Nama : Dewi Fajar Nawulansih

NIM : 211101025

Kelas : TI7A

PENGAMBILAN INFORMASI SCRAPE DATA

1. Download and Install Google Play Scraper Package

pip install google-play-scraper

Output :

Collecting google-play-scraper

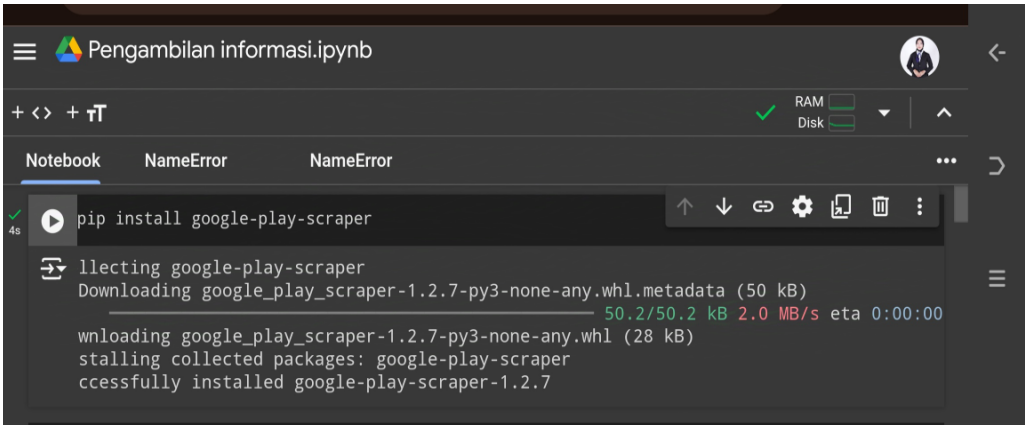
Downloading google_play_scraper-1.2.7-py3-none-any.whl.metadata (50 kB)

━━ 50.2/50.2 kB 2.0 MB/s eta 0:00:00

Downloading google_play_scraper-1.2.7-py3-none-any.whl (28 kB)

Installing collected packages: google-play-scraper

Successfully installed google-play-scraper-1.2.7

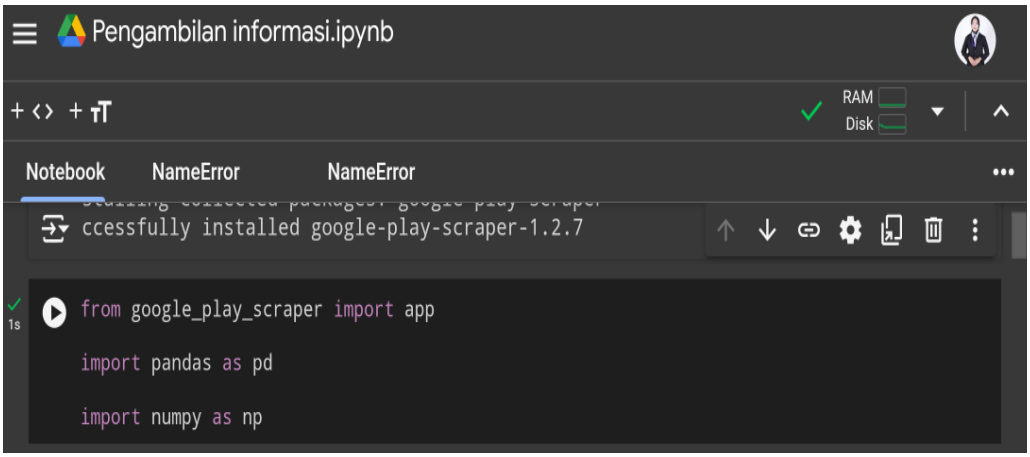


2. Import required packages

from google_play_scraper import app

import pandas as pd

import numpy as np



3. Find the App Id in Google Play Store and scrape

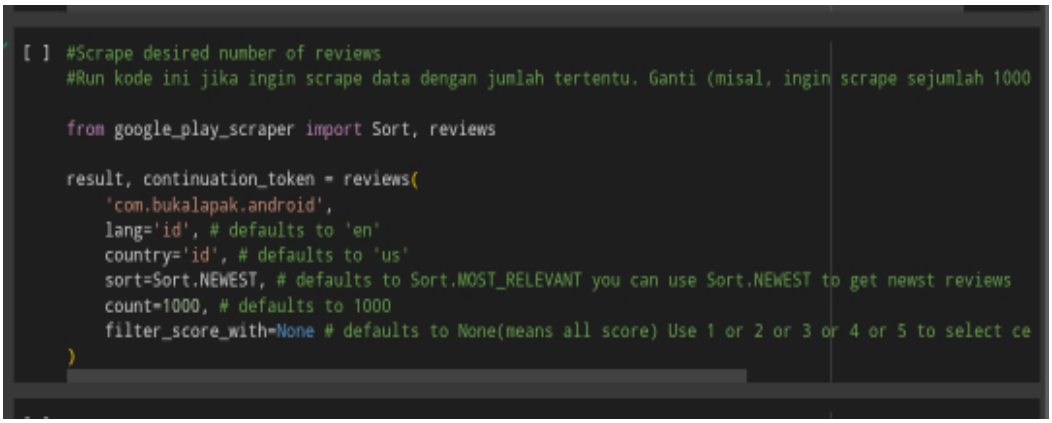
Bisa dilihat dari URL yang akan di ambil datanya, karena saya akan ambil dari Bukalapak maka

'com.bukalapak.android' .

#Scrape desired number of reviews

#Run kode ini jika ingin scrape data dengan jumlah tertentu. Ganti (misal, ingin scrape sejumlah 1000, maka ganti kode , count = 1000)

```
from google_play_scraper import Sort, reviews

result, continuation_token = reviews(
    'com.bukalapak.android',
    lang='id', # defaults to 'en'
    country='id', # defaults to 'us'
    sort=Sort.NEWEST, # defaults to Sort.MOST_RELEVANT you can use Sort.NEWEST to get newst reviews
    count=1000, # defaults to 1000
    filter_score_with=None # defaults to None(means all score) Use 1 or 2 or 3 or 4 or 5 to select certain score
)
```

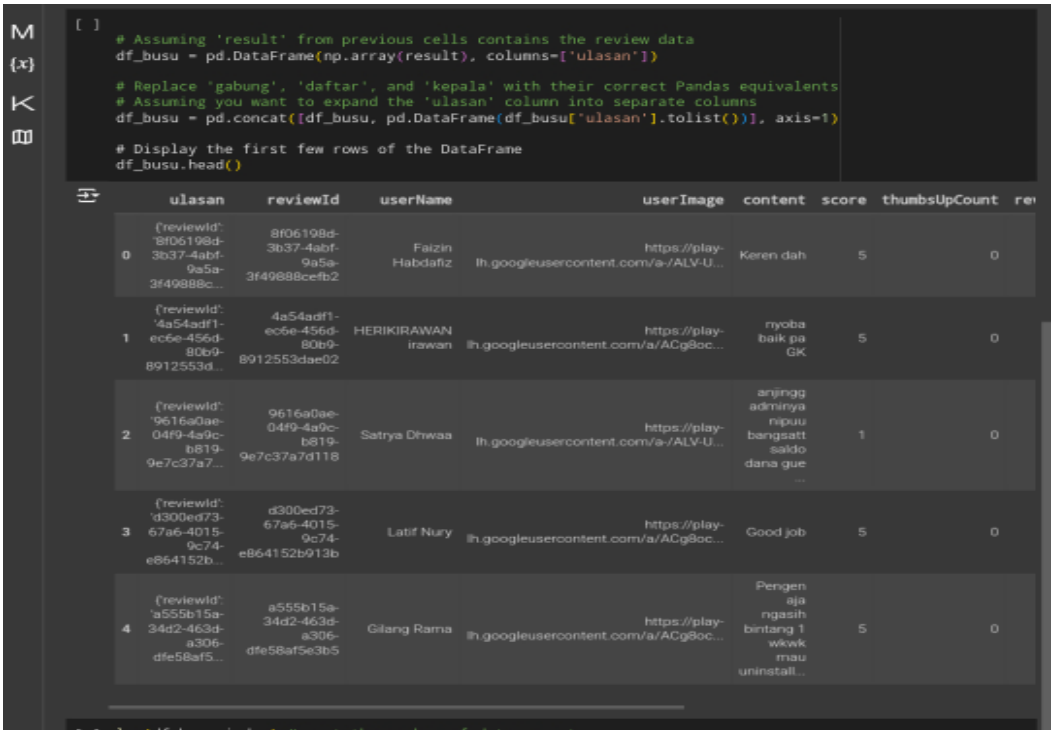


```
[ ] #Scrape desired number of reviews
    #Run kode ini jika ingin scrape data dengan jumlah tertentu. Ganti (misal, ingin scrape sejumlah 1000

    from google_play_scraper import Sort, reviews

    result, continuation_token = reviews(
        'com.bukalapak.android',
        lang='id', # defaults to 'en'
        country='id', # defaults to 'us'
        sort=Sort.NEWEST, # defaults to Sort.MOST_RELEVANT you can use Sort.NEWEST to get newst reviews
        count=1000, # defaults to 1000
        filter_score_with=None # defaults to None(means all score) Use 1 or 2 or 3 or 4 or 5 to select ce
    )
```

4. Put the Reviews into Pandas DataFrame

```
# Assuming 'result' from previous cells contains the review data

df_busu = pd.DataFrame(np.array(result), columns=['ulasan'])


# Replace 'gabung', 'daftar', and 'kepala' with their correct Pandas equivalents

# Assuming you want to expand the 'ulasan' column into separate columns

df_busu = pd.concat([df_busu, pd.DataFrame(df_busu['ulasan'].tolist())], axis=1)


# Display the first few rows of the DataFrame

df_busu.head()
```

Output :



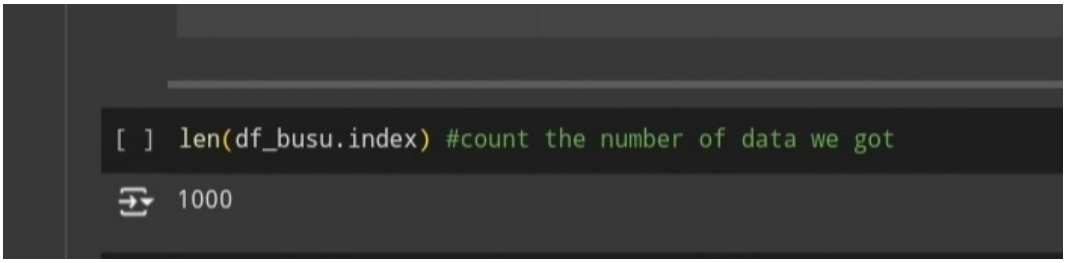
```
[ ]
# Assuming 'result' from previous cells contains the review data
df_busu = pd.DataFrame(np.array(result), columns=['ulasan'])

# Replace 'gabung', 'daftar', and 'kepala' with their correct Pandas equivalents
# Assuming you want to expand the 'ulasan' column into separate columns
df_busu = pd.concat([df_busu, pd.DataFrame(df_busu['ulasan'].tolist())], axis=1)

# Display the first few rows of the DataFrame
df_busu.head()
```

	ulasan	reviewId	userName	userImage	content	score	thumbsUpCount	re
0	{'reviewId': '8f06198d-3b37-4abf-9a5a-3f49888c...	8f06198d-3b37-4abf-9a5a-3f49888cefb2	Faizin Habdafiz	https://play-lh.googleusercontent.com/a-/ALV-U...	Keren dah	5	0	
1	{'reviewId': '4a54adf1-ec6e-456d-80b9-8912553d...	4a54adf1-ec6e-456d-80b9-8912553dae02	HERIKIRAWAN irawan	https://play-lh.googleusercontent.com/a/ACg8oc...	nyoba baik pa GK	5	0	
2	{'reviewId': '9616a0ae-04f9-4a9c-b819-9e7c37a7...	9616a0ae-04f9-4a9c-b819-9e7c37a7d118	Satrya Dhwaa	https://play-lh.googleusercontent.com/a-/ALV-U...	anjingg adminya nipuu bangsatt saldo dana gue ...	1	0	
3	{'reviewId': 'd300ed73-67a6-4015-9c74-e864152b...	d300ed73-67a6-4015-9c74-e864152b913b	Latif Nury	https://play-lh.googleusercontent.com/a/ACg8oc...	Good job	5	0	
4	{'reviewId': 'a555b15a-34d2-463d-a306-dfe58af5...	a555b15a-34d2-463d-a306-dfe58af5e3b5	Gilang Rama	https://play-lh.googleusercontent.com/a/ACg8oc...	Pengen aja ngasih bintang 1 wkwk mau uninstall...	5	0	

5. Jumlah data yang akan didapat

len(df_busu.index) #hitung jumlah data yang kita dapatkan

Output: 1000

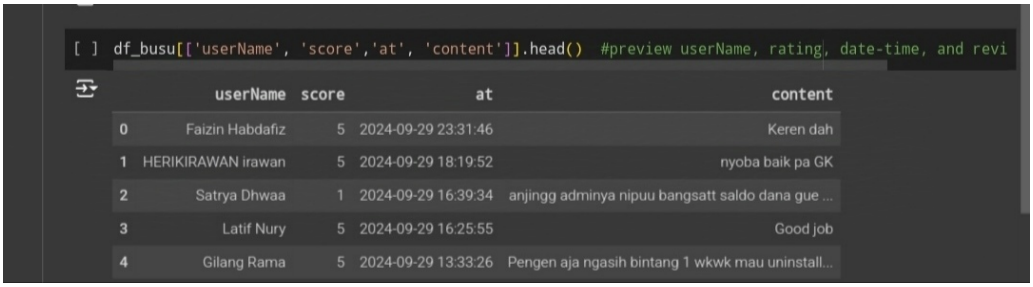


```
[ ] len(df_busu.index) #count the number of data we got
1000
```

6. Filter variabel

df_busu[['userName', 'score','at', 'content']].head() #preview userName, rating, date-time, and reviews only

Output:



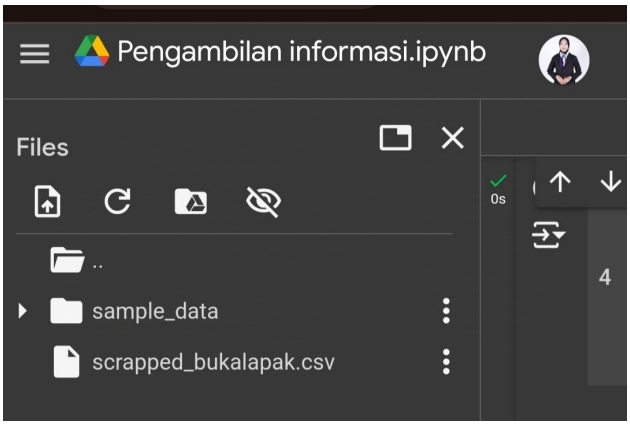
```
[ ] df_busu[['userName', 'score','at', 'content']].head()  #preview userName, rating, date-time, and revi
```

	userName	score	at	content
0	Faizin Habdafiz	5	2024-09-29 23:31:46	Keren dah
1	HERIKIRAWAN irawan	5	2024-09-29 18:19:52	nyoba baik pa GK
2	Satrya Dhwaa	1	2024-09-29 16:39:34	anjingg adminya nipuu bangsatt saldo dana gue ...
3	Latif Nury	5	2024-09-29 16:25:55	Good job
4	Gilang Rama	5	2024-09-29 13:33:26	Pengen aja ngasih bintang 1 wkwk mau uninstall...

6. Simpan Variabel dan download data

my_df = sorted_df[['userName', 'score','at', 'content']] #get userName, rating, date-time, and reviews only

my_df.to_csv ("scrapped_bukalapak.csv" , index = False) #Simpan file sebagai , untuk mengunduh: klik ikon folder di sebelah kiri. file csv seharusnya ada di sana.



Link google colab

PI/211101025/DewiFajarN