MODUL

BIG DATA ANALYTIC



Disusun oleh :

Sri Redjeki, S.Si, M.Kom.

Maria Mediatrix S., S.Kom., M.Eng.

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER

AKAKOM

YOGYAKARTA

2020

# KATA PENGANTAR

# DAFTAR ISI

[KATA PENGANTAR ii](#_Toc15905388)

[DAFTAR ISI iii](#_Toc15905389)

[MODUL 1 NAMA POKOK BAHASAN 1](#_Toc15905390)

# MODUL 4 COLLECTING DATA

|  |  |
| --- | --- |
|  | CAPAIAN PEMBELAJARAN |
| 1. Mampu melakukan *scraping* data menggunakan twitterscraper | |

|  |  |
| --- | --- |
|  | KEBUTUHAN ALAT/BAHAN/SOFTWARE |
| 1. Python (Anaconda / Miniconda) 2. Jupyter Notebook 3. Twitterscraper | |

|  |  |
| --- | --- |
|  | DASAR TEORI |

*Data Scrapping* merupakan kegiatan untuk menemukan data dan mengumpulkannya. Data diambil secara langsung dari sebuah sumber, tidak hanya dari web tetapi bisa dari berbagai sumber dimana terdapat data. Ini mencakup *spreadsheets*, *storage devices*, dll. Proses ini diperlukan untuk menyaring (*filtering*) dan membedakan berbagai jenis data mentah dari berbagai sumber menjadi sesuatu yang berguna dan informatif.

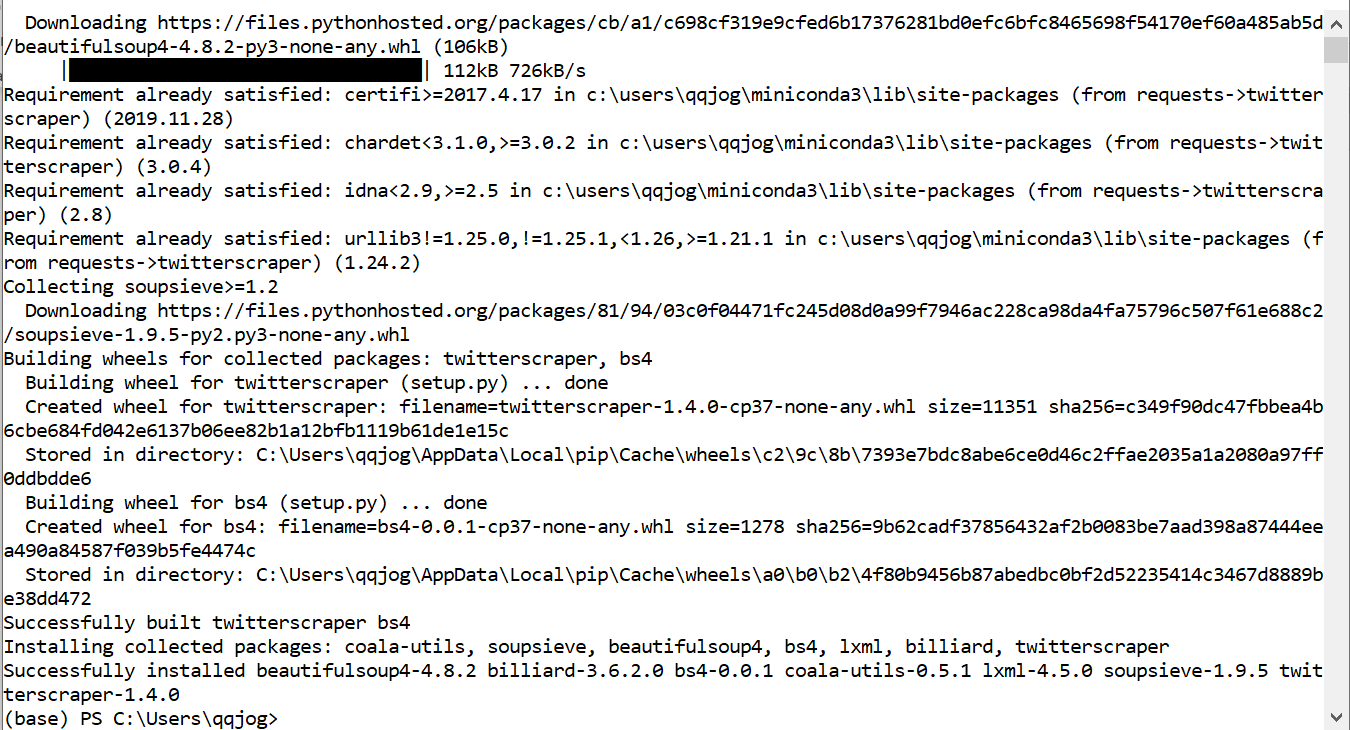
*Data Crawling* adalah cara mengumpulkan semua data dalam suatu website dengan menggunakan metode tertentu dan secara otomatis (*web crawling*). Web crawler akan mengunjungi setiap alamat website yang diberikan kepadanya, kemudian menyerap (*scraping*) dan menyimpan semua informasi yang terkandung didalam website tersebut. Setiap kali web crawler mengunjungi sebuah website, maka dia juga akan mendata semua link yang ada dihalaman yang dikunjunginya itu untuk kemudian dikunjungi lagi satu persatu. Proses crawling dalam suatu website dimulai dari mendata seluruh url dari website, menelusurinya satu-persatu, kemudian memasukkannya dalam daftar halaman pada indeks *search engine*, sehingga setiap kali ada perubahan pada website, akan terupdate secara otomatis. *Web crawling* adalah proses mengambil kumpulan halaman dari sebuah web untuk dilakukan pengindeksan sehingga mendukung kinerja mesin pencari. Salah satu contoh situs yang menerapkan *web crawling* adalah www.webcrawler.com . Di samping situs-situ mesin pencari terkemuka tentunya, seperti Google, Ask, Live, dan lain sebagainya.

Salah satu *social media* yang sangat popular dan hampir dimiliki oleh semua penduduk dunia adalah Twitter. Di Twitter pengguna dapat mengirim dan membaca pesan yang biasanya disebut dengan istilah “tweets”. Setiap hari jutaan tweet muncul untuk mengekspresikan pendapat dalam berbagai hal. Tweet tersebut tentu saja bisa menjadi sumber data dan memiliki nilai yang dapat dimanfaatkan untuk berbagai macam keperluan baik untuk penelitian maupun bisnis. Dalam modul ini, kita akan mencoba mengumpulkan data dari twitter. Python telah menyediakan library khusus yang dapat digunakan untuk mengumpulkan data twitter. Terdapat 2 cara untuk mengumpulkan data twitter, yaitu dengan menggunakan twitterscraper atau bisa juga menggunakan API twitter.

|  |  |
| --- | --- |
|  | PRAKTIK |

1. Install twitterscrapper 🡪 apabila belum ada. Gunakan perintah pip install twitterscraper

Kemudian muncul tampilan seperti dibawah ini :

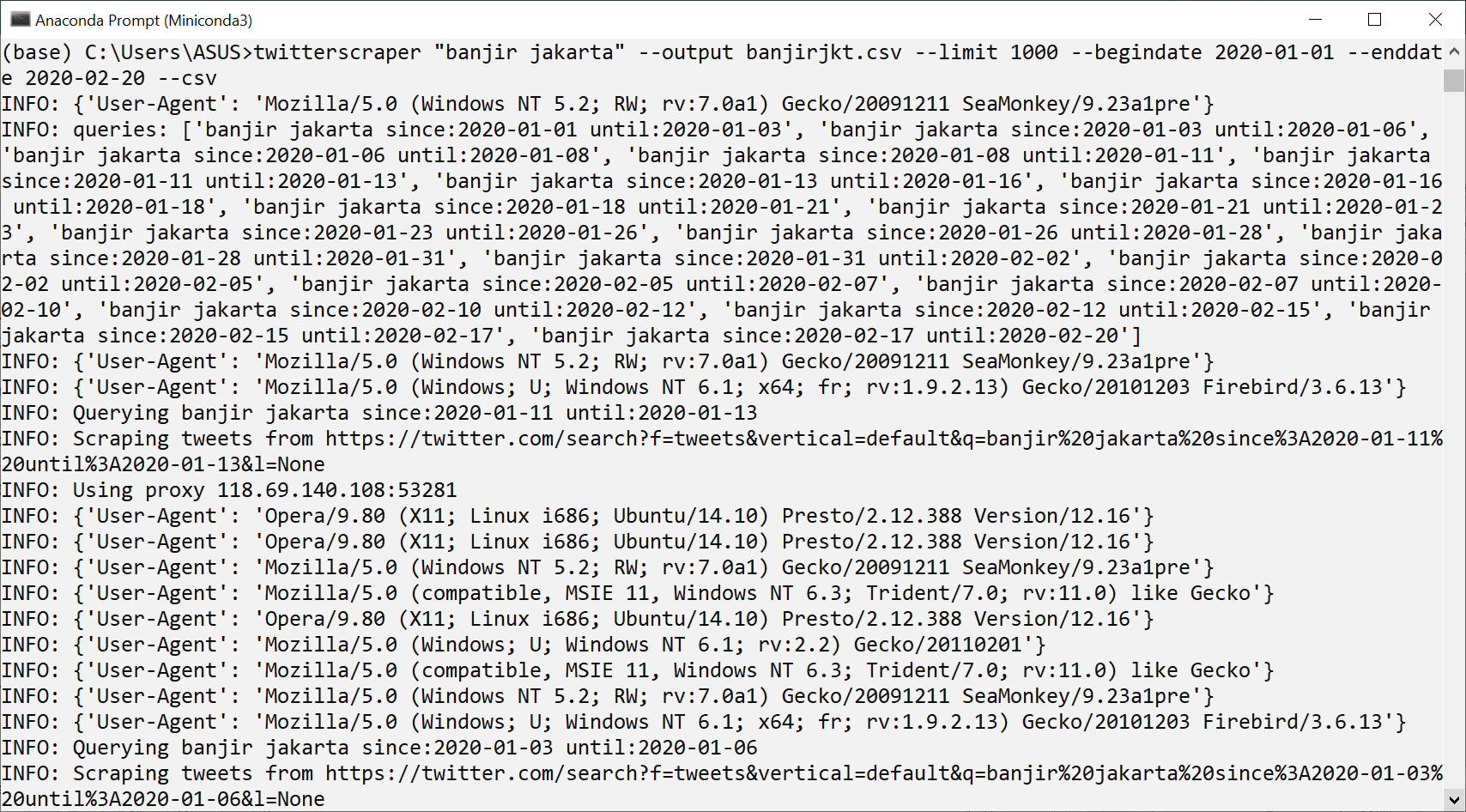


1. Misalnya kita akan mengambil data tweet tentang : “**banjir jakarta**”. Dengan ketentuan :

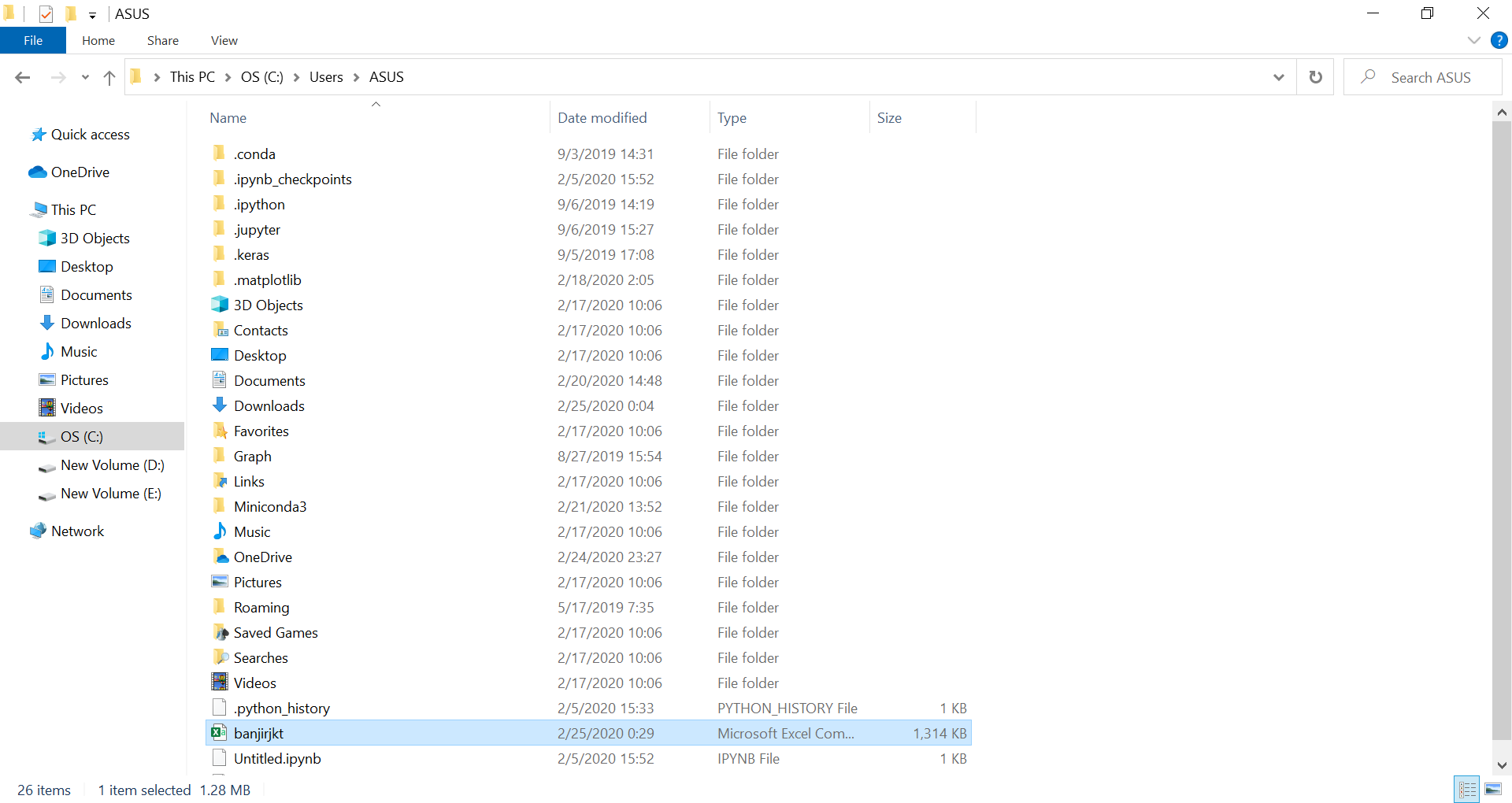
* Membatasi dan mencari tweet sebanyak: **1000**
* Simpan menjadi file: **.csv**
* Nama (output) file: **banjirjkt.csv**
* Ambil Tweet mulai tanggal: **1 Januari 2020 (2020–01–01)**
* Sampai dengan tanggal: **20 Februari 2020 (2020–02–20)**

1. Ketikkan perintah berikut ini :
2. Mengumpulkan dan menyimpan data ke file csv

twitterscraper "banjir jakarta" --output banjirjkt.csv --limit 1000 --begindate 2020-01-01 --enddate 2020-02-20 --csv



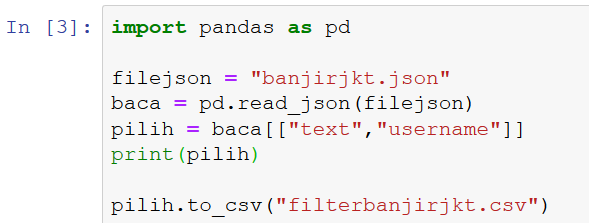
File hasil scraping akan disimpan dalam folder yang sama dengan folder tempat kita bekerja. (karena perintah diketik di direktori C:\Users\ASUS maka file .csv juga akan disimpan di direktori tersebut)



1. Mengumpulkan dan menyimpan data menjadi file json

twitterscraper "banjir jakarta" --output banjirjkt.json --limit 1000 --begindate 2020-01-01 --enddate 2020-02-20

Bukalah Jupyter Notebook, kemudian ketikkan program berikut :



|  |  |
| --- | --- |
|  | LATIHAN |

1. Lakukanlah scraping dengan menggunakan kata kunci yang berbeda. Kemudian lakukan filter untuk menyimpan hanya data “timestamp”, “likes”, dan “retweets”

|  |  |
| --- | --- |
| tugas.png | TUGAS |

|  |  |
| --- | --- |
|  | REFERENSI |