LAPORAN PRAKTIKUM

BIG DATA ANALYTIC

PERTEMUAN KE-04

Disusun Oleh :

NIM : 195610007

NAMA : Ara Widhi Astutik

KELAS : Sistem Informasi-1

UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA

2021/2022

BAB I

COLLECTING DATA

PEMBAHASAN

Praktik

1. Akan dilakukan crawling data twitter menggunakan twint dengan Jupiter notebook atau google colab. Gunakan keyword : Omicron di Indonesia atau vaksin booster
   Codingnya : ketik koding dibawah ini

```
[7]  !git clone --depth=1 https://github.com/twintproject/twint.git

Cloning into 'twint'...
remote: Enumerating objects: 47, done.
remote: Counting objects: 100% (47/47), done.
remote: Compressing objects: 100% (44/44), done.
remote: Total 47 (delta 3), reused 14 (delta 0), pack-reused 0
Unpacking objects: 100% (47/47), done.
```

   Uraian :
   Clone adalah proses untuk menduplikasikan remote repo di GitHub ke komputer lokal.Setelah berhasil membuat repository pada github, berikutnya mengcopy link tersebut (HTTPS) untuk nantinya akan digunakan pada saat proses clone. Dalam hal ini menggunakan repository twintproject/twint.git

```
[8] %cd twint

/content/twint
```

```
[9] !pip3 install . -r requirements.txt

  Downloading multidict-6.0.2-cp37-cp37m-manylinux_2_17_x86_64.manylinux2
     |████████████████████████████████| 94 kB 4.4 MB/s
Collecting async-timeout<5.0,>=4.0.0a3
  Downloading async_timeout-4.0.2-py3-none-any.whl (5.8 kB)
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in /usr/local
Collecting frozenlist>=1.1.1
  Downloading frozenlist-1.3.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1
     |████████████████████████████████| 144 kB 54.8 MB/s
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.7/dist
Collecting pycares>=4.0.0
  Downloading pycares-4.1.2-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x8
     |████████████████████████████████| 291 kB 61.6 MB/s
Requirement already satisfied: cffi>=1.5.0 in /usr/local/lib/python3.7/di
Requirement already satisfied: pycparser in /usr/local/lib/python3.7/dist
Collecting elastic-transport<9,>=8
  Downloading elastic_transport-8.1.0-py3-none-any.whl (58 kB)
     |████████████████████████████████| 58 kB 6.6 MB/s
Requirement already satisfied: certifi in /usr/local/lib/python3.7/dist-p
Collecting urllib3<2,>=1.26.2
  Downloading urllib3-1.26.8-py2.py3-none-any.whl (138 kB)
     |████████████████████████████████| 138 kB 67.8 MB/s
```


Uraian :

Penggunaan file requirements.txt ini akan mencantumkan semua dependensi Python yang di lakukan untuk sebuah proyek juga bisa untuk membuat fungsi tanpa server atau aplikasi web.

```
[10] !pip install aiohttp==3.7.0

Collecting aiohttp==3.7.0
  Downloading aiohttp-3.7.0-cp37-cp37m-manylinux2014_x86_64.whl (1.3 MB)
     |████████████████████████████████| 1.3 MB 7.7 MB/s
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/
Requirement already satisfied: chardet<4.0,>=2.0 in /usr/local/lib/python
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.7
Collecting async-timeout<4.0,>=3.0
  Downloading async_timeout-3.0.1-py3-none-any.whl (8.2 kB)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/pyth
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.7/dist
Requirement already satisfied: typing-extensions>=3.7.4 in /usr/local/lib
Installing collected packages: async-timeout, aiohttp
  Attempting uninstall: async-timeout
    Found existing installation: async-timeout 4.0.2
    Uninstalling async-timeout-4.0.2:
      Successfully uninstalled async-timeout-4.0.2
  Attempting uninstall: aiohttp
    Found existing installation: aiohttp 3.8.1
    Uninstalling aiohttp-3.8.1:
      Successfully uninstalled aiohttp-3.8.1
Successfully installed aiohttp-3.7.0 async-timeout-3.0.1
```

Uraian :
Pada tampilan diatas merupakan proses installasi aiohttp==3.7.0

Setelah success ketik koding berikut ini

```
[11] import nest_asyncio
     nest_asyncio.apply()
     import twint
```

Setelah itu run dan sekarang ketik keyword untuk mengambil data yang diinginkan

```
[12] c = twint.Config()
     c.Search = 'vaksin booster di Indonesia'
     c.Pandas = True
     twint.run.Search(c)
```

Uraian :
Pada source code diatas, akan mencoba mengambil data vaksin booster di Indonesia melalui twint.

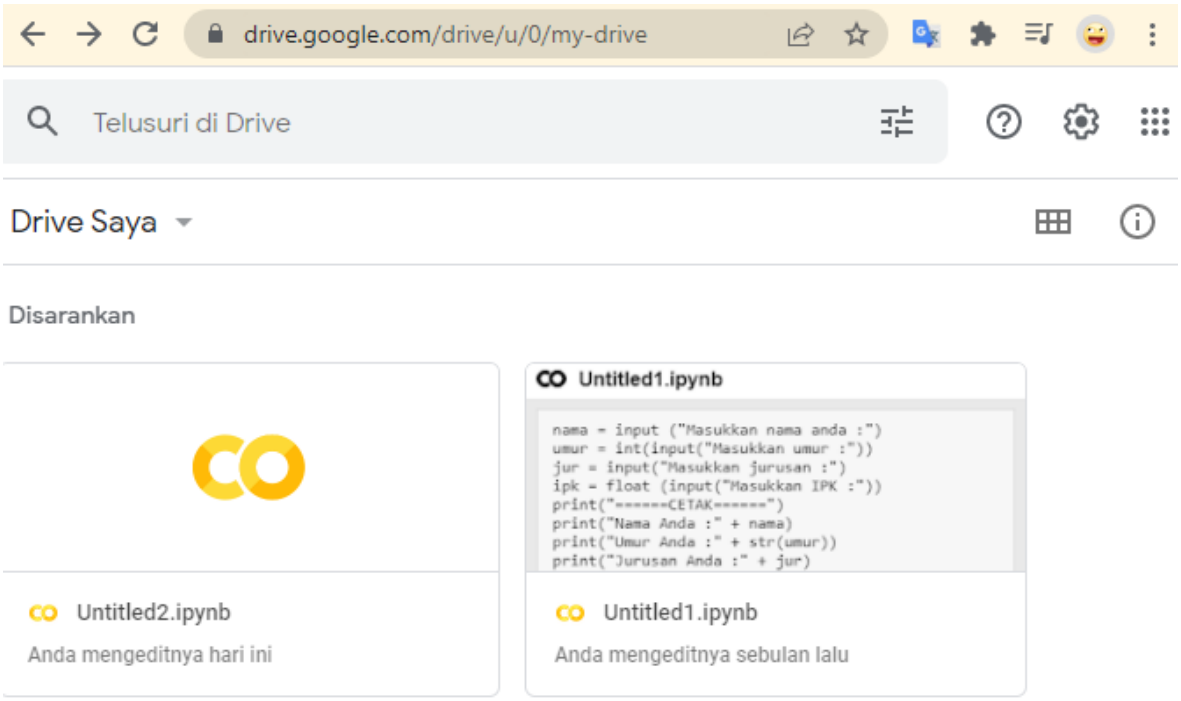
Setelah di run akan tampil hasil crawling yang diperoleh dari twitter

```
1502961573387997191 2022-03-13 10:55:37 +0000 <hasbielrchlv> @hyuga3811 @
1502917402010529795 2022-03-13 08:00:06 +0000 <asumsico> Menurut Ketua Sa
1502863265474383875 2022-03-13 04:24:58 +0000 <kimiafarmaind> bp Indonesi
1502846390442991618 2022-03-13 03:17:55 +0000 <lethalgrey> Wkwk booster d
1502841986591731712 2022-03-13 03:00:25 +0000 <javasatumedia> Akselerasi
1502837124843769859 2022-03-13 02:41:06 +0000 <tribunnews> 14,5 Juta Oran
[!] No more data! Scraping will stop now.
found 0 deleted tweets in this search.
```

Hasil crawling diatas dapat disimpan dalam bentuk csv yang nanti di olah untuk analisis sentimen. Penyimpanan dapat dilakukan di drive kita atau di google drive. Apabila disimpan di drive maka dapat kita buat folder drive dengan koding sebagai berikut

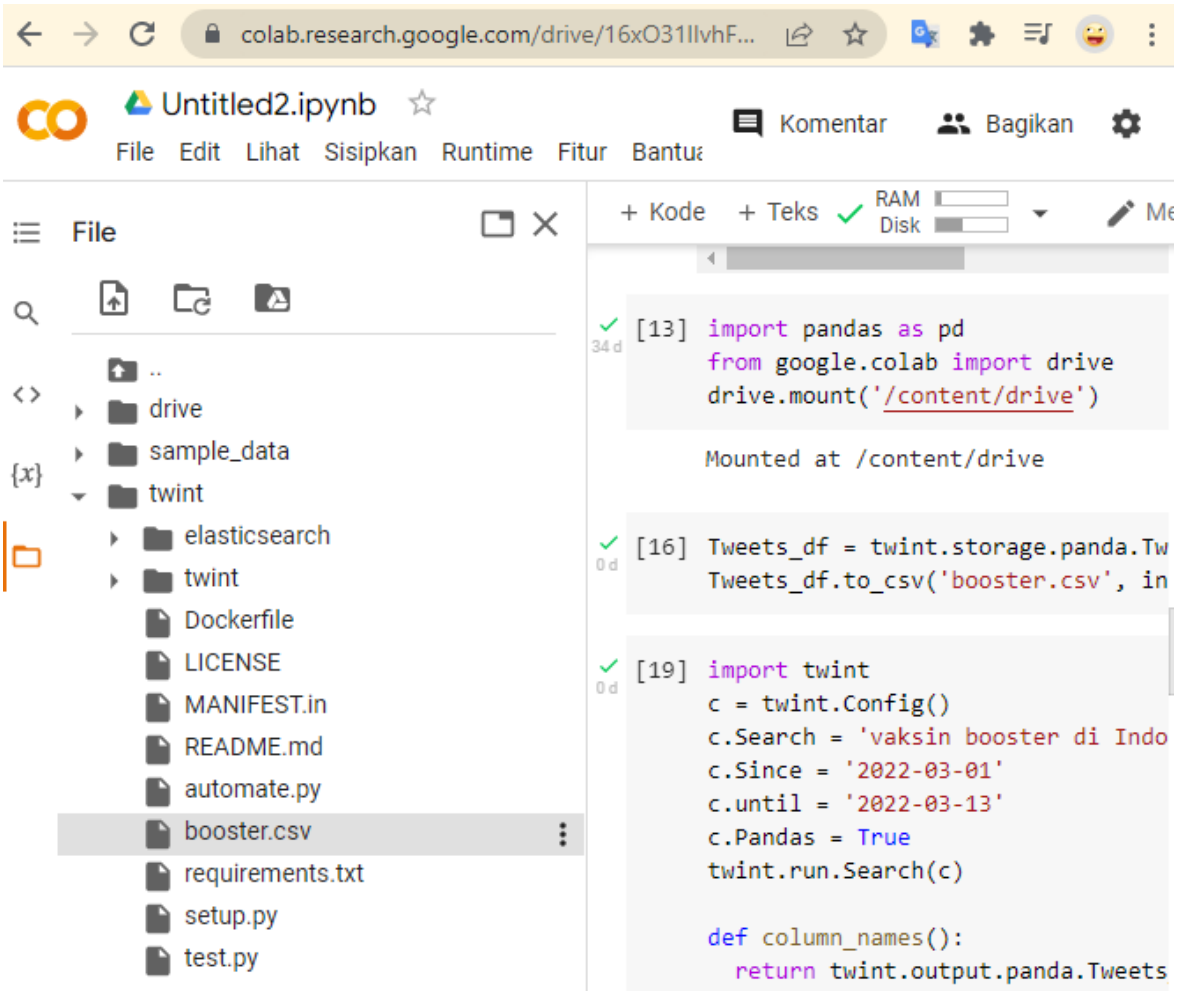
```
[13] import pandas as pd
     from google.colab import drive
     drive.mount('/content/drive')
```

```
Mounted at /content/drive
```



Nanti kita akan diarahkan ke email yang digunakan pada google drive dan pilih aktifkan. Setelah itu dapat kita buat folder pada google drive kita dengan nama dan identitas kita masing2. Pada praktek ini saya gunakan koding sebagai berikut

```
[16] Tweets_df = twint.storage.panda.Tweets_df.head()
     Tweets_df.to_csv('booster.csv', index=False)
```



Uraian :

Pada tahap ini, saya mengalami kendala dimana tidak dapat menyimpan file booster.csv kedalam google drive. Oleh karena itu, saya menyimpannya file csv tersebut ke dalam google colab. Ketika membuka google drive, yang tersimpan hanyalah file projek yang ada pada google.colab.

Latihan

1. Lakukanlah scraping dengan menggunakan kata kunci yang berbeda dan bersifat umum . Kemudian lakukan filter hanya yang berbahasa Indonesia.

   Uraian :

```
[19]  import twint
      c = twint.Config()
      c.Search = 'vaksin booster di Indonesia'
      c.Since = '2022-03-01'
      c.until = '2022-03-13'
      c.Pandas = True
      twint.run.Search(c)

      def column_names():
        return twint.output.panda.Tweets_df.columns

      def twint_to_pd(columns):
        return twint.output().panda.Tweets_df(columns)

      print(column_names())
```

```
1573387997191 2022-03-13 10:55:37 +0000 <hasbielrchlv> @hyuga3811 @Heraloe
7402010529795 2022-03-13 08:00:06 +0000 <asumsico> Menurut Ketua Satgas Pe
3265474383875 2022-03-13 04:24:58 +0000 <kimiafarmaind> bp Indonesia merup
6390442991618 2022-03-13 03:17:55 +0000 <lethalgrey> Wkwk booster disini p
1986591731712 2022-03-13 03:00:25 +0000 <javasatumedia> Akselerasi Vaksin
7124843769859 2022-03-13 02:41:06 +0000 <tribunnews> 14,5 Juta Orang di In
 more data! Scraping will stop now.
0 deleted tweets in this search.
['id', 'conversation_id', 'created_at', 'date', 'timezone', 'place',
 'tweet', 'language', 'hashtags', 'cashtags', 'user_id', 'user_id_str',
 'username', 'name', 'day', 'hour', 'link', 'urls', 'photos', 'video',
```

   Uraian :

   Ketika melakukan scrapping menggunakan data vaksin booster di Indonesia, data yang ingin saya ambil berdasarkan hasil scrapping adalah data vaksin sedari 2022-03-01 sampai dengan 2022-03-13, pada saat proses scrapping dijalankan maka akan menampilkan index dari data tweet vaksin booster di Indonesia. Dalam hal ini , kita dapat memilih beberapa data yang ingin kita tampilkan seperti dibawah ini, dimana saya memilih data (date, username, dan tweet) untuk ditampilkan secara data frame(df) agar mudah untuk memahaminya.

```
import twint
c = twint.Config()
c.Search = 'vaksin booster di Indonesia'
c.Since = '2022-03-01'
c.until = '2022-03-13'
c.Pandas = True
twint.run.Search(c)

def column_names():
  return twint.output.panda.Tweets_df.columns

def twint_to_pd(columns):
  return twint.output.panda.Tweets_df[columns]

print(column_names())
['id', 'conversation_id', 'created_at', 'date', 'timezone', 'place',
       'tweet', 'language', 'hashtags', 'cashtags', 'user_id', 'user_id_s
       'username', 'name', 'day', 'hour', 'link', 'urls', 'photos', 'vide
       'thumbnail', 'retweet', 'nlikes', 'nreplies', 'nretweets', 'quote_
       'search', 'near', 'geo', 'source', 'user_rt_id', 'user_rt',
       'retweet_id', 'reply_to', 'retweet_date', 'translate', 'trans_src'
       'trans_dest']

data = twint_to_pd(['date','username','tweet'])
print(data)
```

```
Traceback (most recent call last):
  File "/usr/lib/python3.7/asyncio/events.py", line 88, in _run
    self._context.run(self._callback, *self._args)
RuntimeError: cannot enter context: <Context object at 0x7fb26afa1960> is
[!] No more data! Scraping will stop now.
found 0 deleted tweets in this search.
Index(['id', 'conversation_id', 'created_at', 'date', 'timezone', 'place',
       'tweet', 'language', 'hashtags', 'cashtags', 'user_id', 'user_id_s
       'username', 'name', 'day', 'hour', 'link', 'urls', 'photos', 'vide
       'thumbnail', 'retweet', 'nlikes', 'nreplies', 'nretweets', 'quote_
       'search', 'near', 'geo', 'source', 'user_rt_id', 'user_rt',
       'retweet_id', 'reply_to', 'retweet_date', 'translate', 'trans_src',
       'trans_dest'],
      dtype='object')
                 date       username  \
0  2022-03-13 10:55:37   hasbielrchlv
1  2022-03-13 08:00:06       asumsico
2  2022-03-13 04:24:58  kimiafarmaind
3  2022-03-13 03:17:55     lethalgrey
4  2022-03-13 03:00:25  javasatumedia
5  2022-03-13 02:41:06     tribunnews

                                               tweet
0  @hyuga3811 @Heraloebss Utk target vaksin indon...
1  Menurut Ketua Satgas Penanganan Covid-19 IDI, ...
2  bp Indonesia merupakan salah satu perusahaan e...
3  Wkwk booster disini pertanyaannya beda yak mal...
4  Akselerasi Vaksin Booster Jatim, 600 Dosis Dis...
5  14,5 Juta Orang di Indonesia Sudah Disuntik Va...
```

Tugas

1. Lakukan crawling data untuk keyword minyak goreng langka hanya untuk bulan Februari 2022

```
[23] import twint
     c = twint.Config()
     c.Search = 'minyak goreng langka bulan Februari 2022'
     c.Pandas = True
     twint.run.Search(c)
```

```
1500996070658084865 2022-03-08 00:45:25 +0000 <grandysofia> Sejak bulan F
1497050087893385216 2022-02-25 03:25:29 +0000 <TehLita_> Jelang akhir bul
1495354074702696450 2022-02-20 11:06:08 +0000 <babang_cky80> Jatuh di bul
[!] No more data! Scraping will stop now.
found 0 deleted tweets in this search.
```

Uraian :

Sama halnya ketika pada sebelumnya yaitu mencari crawling data untuk keyword vaksin booster di Indonesia. Namun dalam hal ini keyword yang ingin dicari adalah minyak goreng langka untuk bulan Februari 2022, maka source code dan tampilan hasil keluaran crawling data seperti pada gambar diatas.

BAB II

KESIMPULAN

Setelah melakukan praktikum ini dapat saya simpulkan bahwa , crawling data twitter menggunakan python sendiri merupakan salah satu cara untuk mengumpulkan data teks. Dalam hal ini , tidak hanya teks yang dapat didapat dari twitter, kita dapat mengumpulkan data waktu interaksi,jumlah like,gambar,audio,dan lain-lain tentunya sesuai dengan keyword yang digunakan dalam mengumpulkan data tersebut.

Selanjutnya dalam praktikum ini diajarkan juga untuk crawling data twitter menggunakan twint. Twint adalah sebuah tools yang digunakan untuk mendapat data dari twitter. Twint dikembangkan menggunakan bahasa pemograman python dan dapat diinstall menggunakan pip ataupun conda. Setelah terinstall, twint bisa digunakan dalam 2 cara yakni dalam source code atau sebagai command line. Source code resmi untuk penggunaan twint dala praktikum ini, dapat dilihat pada praktik langkah 1 (no 1).

BAB III

DAFTAR PUSTAKA

https://elearning.utdi.ac.id/pluginfile.php/16819/mod_resource/content/1/Modul%204_crawlingTwitter.pdf

https://drive.google.com/drive/u/0/my-drive