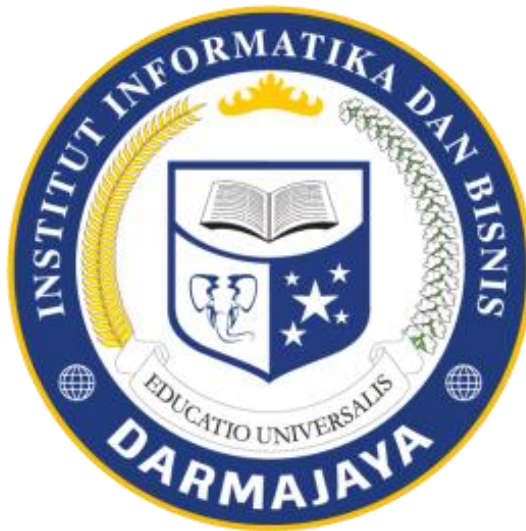


**ANALISIS BIG DATA
KONEKSI GOOGEL DRIVE, GITHUB DAN
CRAWLING**



**DISUSUN OLEH:
FIQQI AHLUDZIKRI (2111010034)**

**DOSEN:
Dr. M. SAID HASIBUAN**

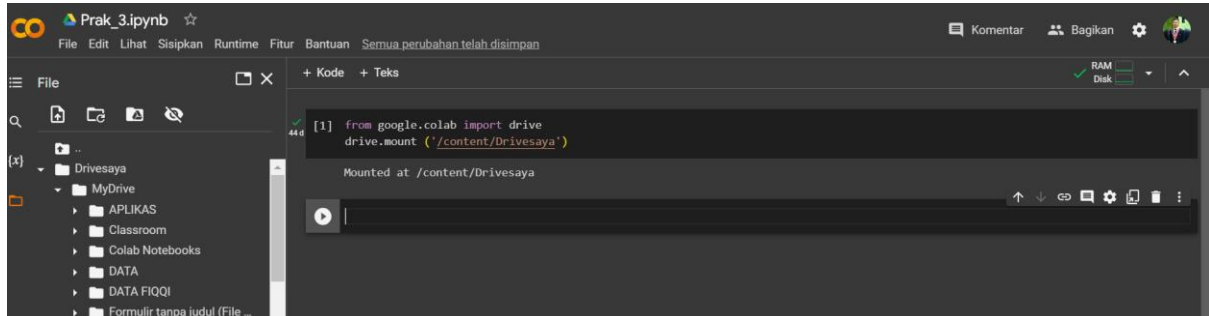
**FAKULTAS ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA
INSTITUT INFORMATIKA DAN BISNIS DARMAJAYA
BANDAR LAMPUNG
TAHUN AJARAN 2023/2024**

Menghubungkan Google Colab dengan Drive

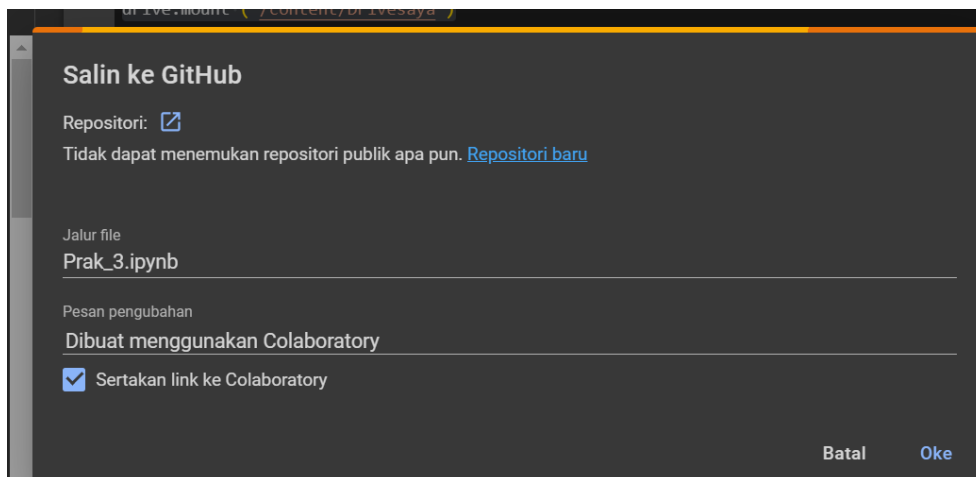
Code:

```
from google.colab import drive
drive.mount ('/content/Drivesaya')
```

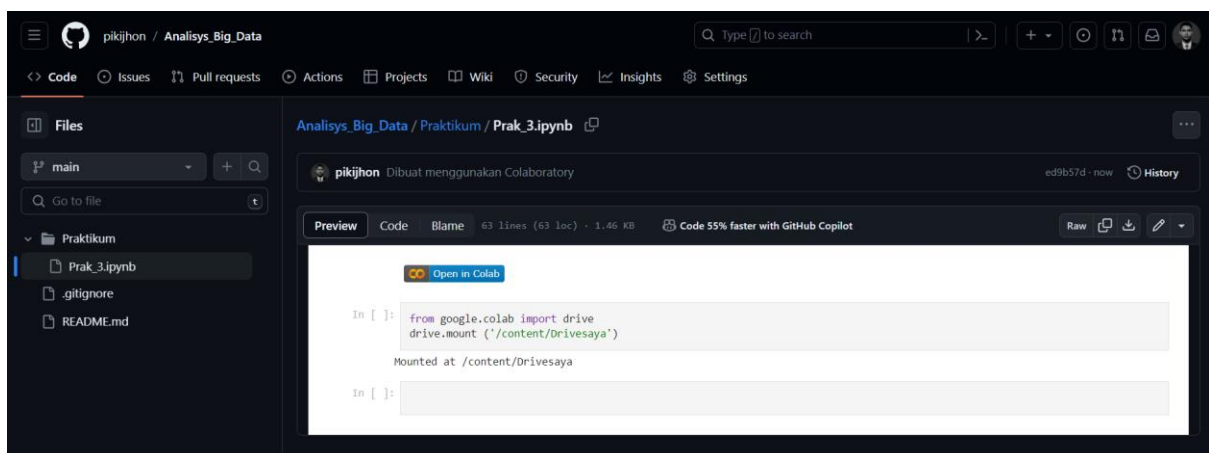
Hasil:



Menghubungkan Google Colab dengan Github



Simpan Salinan ke Github di repository



CRAWLING DATA TWITTER

Langkah 1:

Install Pandas & Instal Node.js (karena tweet-harvest dibuat menggunakan Node.js)

Perintah ini menginstal pustaka Pandas menggunakan pip, yang merupakan manajer paket Python.

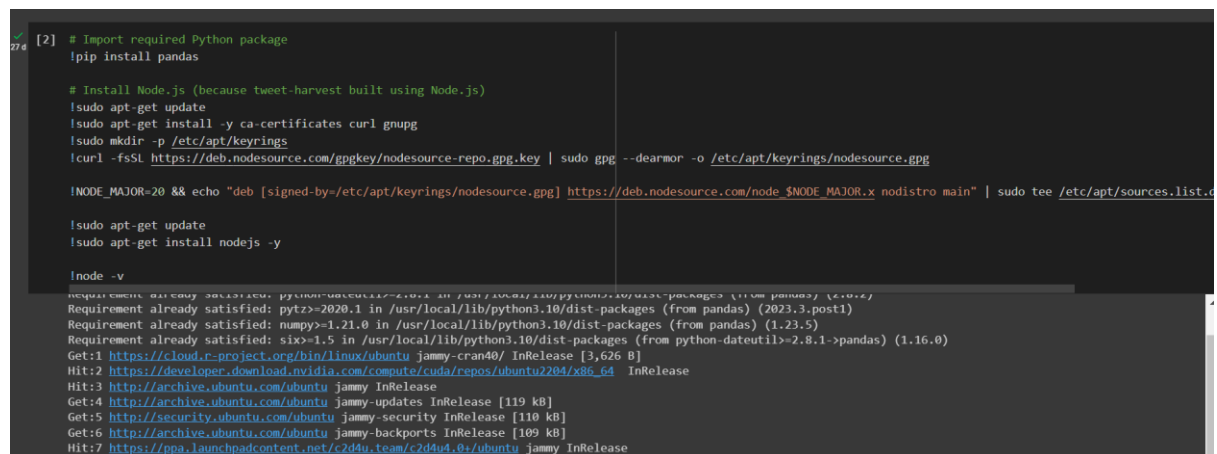
```
!pip install pandas
```

Perintah yang menginstal Node.js, yaitu runtime JavaScript server-side.

```
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --
dearmor -o /etc/apt/keyrings/nodesource.gpg
!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg]
https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" | sudo tee
/etc/apt/sources.list.d/nodesource.list
!sudo apt-get update
!sudo apt-get install nodejs -y
```

Untuk memastikan nodejs telah terinstall menggunakan perintah berikut

```
!node -v
```



```
[2] # Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg

!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" | sudo tee /etc/apt/sources.list.d/nodesource.list

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v
node v20.11.0
```

Langkah 2:

Mengumpulkan tweet dari Twitter menggunakan alat bernama "tweet-harvest" dengan menggunakan Node.js (melalui perintah npx).

1. **filename**: Variabel ini adalah nama file yang akan digunakan untuk menyimpan hasil pengumpulan data.
2. **search_keyword**: Variabel ini digunakan untuk menentukan kata kunci pencarian di Twitter.
3. **limit**: Variabel ini menentukan batasan jumlah tweet yang akan diambil.
4. **--token** : Token akses Twitter.

```
selecting previously unselected package nodejs.
[3] # Crawl Data
filename = 'PDIP.csv'
search_keyword = 'PDIP min_faves:100'
limit = 50

lnpx --yes tweet-harvest@latest -o "{filename}" -s "{search_keyword}" -l {limit} --token ""

Welcome to the Twitter Crawler

This script uses Chromium Browser to crawl data from Twitter with *your* Twitter auth token.
Please enter your Twitter auth token when prompted.

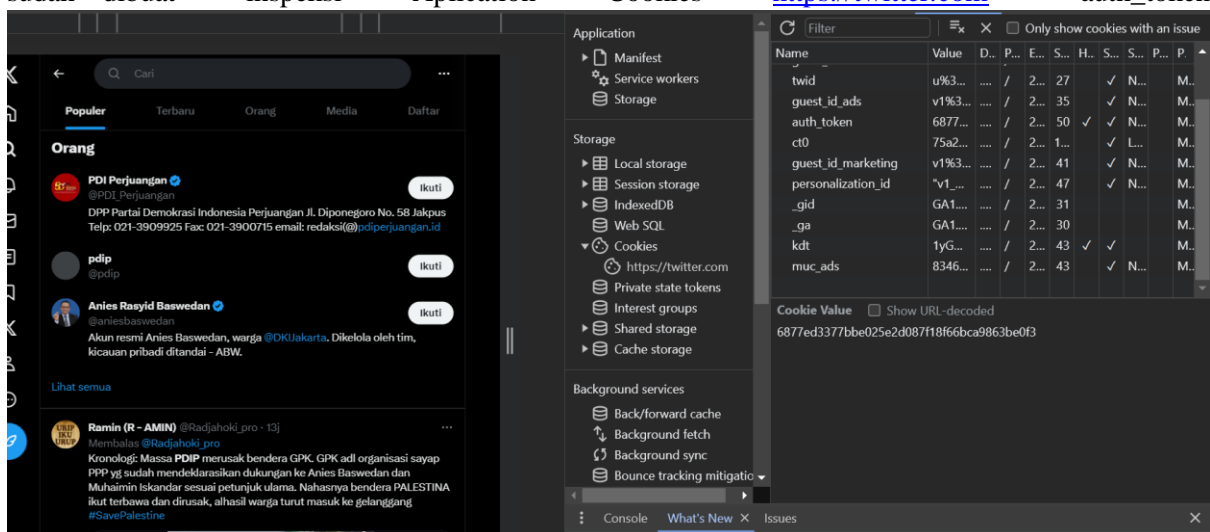
Note: Keep your access token secret! Don't share it with anyone else.
Note: This script only runs on your local device.

? What's your Twitter auth token? : 8788? What's your Twitter auth token? : *8788? What's your Twitter auth token? : **8788? What's your Twitter auth token? : ***8788

added 3 packages in 2s
Installing dependencies...
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:2 https://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64 InRelease
Hit:4 https://deb.nodesource.com/node_20.x nodistro InRelease
Hit:5 https://archive.ubuntu.com/ubuntu jammy InRelease
Hit:6 https://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:7 https://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:8 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u.0/ubuntu jammy InRelease
```

Langkah 3:

1. Untuk mendapatkan token akses ke twitter: Buka twitter.com kemudian login dengan akun yang sudah dibuat – inspeksi – Application – Cookies – <https://twitter.com> - auth_token



Langkah 4:

Membaca data yang telah diambil dari Twitter menggunakan "tweet-harvest" dari file CSV yang telah dihasilkan sebelumnya, dan kemudian menampilkannya dalam bentuk DataFrame menggunakan Pandas.

1. **import pandas as pd**: Pandas digunakan untuk analisis dan manipulasi data dalam Python.
2. **file_path = f'tweets-data/{filename}'**: Variabel **file_path** digunakan untuk menentukan jalur file CSV yang akan dibaca. Nilai dari variabel ini adalah hasil dari menggabungkan direktori 'tweets-data/' dengan nama file yang telah ditentukan sebelumnya dalam variabel **filename**.
3. **df = pd.read_csv(file_path, delimiter=";")**: Perintah untuk membaca file CSV ke dalam sebuah DataFrame menggunakan Pandas. **pd.read_csv** digunakan untuk membaca data dari file CSV. Argumen **file_path** adalah jalur ke file CSV, dan **delimiter=";"** mengindikasikan bahwa pemisah kolom dalam file CSV adalah titik koma (;). Data dari file CSV akan dimuat ke dalam DataFrame yang disimpan dalam variabel **df**.

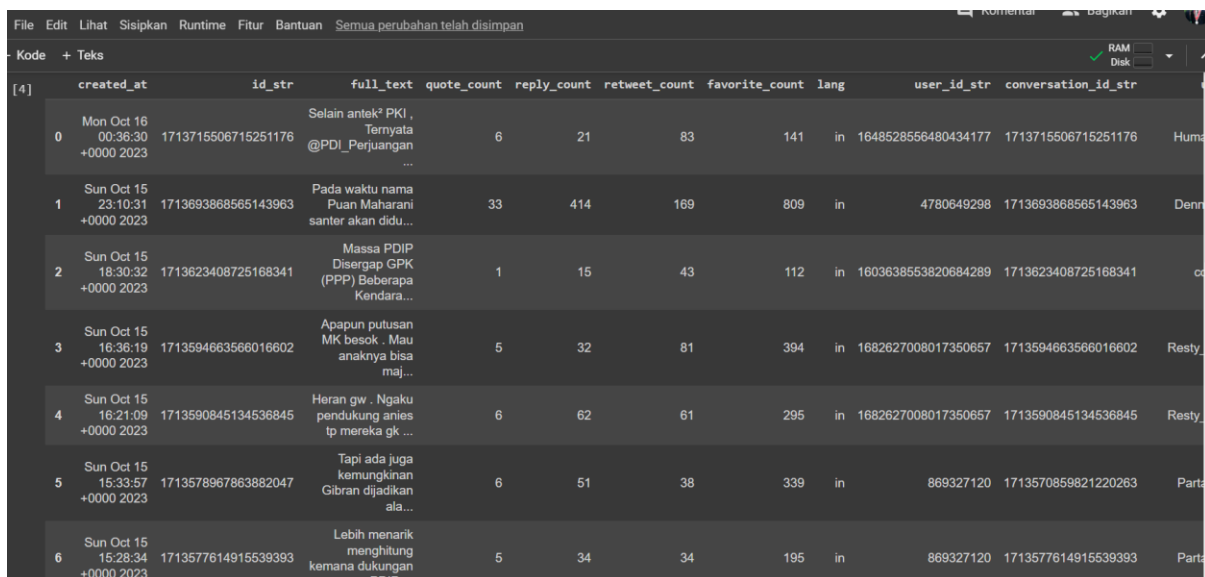
4. **display(df)**: Perintah yang digunakan untuk menampilkan DataFrame **df**. Menampilkan data dalam bentuk tabel, sehingga Anda dapat melihat isi dari data yang telah dibaca.

```
import pandas as pd

# Specify the path to your CSV file
file_path = f"tweets-data/{filename}"

# Read the CSV file into a pandas DataFrame
df = pd.read_csv(file_path, delimiter=";")

# Display the DataFrame
display(df)
```



	created_at	id_str	full_text	quote_count	reply_count	retweet_count	favorite_count	lang	user_id_str	conversation_id_str
0	Mon Oct 16 00:38:30 +0000 2023	1713715506715251176	Selain antek? PKI , Ternyata @PDI_Perjuangan ...	6	21	83	141	in	1648528556480434177	1713715506715251176
1	Sun Oct 15 23:10:31 +0000 2023	1713693868565143963	Pada waktu nama Puan Maharani santer akan didu...	33	414	169	809	in	4780649298	1713693868565143963
2	Sun Oct 15 18:30:32 +0000 2023	1713623408725168341	Massa PDIP Disergap GPK (PPP) Beberapa Kendara...	1	15	43	112	in	1603638553820684289	1713623408725168341
3	Sun Oct 15 16:38:19 +0000 2023	1713594663566016602	Apapun putusan MK besok . Mau anaknya bisa maj...	5	32	81	394	in	1682627008017350657	1713594663566016602
4	Sun Oct 15 16:21:09 +0000 2023	1713590845134536845	Heran gw . Ngaku pendukung anies tp mereka gk ...	6	62	61	295	in	1682627008017350657	1713590845134536845
5	Sun Oct 15 15:33:57 +0000 2023	1713578967863882047	Tapi ada juga kemungkinan Gibran dijadikan ala...	6	51	38	339	in	869327120	1713570859821220263
6	Sun Oct 15 15:28:34 +0000 2023	1713577614915539393	Lebih menarik menghitung kemana dukungan PDIP	5	34	34	195	in	869327120	1713577614915539393

Langak 4:

Menghitung jumlah tweet yang ada dalam DataFrame yang telah dibaca sebelumnya menggunakan Pandas dan menampilkannya sebagai output.

1. `num_tweets = len(df)`: Baris ini menghitung jumlah baris (entri) dalam DataFrame `df` menggunakan fungsi `len()`. Setiap baris dalam DataFrame mewakili satu tweet.
2. `print(f"Jumlah tweet dalam dataframe adalah {num_tweets}.")`: Baris ini mencetak jumlah tweet yang telah dihitung sebelumnya ke layar. Ini digunakan untuk memberikan informasi kepada pengguna tentang berapa banyak tweet yang ada dalam DataFrame.

```
[5] # Cek jumlah data yang didapatkan

num_tweets = len(df)
print(f"Jumlah tweet dalam dataframe adalah {num_tweets}.")

Jumlah tweet dalam dataframe adalah 58.
```