

Tugas Pertemuan 3
Koneksi google colab ke Gdrive dan GitHub, Crawling

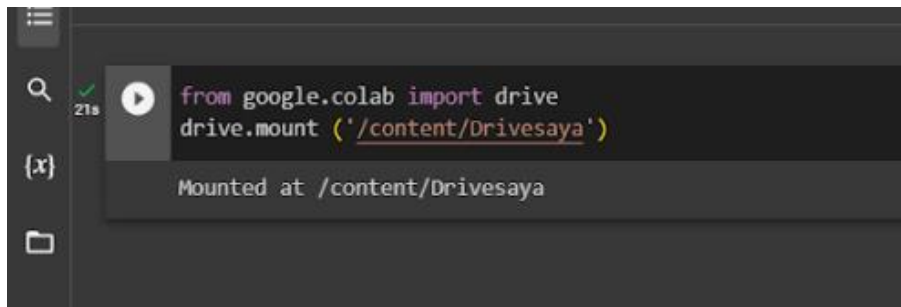


Disusun Oleh:
Latifah Nabilah (2011010045)

Dosen Pengampu:
Dr. Muhammad Said Hasibuan S.Kom., M.Kom

Program Studi Teknik Informatika Fakultas Ilmu Komputer
Institut Informatika dan Bisnis Darmajaya 2023

Menghubungkan google colab dengan google drive

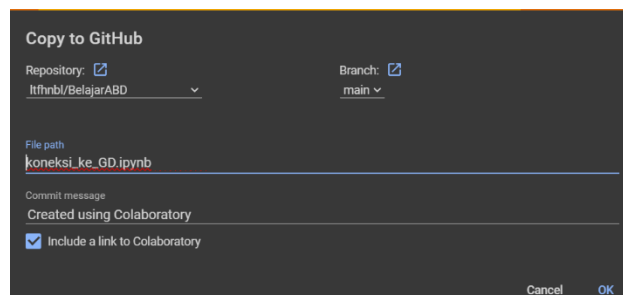


```
from google.colab import drive
drive.mount('/content/Drivesaya')

Mounted at /content/Drivesaya
```

1. **from google.colab import drive:** Mengimpor modul drive yang disediakan oleh Google Colab. Modul ini untuk terhubung ke Google Drive dari Colab.
2. **drive.mount('/content/Drivesaya'):** Perintah yang menginisialisasi proses penghubungan antara Google Colab dan Google Drive. /content/Drivesaya adalah direktori tempat Google Drive akan di-mount atau dihubungkan.
3. **Mounted at /content/Drivesaya:** Setelah menjalankan perintah di atas, Google Colab akan meminta Anda untuk memberikan izin untuk mengakses Google Drive. Setelah memberikan izin, Google Colab akan mem-mount Google Drive ke direktori ditentukan (/content/Drivesaya). Ini berarti sudah memiliki akses ke file dan folder di Google Drive langsung dari Colab.

Menghubungkan google colab dengan GitHub



Copy to GitHub

Repository: ☐ Itfhnbl/BelajarABD

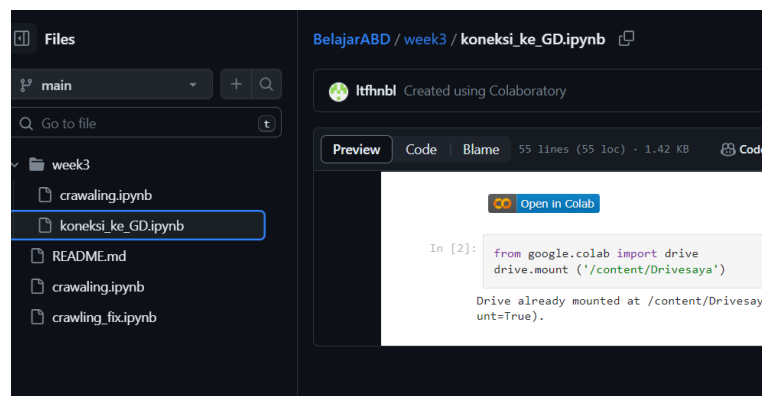
Branch: ☐ main

File path: koneksi_ke_GD.ipynb

Commit message: Created using Colaboratory

☒ Include a link to Colaboratory

Cancel OK



Crawling

Data crawling adalah proses otomatis untuk mengumpulkan dan mengindeks data dari berbagai sumber seperti situs web, database, atau dokumen

```

!pip3 install git+https://github.com/JustAnotherArchivist/snscrape.git
!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

```

!pip3 install git+https://github.com/JustAnotherArchivist/snscrape.git: Perintah yang digunakan untuk menginstal snscrape dari repositori GitHub. Snscrape adalah alat yang untuk mengambil data dari Twitter. Git+ dalam perintah pip, untuk menginstal snscrape langsung dari repositori GitHub.

!pip install pandas: Perintah yang digunakan untuk menginstal pandas, yang merupakan salah satu pustaka yang paling umum digunakan untuk manipulasi data dalam Python.

!sudo apt-get update: Untuk mengupdate daftar paket yang tersedia dalam sistem.

!sudo apt-get install -y ca-certificates curl gnupg: Menginstal paket-paket yang diperlukan, termasuk ca-certificates, curl, dan gnupg.

!sudo mkdir -p /etc/apt/keyrings: Perintah untuk membuat direktori /etc/apt/keyrings, yang akan digunakan untuk menyimpan kunci untuk repo Node.js.

!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg: Untuk mengunduh dan mengimpor kunci GPG yang diperlukan untuk mengakses repo Node.js.

!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: Perintah untuk menambahkan repo Node.js ke daftar sumber perangkat lunak pada sistem.

!sudo apt-get update: Untuk mengupdate daftar paket lagi setelah menambahkan repo Node.js.

!sudo apt-get install nodejs -y: Menginstal Node.js pada sistem.

!node -v: Digunakan untuk memeriksa versi Node.js yang telah diinstal pada sistem. Ini akan menampilkan versi Node.js yang digunakan.

```
# Crawl Data

filename = 'sianida.csv'
search_keyword = 'sianida until:2023-10-13 since:2023-01-01'
limit = 100

!npx --yes tweet-harvest@latest -o "{filename}" -s "{search_keyword}" -l {limit} --token "xxx"
```

filename = 'sianida.csv': Menentukan nama file CSV yang akan digunakan untuk menyimpan data tweet yang akan diambil. Data akan disimpan dalam file dengan nama "sianida.csv" di direktori kerja saat ini.

search_keyword = 'sianida until:2023-10-13 since:2023-01-01': Kunci pencarian yang akan digunakan untuk mengambil data tweet. Pencarian ini akan mencakup tweet-tweet yang mengandung kata "sianida" dan ditulis di antara tanggal 1 Januari 2023 hingga 13 Oktober 2023.

limit = 100: Jumlah maksimal tweet yang akan diambil. Dalam hal ini, Membatasi mengambil hingga 100 tweet.

tweet-harvest@latest: Nama paket yang akan dijalankan, dengan versi terbaru.

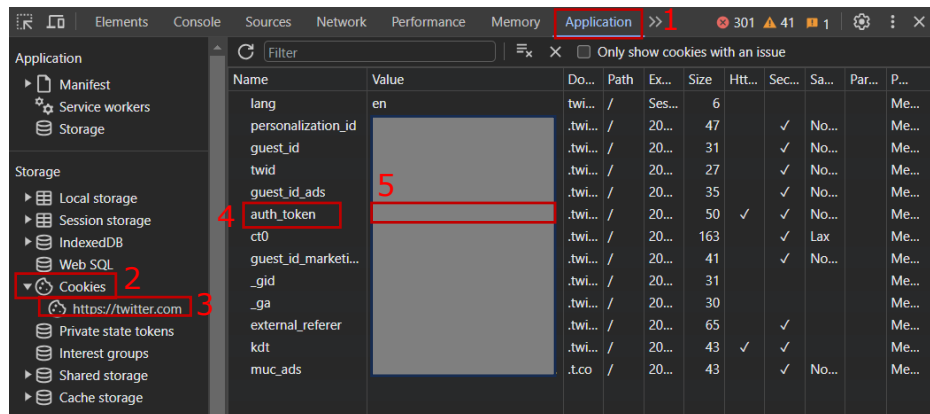
-o "{filename}": Digunakan untuk menentukan nama file keluaran (file CSV) yang akan digunakan untuk menyimpan data tweet.

-s "{search_keyword}": Digunakan untuk menentukan kata kunci pencarian.

-l {limit}: Digunakan untuk menentukan batas jumlah tweet yang akan diambil.

--token "xxx": Memberikan token akses Twitter yang diperlukan untuk mengambil data dari Twitter, yang dimana token tersebut bersifat rahasia.

Untuk mendapatkan token, kita harus login dengan akun twitter, lalu inspect, pada halaman inspect dengan cara klik kanan pada halaman web twitter lalu, ikuti langkah di bawah ini



Penjelasan :

1. Klik icon application
2. Lalu pilih cookies
3. Pada menu cookies pilih twitter
4. Selanjutnya pilih auth_token
5. Dan akan tampil token yang akan di gunakan di sampingnya

```
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)
```

file_path = f"tweets-data/{filename}": Menentukan path file CSV yang akan dibaca dalam variabel file_path. Ini menggabungkan direktori "tweets-data" dengan nama file yang telah ditentukan sebelumnya dalam filename.

df = pd.read_csv(file_path, delimiter=";"): Parameter delimiter=";" untuk menentukan bahwa delimiter (pemisah) dalam file CSV adalah titik koma (";").

display(df): Perintah untuk menampilkan DataFrame df dalam lingkungan seperti Jupyter Notebook atau Google Colab. Ini akan menampilkan data dalam bentuk tabel yang mudah dibaca.

Output:

index	created_at	id_str	full_text	quote_count	reply_count	retweet_count	favorite_count	lang	user_id_str	conversation_id_str	username	
0	Thu Oct 12 23:56:21 +0000 2023	1712618239736607062	Kejanggalan CCTV Jessica yg tertera dalam BAP DI jam2 terusan (15:35-18:59) yg "kalanya" Jessica memasukan sianida, resolusinya: 960x576, sementara selesai kejadian (18:58-18:25) resolusinya malah naik 1920x1080 Bkn modal ntn netflix ya, tp saya memang mengikut sidang 2016 https://t.co/RDXAMcZd4D	0	1	0	0	in	1114048636151668737	1712618239736607062	nkippings	https://twitter.com/nkippings
1	Thu Oct 12 23:54:54 +0000 2023	1712617874014244874	Skrang kta disuguhi DRAMA TOL OL, penegak hukum tentang kasus sianida.	0	0	0	0	in	352092774	1712617874014244874	barnescoy	https://twitter.com/barnescoy
2	Thu Oct 12 23:51:11 +0000 2023	1712616939254948173	@Jerhermyy tapi emang aneh banget dah. Sianida tuh kalau masuk ke tubuh bikin kulit berubah warna, gak mungkin membru karena derahnya udah rusak. Reaksi tubuh ke sianida juga ospet. Heran sama yg urus kasusnya, semua bukti pakernya dukun dukunan, gak ada tenaga ahli yg dibolehin ngomong.	0	0	0	0	in	1561029855634141185	1712541158696094123	omitmenot	https://twitter.com/omitmenot

```
# Cek jumlah data yang didapatkan
```

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

Kode ini menggunakan fungsi `len(df)` untuk menghitung jumlah baris (tweet) dalam DataFrame `df`. Kemudian, dengan menggunakan f-string (`f"..."`), pesan yang mencetak jumlah tweet dihasilkan dan dicetak ke layar.

Outputnya:

```
Jumlah tweet dalam dataframe adalah 109.
```