list(set(temp_list))
8
9 len(ig_posts)
```

Filtered ig post urls as Input

Web Scraper

Select key fields from
outputs and save into
Hive table



Table <ig_comment>



Table <ig_comment>

step_proj >

step_proj.ig_comment

Create

Owner: Not set Size: 7.8MiB, 16 files Last Updated: 11 hours ago

Comment: Add comment

Columns Sample Data Details Permissions History

| postUrl | id | text | ownerUsername | timestamp | likesCount |
|--|-------------------|--|------------------|------------------------------|------------|
| https://www.instagram.com/p/CxE54AIS7Ju/ | 18224244145217060 | @hkdisneyplus | hymelonmelon | 2023-09-12T04:04:57.000+0000 | 1 |
| https://www.instagram.com/p/CxE54AIS7Ju/ | 17969935544425455 | @tseannatse | hiu_tinggggg | 2023-09-12T05:06:38.000+0000 | 1 |
| https://www.instagram.com/p/CxE54AIS7Ju/ | 17846064723061032 | @hiu_tinggggg I buy gold card | tseannatse | 2023-09-13T01:15:00.000+0000 | 0 |
| https://www.instagram.com/p/CxE54AIS7Ju/ | 17981408735421937 | @tseannatse will you go with me? | hiu_tinggggg | 2023-09-13T05:17:32.000+0000 | 0 |
| https://www.instagram.com/p/CpOmx_TJWNH/ | 17985290566792298 | ❤❤ | theformalesthete | 2023-03-10T17:08:13.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 18041963518396719 | 👍👍👍👍 | antonsd | 2022-12-28T07:15:19.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 17983897255806689 | Congratulations po Ma'am @chayecabalrevilla 💜👉👊 | thejeffcira | 2022-12-28T07:18:34.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 17897729165722601 | Congrats boss! 🍌🍌 | hechyboy | 2022-12-28T07:20:32.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 17951691644462106 | Congrats Miss @chayecabalrevilla 🌟👍👍👍 | misskeithmanila | 2022-12-28T07:23:56.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 17986610809649634 | 👍👍👍👍 | gsingzon | 2022-12-28T07:47:27.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 17950050011247100 | 🍌🍌🍌congratulations Ms Chaye❤❤❤ | mjpestrada | 2022-12-28T07:48:17.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 18226164955163196 | Congratulations po, Ma'am @chayecabalrevilla !! 😊❤ | mrskjtortes | 2022-12-28T07:48:50.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 17955060059349725 | Congratulations, Maam Chaye!! 🍌 | kathrinyu | 2022-12-28T07:51:48.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 17988037210703260 | 🍌🍌🍌🍌 Bravo bravo !!! 🍌🍌 | thejojieworld | 2022-12-28T07:53:57.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 17984079073815327 | Congratulations!!! | ricamisra | 2022-12-28T07:58:19.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 18073273276331304 | Congratulations! 🌟❤ | _sweetlifecandy_ | 2022-12-28T08:22:37.000+0000 | 0 |
| https://www.instagram.com/p/Cms-HQ0PxJk/ | 18236294503152668 | 🍌🍌🍌🍌 | cacaimitra | 2022-12-28T09:30:14.000+0000 | 0 |



02&03

Data Cleaning & Feature engineering



1. Combine Post caption and comments into a single Dataframe



| | text | postUrl | is_post | emojis | text_cleaned |
|--------|--|--|---------|--------|---|
| 0 | /\n\nBringing back the trendy rectangular eyew... | https://www.instagram.com/p/CugrZRvrsSI/ | True | 👁️ | 👁️ |
| 4 | 【#熱貓生活】🐱\n日本一位貓奴分享了自家貓咪 (Goma) 絕育後的趣事。她發現Goma的身... | https://www.instagram.com/p/Cxc2EHUrBUQ/ | True | 🐱🐱🐱🐱 | 熱貓生活🐱🐱日本一位貓奴分享了自家貓咪絕育後的趣事她發現的身體狀況正常但就會不停發出汪汪的叫... |
| 5 | 【恭喜中大同學麥世霆成香港首位奪世青賽金牌男泳手! 🏆🏆】\n#中大健康與體育運動科學 一年... | https://www.instagram.com/p/Cw7V6EltxKX/ | True | 🏆🏆🏆🏆 | 恭喜中大同學麥世霆成香港首位奪世青賽金牌男泳手🏆🏆 🏆中大健康與體育運動科學一年級學生麥世霆日... |
| 6 | 應該就冇可能\n\n不管都差唔多\n\n直上直落咁解 🐱\n\n#貓 #插畫 | https://www.instagram.com/p/Cj5Q9BeSAIO/ | True | 🐱 | 應該就冇可能不管都差唔多直上直落咁解🐱#貓插畫 |
| 7 | My skin care routine for long flights ✈️\n50%... | https://www.instagram.com/p/CxSev1HP5rA/ | True | ✈️ | ✈️半價搶👁️機上護膚 |
| ... | ... | ... | ... | ... | ... |
| 124559 | 🙏🙏🙏 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏 | 🙏🙏🙏 |
| 124561 | 太慘，所以應該要有安樂死合法化🙏 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏 | 太慘所以應該要有安樂死合法化🙏 |
| 124562 | 🙏🙏all the best 🙏 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏 | 🙏🙏🙏 |
| 124563 | 🙏🙏🙏🙏🙏 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏🙏🙏 | 🙏🙏🙏🙏🙏 |
| 124565 | 🙏🙏🙏R.I.P | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏 | 🙏🙏🙏 |





2. Distinguish posts and comments

| | text | postUrl | is_post | emojis | text_cleaned |
|--------|--|---|---------|---------|---|
| 0 | /\n\nBringing back the trendy rectangular eyew... | https://www.instagram.com/p/CugrZRvrsS | True | 👁️ | 👁️ |
| 4 | 【#熱貓生活】🔥🐱\n日本一位貓奴分享了自家貓咪 (Goma) 絕育後的趣事。她發現Goma的身... | https://www.instagram.com/p/Cxc2EHUrbuQ | True | 🔥🐱👁️👂 | 熱貓生活🔥🐱日本一位貓奴分享了自家貓咪絕育後的趣事她發現的身體狀況正常但就會不停發出汪汪的叫... |
| 5 | 【恭喜中大同學麥世霆成香港首位奪世青賽金牌男泳手! 🏆🏊‍♂️】\n#中大健康與體育運動科學 一年... | https://www.instagram.com/p/Cw7V6EltxKX | True | 🏆🏊‍♂️👏👏 | 恭喜中大同學麥世霆成香港首位奪世青賽金牌男泳手🏆🏊‍♂️中大健康與體育運動科學一年級學生麥世霆日... |
| 6 | 應該就冇可能\n\n不管都差唔多\n\n直上直落咁解 🐱\n\n#貓 #插畫 | https://www.instagram.com/p/Cj5Q9BeSAIO | True | 🐱👁️ | 應該就冇可能不管都差唔多直上直落咁解🐱👁️貓插畫 |
| 7 | My skin care routine for long flights ✈️ \n50%... | https://www.instagram.com/p/CxSev1HP5rA | True | ✈️👁️ | ✈️半價搶👁️機上護膚 |
| ... | ... | ... | ... | ... | ... |
| 124559 | 🙏🙏🙏 | https://www.instagram.com/p/CymmM1zu26 | False | 🙏🙏🙏 | 🙏🙏🙏 |
| 124561 | 太慘，所以應該要有安樂死合法化🥲 | https://www.instagram.com/p/CymmM1zu26 | False | 🥲 | 太慘所以應該要有安樂死合法化🥲 |
| 124562 | 🙏🙏all the best 🥲 | https://www.instagram.com/p/CymmM1zu26 | False | 🙏🙏🥲 | 🙏🙏🥲 |
| 124563 | 🙏🙏🙏🙏🙏 | https://www.instagram.com/p/CymmM1zu26 | False | 🙏🙏🙏🙏🙏 | 🙏🙏🙏🙏🙏 |
| 124565 | 🙏🙏🙏R.I.P | https://www.instagram.com/p/CymmM1zu26 | False | 🙏🙏🙏 | 🙏🙏🙏 |

3. Extract Emojis from texts

- import *emoji*
- *emoji.EMOJI_DATA*: A dictionary of existing emoji

| | text | postUrl | is_post | emojis | text_cleaned |
|--------|---|--|---------|--------|---|
| 0 | /\n\nBringing back the trendy rectangular eyew... | https://www.instagram.com/p/CugrZRvrsSI/ | True | 👁️ | 👁️ |
| 4 | 【#熱貓生活】🐱🐱\n日本一位貓奴分享了自家貓咪 (Goma) 絕育後的趣事。她發現Goma的身... | https://www.instagram.com/p/Cxc2EHUrbuQ/ | True | 🐱🐱🗣️🗣️ | 熱貓生活🐱🐱日本一位貓奴分享了自家貓咪絕育後的趣事她發現的身體狀況正常但就會不停發出汪汪的叫... |
| 5 | 【恭喜中大同學麥世霆成香港首位奪世青賽金牌男泳手! 🏆🏆】\n#中大健康與體育運動科學 一年... | https://www.instagram.com/p/Cw7V6EltxKX/ | True | 🏆🏆🏆🏆🏆🏆 | 恭喜中大同學麥世霆成香港首位奪世青賽金牌男泳手🏆🏆 🏆中大健康與體育運動科學一年級學生麥世霆日... |
| 6 | 應該就冇可能\n\n不管都差唔多\n\n直上直落咁解 🐱\n\n#貓 #插畫 | https://www.instagram.com/p/Cj5Q9BeSAIO/ | True | 🐱🗣️ | 應該就冇可能不管都差唔多直上直落咁解🐱🗣️貓插畫 |
| 7 | My skin care routine for long flights ✈️ \n50%... | https://www.instagram.com/p/CxSev1HP5rA/ | True | ✈️👁️ | ✈️半價搶👁️機上護膚 |
| ... | ... | ... | ... | ... | ... |
| 124559 | 🙏🙏🙏 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏 | 🙏🙏🙏 |
| 124561 | 太慘，所以應該要有安樂死合法化🥲 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🥲 | 太慘所以應該要有安樂死合法化🥲 |
| 124562 | 🙏🙏all the best 🥲 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🥲 | 🙏🙏🥲 |
| 124563 | 🙏🙏🙏🙏🙏🙏 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏🙏🙏🙏 | 🙏🙏🙏🙏🙏🙏 |
| 124565 | 🙏🙏🙏R.I.P | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏 | 🙏🙏🙏 |

4. Extract Chinese characters and emoji only

```
def filter_text(row):  
    return ''.join(c for c in row if (c in emoji.EMOJI_DATA or  
    re.match(r'[\u4E00-\u9FA5]+', c)))
```

| | text | postUrl | is_post | emojis | text_cleaned |
|--------|--|--|---------|--------|--|
| 0 | /\n\nBringing back the trendy rectangular eyew... | https://www.instagram.com/p/CugrZRvrsSI/ | True | 👁️ | 👁️ |
| 4 | 【#熱貓生活】🐱🐱\n日本一位貓奴分享了自家貓咪 (Goma) 絕育後的趣事。她發現Goma的身... 【恭喜中大同學麥世霆成香港首位奪世青賽金牌泳手 | https://www.instagram.com/p/Cxc2EHUrbuQ/ | True | 🐱🐱🐱🐱🐱🐱 | 熱貓生活🐱🐱日本一位貓奴分享了自家貓咪絕育後的趣 事她發現的身體狀況正常但就會不停發出汪汪的叫... 恭喜中大同學麥世霆成香港首位奪世青賽金牌泳手🏆 🏆中大健康與體育運動科學一年級學生麥世霆日... 應該就冇可能不管都差唔多直上直落咁解🐱貓插畫 |
| | | | | | →半價搶🐱機上護膚 |
| | | | | | ... |
| | | | | | 太慘所以應該要有安樂死合法化🐱 |
| | | | | | 🐱🐱🐱🐱🐱🐱 |
| 124563 | 🐱🐱🐱🐱🐱🐱 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🐱🐱🐱🐱🐱🐱 | 🐱🐱🐱🐱🐱🐱 |
| 124565 | 🐱🐱🐱R.I.P | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🐱🐱🐱 | 🐱🐱🐱 |

5. Filter out null data

| | text | postUrl | is_post | emojis | text_cleaned |
|--------|--|--|---------|----------|--|
| 0 | /\n\nBringing back the trendy rectangular eyew... | https://www.instagram.com/p/CugrZRvrsSI/ | True | 👁️ | 👁️ |
| 4 | 【#熱貓生活】🔥🐱\n日本一位貓奴分享了自家貓咪 (Goma) 絕育後的趣事。她發現Goma的身... | https://www.instagram.com/p/Cxc2EHUrbuQ/ | True | 🔥🐱😄🗣️ | 熱貓生活🔥🐱日本一位貓奴分享了自家貓咪絕育後的趣事她發現的身體狀況正常但就會不停發出汪汪的叫... |
| 5 | 【恭喜中大同學麥世霆成香港首位奪世青賽金牌男泳手! 🏆🏊‍♂️】\n#中大健康與體育運動科學 一年... | https://www.instagram.com/p/Cw7V6EltXKX/ | True | 🏆🏊‍♂️🏠👦👦 | 恭喜中大同學麥世霆成香港首位奪世青賽金牌男泳手🏆🏊‍♂️ 🏠中大健康與體育運動科學一年級學生麥世霆日... |
| 6 | 應該就冇可能\n\n不管都差唔多\n直上直落咁解 🐱\n\n#貓 #插畫 | https://www.instagram.com/p/Cj5Q9BeSAIO/ | True | 🐱👦 | 應該就冇可能不管都差唔多直上直落咁解🐱👦貓插畫 |
| 7 | My skin care routine for long flights ✈️ \n50%... | https://www.instagram.com/p/CxSev1HP5rA/ | True | ✈️👁️ | ✈️半價搶👁️機上護膚 |
| ... | ... | ... | ... | ... | ... |
| 124559 | 🙏🙏🙏 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏 | 🙏🙏🙏 |
| 124561 | 太慘，所以應該要有安樂死合法化😞 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 😞 | 太慘所以應該要有安樂死合法化😞 |
| 124562 | 🙏🙏all the best 😞 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏😞 | 🙏🙏😞 |
| 124563 | 🙏🙏🙏🙏🙏🙏 | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏🙏🙏🙏 | 🙏🙏🙏🙏🙏🙏 |
| 124565 | 🙏🙏🙏R.I.P | https://www.instagram.com/p/CymmM1zu2-6/ | False | 🙏🙏🙏 | 🙏🙏🙏 |

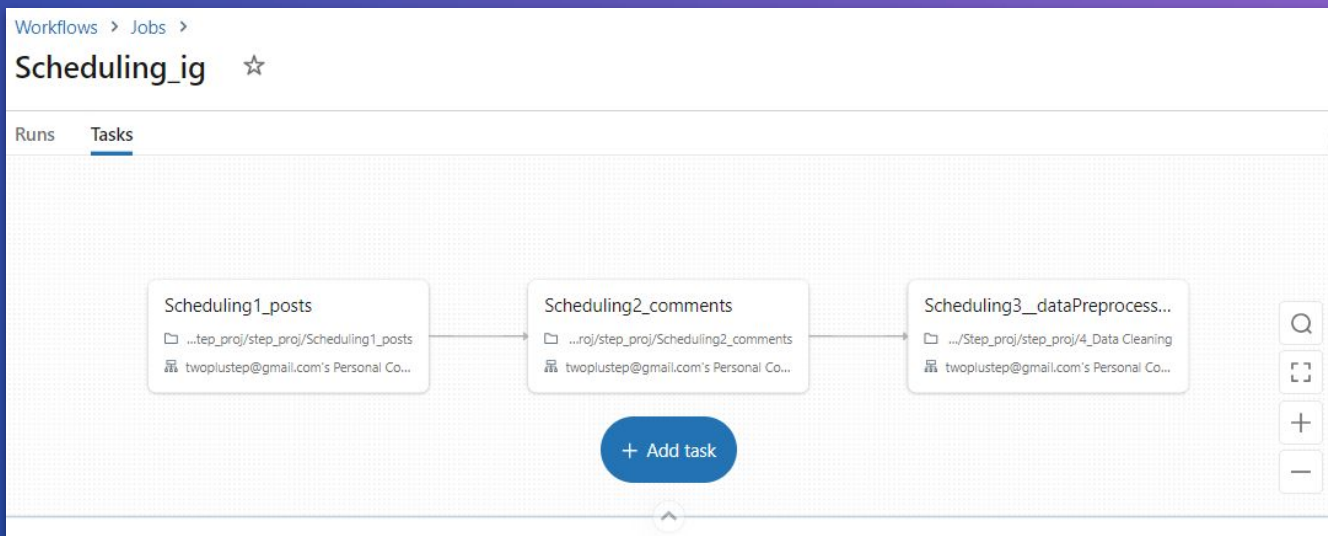


04

Scheduling



Set up scheduler



Schedule

At 0 minutes past the hour, every 4 hours (UTC+08:00 — undefined)

Edit schedule

Pause

Delete

Schedule:
6 times a day

Adjustments:
1. Re-select ig accounts

2. Reduce resultLimit to 90 posts per account and 30 comments per post

3. Identify new posts and Comments

4. Create Schedule Logs to Monitor the condition of Tables

5. Follow-up for failure jobs
- Using Pickle to store Python objects



05

Data Analytics



Analysis

- Most frequently used emojis
- Most intense emojis



- Sentiment Analysis on the use of emojis
- Basket Analysis on the combination of emojis

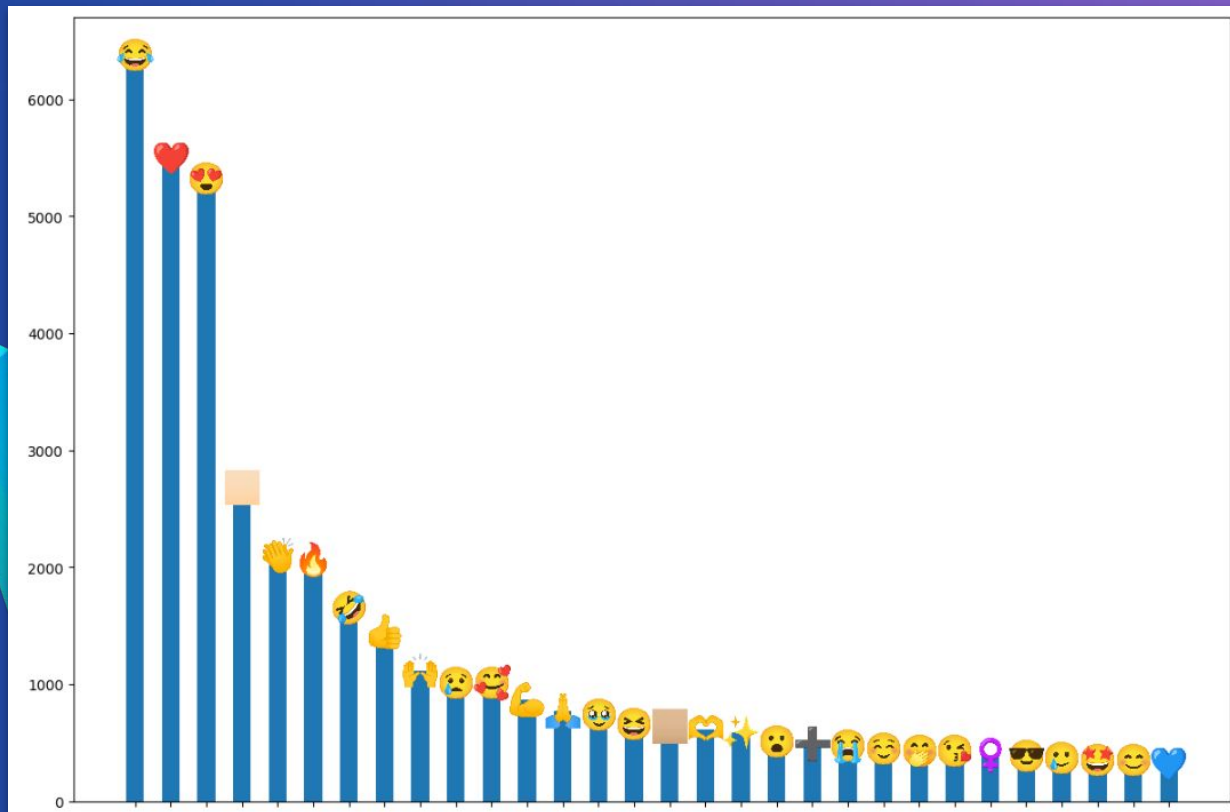


06

Data Visualization



Top 30 Emojis used on Instagram



| Emoji | Demojize | Count | % |
|-------|----------------------------------|-------|-----------|
| 😄 | :face_with_tears_of_joy: | 6373 | 15.991268 |
| ❤️ | :red_heart: | 5493 | 13.783153 |
| 😍 | :smiling_face_with_heart-eyes: | 5309 | 13.321456 |
| 👤 | :light_skin_tone: | 2677 | 6.717186 |
| 👏 | :clapping_hands: | 2092 | 5.249291 |
| 🔥 | :fire: | 2059 | 5.166487 |
| 🤣 | :rolling_on_the_floor_laughing: | 1640 | 4.115123 |
| 👍 | :thumbs_up: | 1445 | 3.625825 |
| 🙌 | :raising_hands: | 1112 | 2.790254 |
| 😭 | :crying_face: | 1000 | 2.509221 |
| 💪 | :smiling_face_with_hearts: | 999 | 2.506712 |
| 🙏 | :flexed_biceps: | 863 | 2.165458 |
| 🤔 | :folded_hands: | 766 | 1.922064 |
| 😓 | :face_holding_back_tears: | 729 | 1.829222 |
| 😏 | :grinning_squinting_face: | 645 | 1.618448 |
| 👤 | :medium-light_skin_tone: | 627 | 1.573282 |
| 🙌 | :heart_hands: | 611 | 1.533134 |
| ✨ | :sparkles: | 581 | 1.457858 |
| 😮 | :face_with_open_mouth: | 495 | 1.242065 |
| + | :plus: | 479 | 1.201917 |
| 🤯 | :loudly_crying_face: | 463 | 1.161770 |
| 😊 | :smiling_face: | 436 | 1.094021 |
| 🤔 | :face_with_hand_over_mouth: | 423 | 1.061401 |
| 💋 | :face_blowing_a_kiss: | 417 | 1.046345 |
| ♀️ | :female_sign: | 388 | 0.973578 |
| 🕶️ | :smiling_face_with_sunglasses: | 374 | 0.938449 |
| 😏 | :smiling_face_with_tear: | 356 | 0.893283 |
| 🌟 | :star-struck: | 345 | 0.865681 |
| 😏 | :smiling_face_with_smiling_eyes: | 343 | 0.860663 |
| 💙 | :blue_heart: | 313 | 0.785386 |

Sentiment Analysis

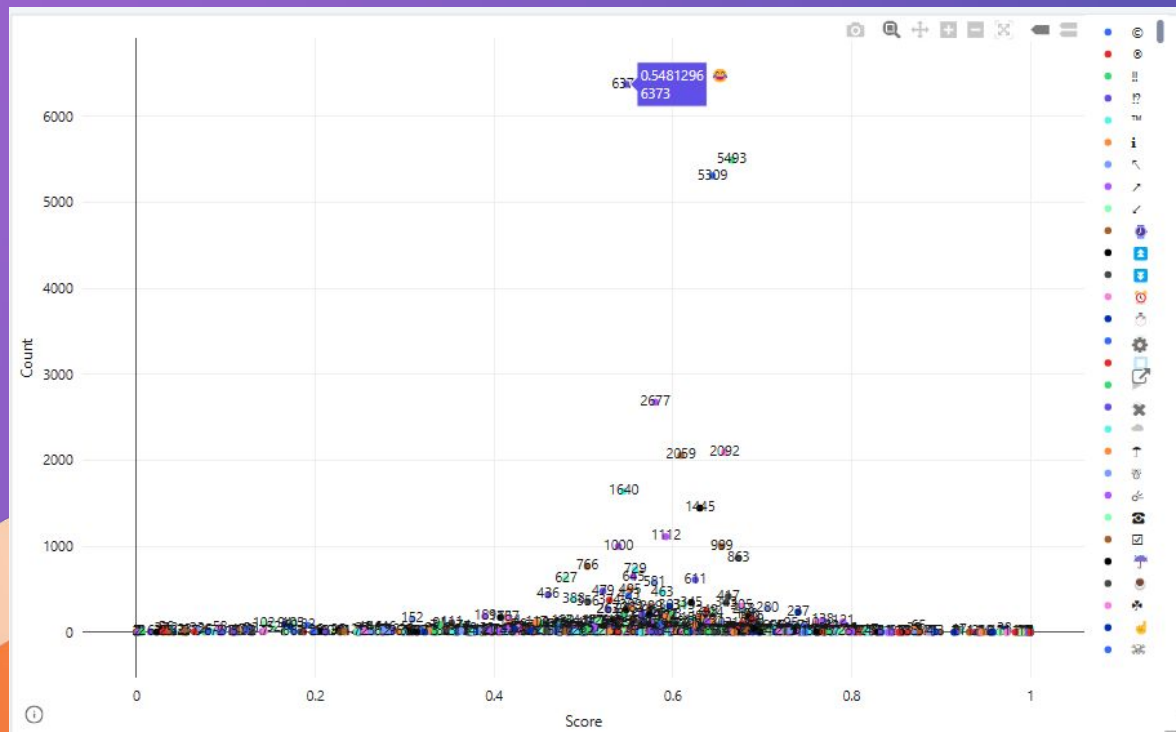
from snownlp import SnowNLP - Support Sentiment Analysis on Chinese Characters

| | text | postUrl | is_post | emojis | text_cleaned | text_only | sentiment_score |
|--------|--|--|---------|--------|---|---|-----------------|
| 4 | 【#熱貓生活】🔥🐱🇯🇵日本一位貓奴分享了自家貓咪 (Goma) 絕育後的趣事。她發現Goma的身... | https://www.instagram.com/p/Cxc2EHUrbuQ/ | True | 🔥🐱🇯🇵 | 熱貓生活🔥🐱日本一位貓奴分享了自家貓咪絕育後的趣事她發現的身體狀況正常但就會不停發出汪汪的叫聲... | 熱貓生活日本一位貓奴分享了自家貓咪絕育後的趣事她發現的身體狀況正常但就會不停發出汪汪的叫聲... | 7.086497e-02 |
| 5 | 【恭喜中大同學麥世堃成香港首位奪世青賽金牌男泳手！🏆🏊🇺🇸】🇺🇸中大健康與體育運動科學一年級學生麥世堃日前代表... | https://www.instagram.com/p/Cw7V6EltXKX/ | True | 🏆🏊🇺🇸 | 恭喜中大同學麥世堃成香港首位奪世青賽金牌男泳手🏆🏊中大健康與體育運動科學一年級學生麥世堃日前代表... | 恭喜中大同學麥世堃成香港首位奪世青賽金牌男泳手中大健康與體育運動科學一年級學生麥世堃日前代表... | 1.000000e+00 |
| 6 | 應該就方可能🇺🇸🇺🇸不管都差唔多🇺🇸直上直落咁解🇺🇸🇺🇸#貓畫 | https://www.instagram.com/p/Cj5Q9BeSAIO/ | True | 🇺🇸🇺🇸 | 應該就方可能不管都差唔多直上直落咁解🇺🇸🇺🇸貓畫 | 應該就方可能不管都差唔多直上直落咁解貓畫 | 2.481876e-01 |
| 7 | My skin care routine for long flights 🇺🇸🇺🇸50%... | https://www.instagram.com/p/CxSev1HP5rA/ | True | 🇺🇸🇺🇸 | 🇺🇸半價搶🇺🇸機上護膚 | 半價搶機上護膚 | 4.572158e-02 |
| 8 | 每個女人都應該有一個 Precious Moment 比自己，而我作為一個working m... | https://www.instagram.com/p/CqE8v5evs-O/ | True | 💖🇺🇸 | 每個女人都應該有一個比自己而我作為一個都一樣每日都會為自己嘅至臻凝眸系列幫到我特別係裡面嘅... | 每個女人都應該有一個比自己而我作為一個都一樣每日都會為自己嘅至臻凝眸系列幫到我特別係裡面嘅... | 2.273359e-11 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 124594 | 香水🇺🇸 | https://www.instagram.com/p/CynBqtmp5LK/ | False | 🇺🇸 | 香水🇺🇸 | 香水 | 3.571429e-01 |
| 124602 | 點解無人係圈內買股票，係有原因的🇺🇸 | https://www.instagram.com/p/CymfU6Zu0l1/ | False | 🇺🇸 | 點解無人係圈內買股票係有原因的🇺🇸 | 點解無人係圈內買股票係有原因的 | 9.756941e-01 |
| 124605 | 第一眼睇左陰江市🇺🇸...真係陰乾左 | https://www.instagram.com/p/CymfU6Zu0l1/ | False | 🇺🇸 | 第一眼睇左陰江市🇺🇸真係陰乾左 | 第一眼睇左陰江市真係陰乾左 | 7.606363e-01 |
| 124612 | 好無私🇺🇸🇺🇸🇺🇸 | https://www.instagram.com/p/CymnM1zu2-6/ | False | 🇺🇸🇺🇸🇺🇸 | 好無私🇺🇸🇺🇸🇺🇸 | 好無私 | 8.607041e-01 |
| 124618 | 太慘，所以應該要有安樂死合法化🇺🇸 | https://www.instagram.com/p/CymnM1zu2-6/ | False | 🇺🇸 | 太慘所以應該要有安樂死合法化🇺🇸 | 太慘所以應該要有安樂死合法化 | 3.513378e-01 |

Improvement: Split the sentence into smaller pieces for each emoji

Sentiment Analysis

Explode the dataframe for each emoji, Group by emoji and Avg()
(Dashboard)



| | Emoji | Score | Count |
|----|-------|--------------------|-------|
| 1 | 😮 | 0.548129623397084 | 6373 |
| 2 | ❤️ | 0.6657225508554004 | 5493 |
| 3 | 😄 | 0.644119639856488 | 5309 |
| 4 | 👉 | 0.5802241639874075 | 2677 |
| 5 | 👉 | 0.6577485683140423 | 2092 |
| 6 | 👉 | 0.6090136006169505 | 2059 |
| 7 | 👉 | 0.5450546577835826 | 1640 |
| 8 | 👉 | 0.6303569053159175 | 1445 |
| 9 | 👉 | 0.5921657793668582 | 1112 |
| 10 | 👉 | 0.5390300068769259 | 1000 |
| 11 | 👉 | 0.6545477421361942 | 999 |
| 12 | 👉 | 0.6734214743030926 | 863 |
| 13 | 👉 | 0.504213021851671 | 766 |
| 14 | 👉 | 0.5577628089643304 | 729 |
| 15 | 👉 | 0.5558975729052281 | 645 |
| 16 | 👉 | 0.4804377142132092 | 627 |
| 17 | 👉 | 0.6251453497825384 | 611 |
| 18 | 👉 | 0.5792716521638435 | 581 |
| 19 | 👉 | 0.552085933913154 | 495 |
| 20 | 👉 | 0.5218224661597759 | 479 |
| 21 | 👉 | 0.5881611643726654 | 463 |
| 22 | 👉 | 0.4604814021201426 | 436 |

Basket Analysis

```
from mlxtend.frequent_patterns import apriori
```

```
from mlxtend.frequent_patterns import association_rules
```

| | antecedents ▲ | consequents ▲ | antecedent_support ▲ | consequent_support ▲ | support ▼ | confidence ▲ | lift |
|----|---------------|---------------|----------------------|----------------------|----------------------|---------------------|-------|
| 1 | 🍌 | ❤️ | 0.08405713061126728 | 0.23104365474554692 | 0.020383980233716136 | 0.24250149970006 | 1.049 |
| 2 | ❤️ | 🍌 | 0.23104365474554692 | 0.08405713061126728 | 0.020383980233716136 | 0.08822566564818855 | 1.049 |
| 3 | 🍌🍌 | 🍌 | 0.18547278984456742 | 0.07867434795214744 | 0.013753198784776936 | 0.07415211037857677 | 0.942 |
| 4 | 🍌 | 🍌🍌 | 0.07867434795214744 | 0.18547278984456742 | 0.013753198784776936 | 0.17481172888960103 | 0.942 |
| 5 | 🍌 | 🍌🍌 | 0.08405713061126728 | 0.18547278984456742 | 0.01294641168832806 | 0.15401919616076784 | 0.830 |
| 6 | 🍌🍌 | 🍌 | 0.18547278984456742 | 0.08405713061126728 | 0.01294641168832806 | 0.06980221572758785 | 0.830 |
| 7 | 🍌 | 🍌 | 0.05269328223681723 | 0.0398729310323093 | 0.011509322172778499 | 0.21842105263157893 | 5.477 |
| 8 | 🍌 | 🍌 | 0.0398729310323093 | 0.05269328223681723 | 0.011509322172778499 | 0.28865001580777744 | 5.477 |
| 9 | 🍌🍌 | ❤️ | 0.033153907244696 | 0.23104365474554692 | 0.009643627012240473 | 0.2908745247148289 | 1.258 |
| 10 | ❤️ | 🍌🍌 | 0.23104365474554692 | 0.033153907244696 | 0.009643627012240473 | 0.04173941510257529 | 1.258 |
| 11 | 🍌 | 🍌 | 0.015114652010034415 | 0.05269328223681723 | 0.00932847580269013 | 0.6171809841534612 | 11.71 |
| 12 | 🍌 | 🍌 | 0.05269328223681723 | 0.015114652010034415 | 0.00932847580269013 | 0.1770334928229665 | 11.71 |
| 13 | 🍌🍌 | ❤️ | 0.036746631033569906 | 0.23104365474554692 | 0.008988112496375761 | 0.24459691252144083 | 1.058 |
| 14 | ❤️ | 🍌🍌 | 0.23104365474554692 | 0.036746631033569906 | 0.008988112496375761 | 0.0389022261021388 | 1.058 |
| 15 | 🍌 | 🍌 | 0.08405713061126728 | 0.07867434795214744 | 0.00833259798051105 | 0.09913017396520697 | 1.260 |
| 16 | 🍌 | 🍌 | 0.07867434795214744 | 0.08405713061126728 | 0.00833259798051105 | 0.10591251402018909 | 1.260 |
| 17 | 🍌 | 🍌 | 0.05269328223681723 | 0.01947634475021115 | 0.006996356852017598 | 0.13277511961722488 | 6.817 |
| 18 | 🍌 | 🍌 | 0.01947634475021115 | 0.05269328223681723 | 0.006996356852017598 | 0.35922330097087385 | 6.817 |
| 19 | 🍌 | 🍌 | 0.0398729310323093 | 0.07867434795214744 | 0.005924842739546434 | 0.14859310780904206 | 1.888 |
| 20 | 🍌 | 🍌 | 0.07867434795214744 | 0.0398729310323093 | 0.005924842739546434 | 0.07530844415958982 | 1.888 |
| 21 | 🍌 | 🍌🍌 | 0.05269328223681723 | 0.01708119555762855 | 0.00563490362676012 | 0.1069377990430622 | 6.260 |
| 22 | 🍌🍌 | 🍌 | 0.01708119555762855 | 0.05269328223681723 | 0.00563490362676012 | 0.3298892988929889 | 6.260 |