MODUL DATA MINING

TEXT EXTRACTION

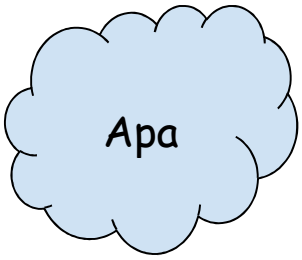
Pada modul ini dijelaskan mengenai konsep ekstraksi data teks dan menerapkannya dengan mempergunakan bahasa pemrograman python.
Diharapkan setelah mempelajari modul ini, mahasiswa mampu memahami tujuan ekstraksi data teks, dan dapat melakukan ekstraksi data teks dari berbagai sumber.

EPS 8

Ekstraksi Data Teks

Konsep Ekstraksi Data

Ekstraksi Data mengacu pada proses pengambilan data dari suatu sumber data (bisa dari banyak sumber) dan mengubahnya dari satu format ke format yang lain agar bisa diproses lebih lanjut, umumnya data dari website, diambil kemudian disimpan menjadi bentuk excel, csv atau txt. Ekstraksi data dilakukan sebelum pemrosesan atau analisis data. Ekstraksi Data dapat digunakan dalam banyak skenario yang berbeda, seperti pengarsipan, pemindahan/transfer data atau analisis.



- Pengarsipan

  Ekstraksi data dipergunakan untuk membuat salinan dari suatu data untuk diamankan atau sebagai cadangan. Contoh umum adalah menggunakan ekstraksi data untuk mengkonversi data dari format fisik ke format digital agar dapat dapat disimpan dengan lebih aman dan lebih lagi.

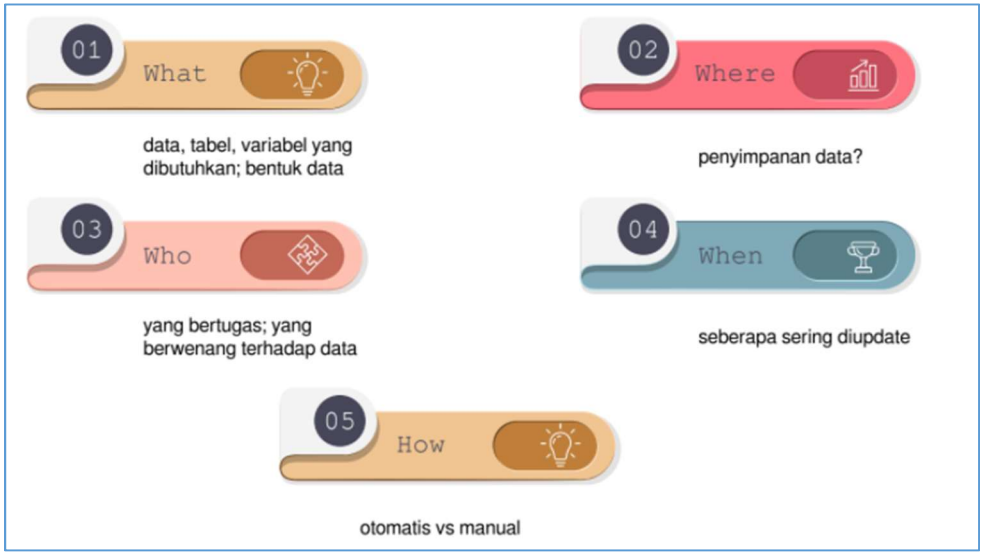
- Pemindahan/Transfer data

  Ekstraksi data digunakan untuk mentransfer atau memindahkan satu set data dari satu format ke format lain tanpa membuat perubahan apa pun pada data itu sendiri. Misalnya, mengekstrak data dari versi situs web saat ini ke versi yang lebih baru dari situs yang saat ini sedang dalam pengembangan.

- Analisis

  Penggunaan paling umum dari ekstraksi data adalah untuk analisis data. Ini merujuk pada pengetahuan/wawasan apa pun yang dapat ditemukan dari hasil analisis data yang sudah terkestrak. Misalnya, mengekstrak harga dan rating produk untuk semua komputer atau laptop di tokopedia dan mendapatkan wawasan bahwa harga dari suatu item berkorelasi dengan peringkat item itu sendiri.

Untuk mengekstrak suatu data, perlu diperhatikan apa, dimana, siapa, kapan dan bagaimana dari data tersebut.



- Apa

  Data apa saja yang diperlukan, tabel apa saja yang akan diakses, variabel apa saja yang akan diambil, dan bentuk dari data itu sendiri. Ada berbagai macam bentuk sumber data, secara garis besar dapat dikelompokkan menjadi dua kategori yaitu :

  1. Sumber Digital

     Data digital adalah salah satu sumber data yang paling umum di zaman modern. Ini merujuk pada segala jenis kumpulan data yang disimpan dalam suatu file baik secara online atau di penyimpanan lokal dari suatu perangkat. Bisa disimpan dalam bentuk database maupun struktur data lain yang lebih kompleks.

2. Sumber Fisik

   Data fisik biasanya ada di media cetak atau fisik. Dalam hal ini, mengacu pada buku, surat kabar, laporan keuangan (versi cetak), faktur, dll.

- Dimana

  Dimana lokasi data tersebut berada? Apakah dari suatu database yang ada di suatu server? Apakah data berasal dari suatu website? Apakah data berasal dari aplikasi desktop dan lain-lain.

- Siapa

  Pertanyaan-pertanyaan yang perlu diperhatikan terkait kepemilikan data adalah : Siapa yang memiliki data tersebut? Apakah data tersebut untuk konsumsi publik?(Tashea, 2019) Bagaimana *term&conditions* dari website tersebut? Apakah perlu mengurus *license agreement* untuk memakai data dari website tersebut ? jika iya, siapa yang harus dikontak? Pertanyaan tersebut perlu dijawab agar tidak ada kendala terkait hukum dikemudian hari (Bode, 2019; *How to Use Terms and Conditions for Web Scraping Protection*, n.d.; Kernel, 2019).

- Kapan

  Apakah hanya akan mengambil data sekali saja untuk keperluan analisa, setelah analisa selesai, tidak akan mengambil data terbaru dari sumber data tersebut? Atau data sering berubah, dan data terbaru diperlukan untuk melakukan analisa? Jika iya, Seberapa sering data tersebut mengalami perubahan? Seberapa sering diperlukan ekstraksi data yang baru? Apakah perlu ektraksi data secara realtime? (Naeem, 2020) Bagaimana cara membedakan data yang lama dengan data terbaru agar tidak terjadi duplikasi data dan lain-lain.

- Bagaimana

  Setelah mengetahui data apa yang akan diambil, dimana lokasi data tersebut berada dan kapan mengambilnya, langkah berikutnya adalah mengambil data tersebut. Teknik untuk mengambil data terbagi menjadi 3, yaitu manual, semi otomatis dan otomatis.

  1. Manual
     Biasanya dilakukan jika sumber data berupa fisik (berupa cetak/hardcopy). Teknik yang dilakukan adalah membaca dan mengetik/menginputkan data secara manual, resikonya adalah salah input data, dan menghabiskan waktu cukup banyak (Naeem, 2020). Sumber data digital pun mungkin dieksrtaksi secara manual, seperti kasus mendownload/mengkopi content/isi dari suatu website satu persatu ke dalam suatu file kemudian menyimpan data tersebut dalam format lain yang lebih mudah diakses (txt atau csv).
  2. Semi otomatis

Jika sumber data berupa fisik (file pdf atau print out suatu dokumen), untuk mengekstrak informasi yang terkandung didalamnya dan meminimalkan kesalahan inputan user (jika dilakukan secara manual), maka teknologi OCR dapat digunakan untuk menscan (membaca), menangkap konten dari suatu media cetak kemudian menginputkan nilai yang sudah terbaca ini ke dalam suatu halaman tertentu secara otomatis. Metode ini termasuk semi otomatis, karena masih ada keterlibatan manusia didalamnya, seperti secara manual menscan dokumen. Jika sumber data berupa digital, dan berasal dari website atau aplikasi desktop (besar kemungkinan berupa data tidak terstruktur), teknik semi otomatis seperti pada kasus membuat script untuk melakukan crawling/scrapping sumber data diperlukan. Script ini haris dijalankan/dieksekusi/dipanggil secara manual agar dapat mengambil data yang diperlukan (Akbar et al., 2016). Atau mempergunakan API yang memang sudah disediakan oleh sumber data untuk mengakses data-data yang diperlukan. Tetapi jika sumber data dari suatu database, berarti data yang diakses adalah data terstruktur, dan alamat server database tersebut diketahui, perintah SQL dapat dipergunakan untuk mengambil data yang diperlukan.

3. Otomatis

   Ekstraksi data secara otomatis, hanya bisa dilakukan jika sumber datanya berupa digital, baik itu dari suatu website ataupu dari aplikasi desktop. Ada berbagai macam data extraction tools yang beredar dipasaran, baik yang gratis maupun berbayar (*Data Extraction Tools: Improving Data Warehouse Performance - Talend*, n.d.; Naeem, 2020; Sharma, 2020). Atau bisa juga dengan membuat sendiri script untuk ekstraksi data dari sumber data (bisa mempergunakan PHP, Python, ataupun R) kemudian menjalankan script tersebut secara periodik melalui fitur task scheduler di windows atau fitur sejenis di OS lainnya

Ekstraksi Data Menggunakan web scrapping

Web scraping adalah mengekstrak data dari suatu website secara langsung dengan mempergunakan protokol http. Web Scraping bisa menjadi solusi dalam mendapatkan informasi dari sebuah situs web jika situs tersebut tidak menyediakan API untuk pengambilan informasi (Pernanda, 2018). Jika ingin mengekstrak artikel berita (atau teks apapun) dari sebuah situs web, langkah pertama adalah mengetahui cara kerja situs web.

1. Pengenalan singkat tentang HTML dan desain halaman web

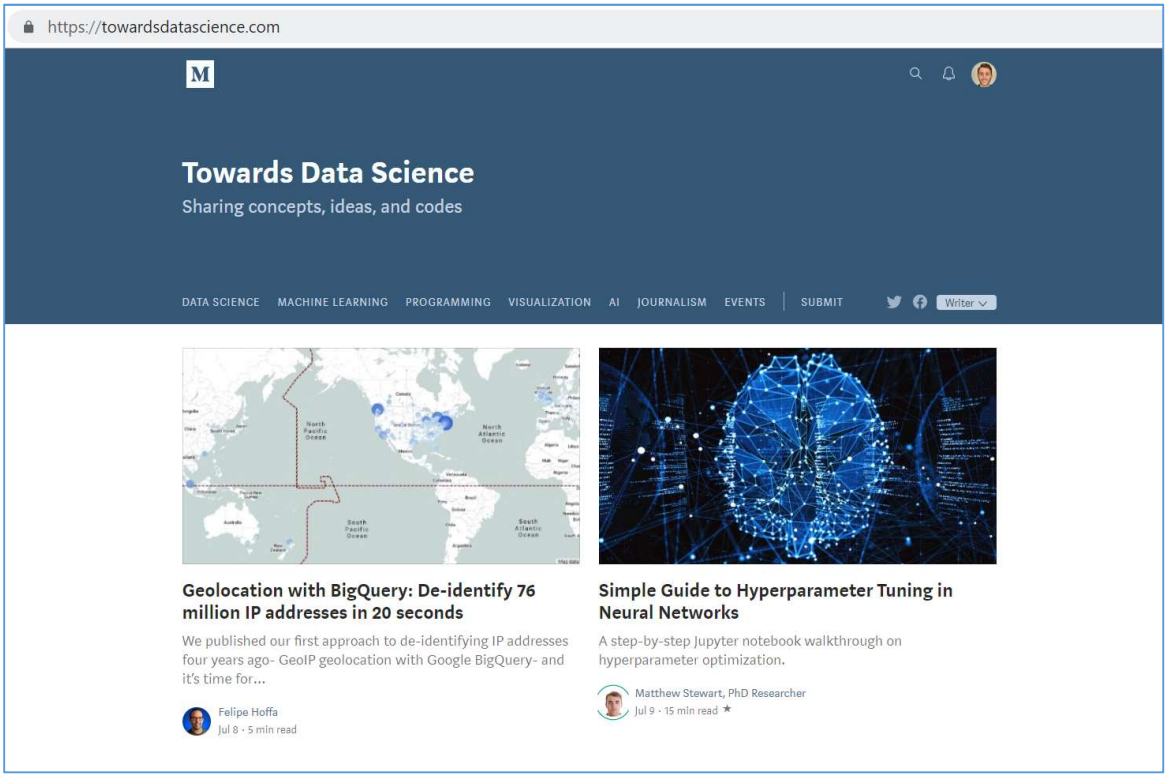
Saat sebuah url dimasukkan ke browser web (mis. Google Chrome, Firefox, dll…) dan pengguna mengakses situs tersebut, ada kombinasi dari tiga teknologi yang bekerja bersamaan :

1. HTML (HyperText Markup Language): merupakan bahasa standar untuk menambahkan suatu konten ke situs web. Dengan memakai HTML, memungkinkan gambar, teks dan file jenis lain dimasukkan ke dalam suatu situs. HTML menentukan konten dari suatu web.

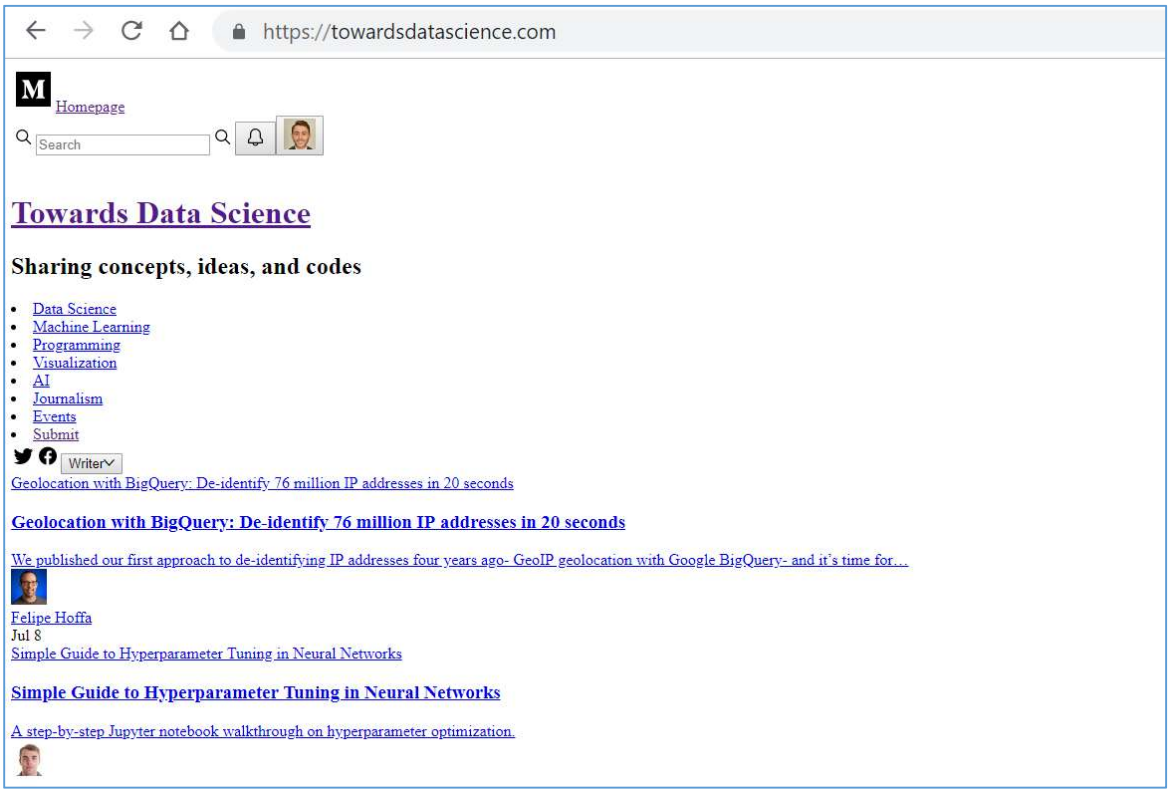
2. CSS (Cascading Style Sheets): bahasa ini dipergunakan untuk mengatur desain visual dari sebuah website. Teknologi ini yang menentukan gaya dari suatu halaman web.

3. JavaScript: JavaScript memungkinkan konten dan gaya dari suatu website menjadi lebih interaktif.

Ketiga teknologi tersebut memungkinkan untuk membuat dan memanipulasi setiap aspek desain halaman web. Ilustrasi konsep diatas dalam sebuah contoh adalah sebagai berikut. Saat ada pengguna yang mengakses halaman beranda dari Towards Data Science, halaman yang terlihat adalah sebagai berikut:



Jika CSS dihapus dari halaman web, halaman yang tampil adalah seperti berikut:



Dan jika javascript dinonaktifkan, pop up seperti berikut, tidak akan bisa muncul lagi

Jika ingin mengekstrak konten laman web melalui scraping web, bagian mana yang perlu dicari?

Jawaban : HTML

HTML, dari sudut pandang yang sangat dasar, terdiri dari elemen-elemen yang memiliki atribut. Elemen bisa berupa paragraf, dan atributnya bisa jadi paragraf tersebut dicetak tebal. Ada banyak jenis elemen, masing-masing dengan atributnya sendiri. Untuk mengidentifikasi elemen dipergunakan tag. Tag ini direpresentasikan dengan simbol <> (misalnya, tag <p> berarti teks tertentu akan berperilaku sebagai paragraf).

Misalnya, kode HTML di bawah ini memungkinkan kita untuk mengubah rata kanan atau tengah atau kiri dari suatu paragraf:

```
<!DOCTYPE html>
<html>

   <body>
      <p align = "left">This is left aligned</p>
      <p align = "center">This is center aligned</p>
      <p align = "right">This is right aligned</p>
   </body>

</html>
```

Hasil dari html tersebut adalah berikut :

This is left aligned

This is center aligned

This is right aligned

sehingga, konten dan properti dari suatu web, dapat ditemukan dalam kode HTML.

Setelah konsep tersebut dapat dikuasai dengan baik, web scrapping dapat mulai dijalankan. Buku ini akan memberikan contoh bagaimana melakukan web scrapping dari artikel di kompas dengan mempergunakan library BeautifulSoup dan artikel dari detik dengan mempergunakan library Newspaper3k.

2. Beautifulsoup4

BeautifulSoup adalah library python yang dirancang untuk project web-scraping. Keunggulan dari library ini dibanding yang lain adalah :

1. metode-metode yang dipergunakan cukup sederhana sehingga proses navigasi, pencarian dan juga modifikasi struktur data situs yang akan discraping bisa dilakukan dengan mudah.

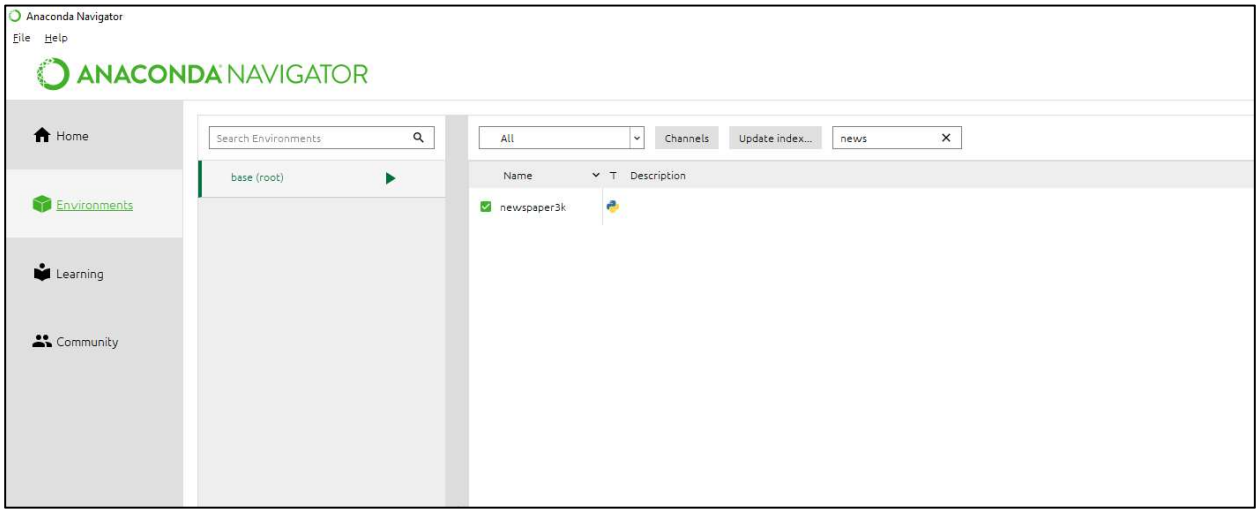
2. mampu menkonversi dokumen ke dalam format UTF-8 secara otomatis
3. Bekerja dengan baik dengan library Python lxml dan html5lib untuk melakukan parsing dokumen.(Pernanda, 2018)

Instalasi

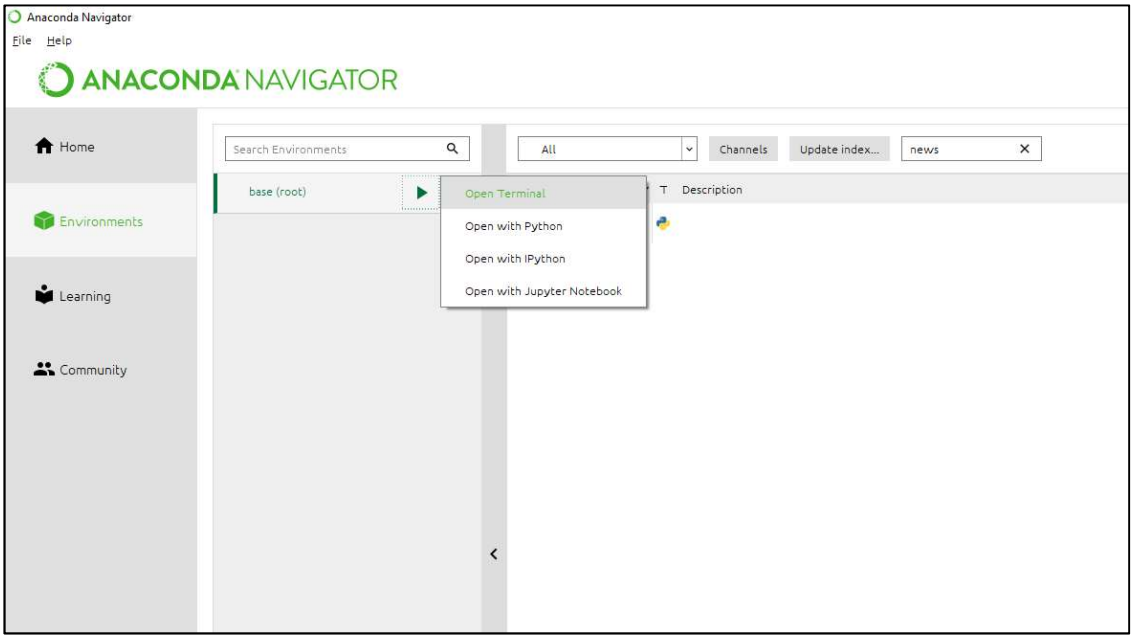
Jika menginstal Anaconda Navigator dan library Pandas, umumnya library ini sudah ikut terinstal, tetapi dalam beberapa kasus, mungkin library perlu diinstal sendiri. Cara untuk menginstal library sebagai berikut :

Instalasi dapat dilakukan melalui pip, untuk mengakses pip, dapat melalui Anaconda-Navigator

1. Klik menu environment,



2. Klik base root lalu open terminal



3. Setelah muncul terminal seperti berikut

Ketikkan perintah berikut didalam terminal

```
pip install beautifulsoup4
```

4. Tunggu hingga proses instalasi selesai

Instalasi parser

Objek dari BeautifulSoup dapat menerima dua argumen. Argumen pertama adalah markup (tag html) yang sebenarnya, dan argumen kedua adalah parser yang ingin dipergunakan. Parser dalam BeautifulSoup adalah : html.parser, lxml, dan html5lib. Parser lxml memiliki dua versi, parser HTML dan parser XML.

Html.parser adalah parser bawaan, sayangnya, parser ini tidak bekerja dengan baik di Python versi lama. Untuk menginstal parser lain menggunakan perintah berikut:

```
pip install lxml
pip install html5lib
```

Parser lxml sangat cepat dalam mengurai HTML yang diberikan dan tidak mengkonsumsi memori yang cukup banyak. Parser ini juga cukup bagus untuk menangani markup html yang tidak lengkap. Sedagkan parser html5lib sangat lambat ketika memproses.

Langkah-langkah untuk scrapping kompas dengan mempergunakan BeautifulSoup adalah

1. Import library pandas

```
import pandas as pd
```

2. Import BeautifulSoup

```
import urllib
from bs4 import BeautifulSoup
```

3. Buat fungsi bernama getUrl yang menerima 1 parameter (parameter ini antinya akan diisi oleh url dari situs yang ingin discrap)

```
# langkah 1
# membuat fungsi getUrl yang menerima 1 parameter url, dialiaskan menjadi a

def getUrl(a):

# langkah 2
# inisialisasi 4 variabel yang bertype array, yang bernama link_berita, judul_berita, tanggal_berita
# dan kategori yang akan dipergunakan untuk menampung sementara hasil scrap

  link_berita = []
  judul_berita = []
  tanggal_berita = []
  kategori=[]

# membuka url (yang dialiaskan menjadi a) mempergunakan urllib.request dan menyimpan HTML yang didapatkan
# kedalam  variabel  url

  url = urllib.request.urlopen(a)
  result = url.read()
  url.close()

# memparsing HTML dari url mempergunakan parser HTML
  soup = BeautifulSoup(result, "html.parser")

# mencari elemen html h3, untuk kompas, elemen html <h3> mengindikasikan judul
  for judul in soup.find_all('h3'):

# mencari elemen html h3, untuk kompas, elemen html <h3> mengindikasikan judul
    for l in judul.find_all('a'):

# mengambil link dan menambahkan hasilnya pada variabel link_berita
      link_berita.append(l.get('href'))
    judul_berita.append(judul.get_text())
  for kateg in soup.find_all('div',class_='article__subtitle'):
    kategori.append(kateg.get_text())
  for tanggal in soup.find_all('div',class_='article__date'):
    tanggal_berita.append(tanggal.get_text())

  print(judul_berita)
  print(link_berita)
  print(tanggal_berita)
  print(kategori)
  return judul_berita, link_berita, tanggal_berita, kategori, soup
```

4. Membuaat variabel bernama judul_berita yang berupa dataframe dari pandas untuk menampung hasil scrapping

```
judul_berita = pd.DataFrame(columns=['judul','link','tanggal','kategori'])
```

5. Menentukan url dari artikel yang ingin diekstrak dan menjalankan fungsi getUrl

```
kompas='https://indeks.kompas.com/?site=all&date=2020-10-30'
judul, link, tanggal, kategori, semua = getUrl(kompas)
for a in range(len(judul)):
  judul_berita = judul_berita.append(pd.DataFrame([[judul[a],
    link[a],tanggal[a],kategori[a]]],
    columns=['judul','link','tanggal','kategori']),ignore_index=True)
```

Sebagian output dari baris perintah diatas adalah sebagai berikut
(output dari fungsi getUrl bisa bervariasi, tergantung tanggal yang dipilih, ingin menscrape tanggal berapa, dalam studi kasus ini, fungsi getUrl akan mesncrap data pada 30 Oktober 2020)

```
['Hari Ketiga Libur Panjang, Jumlah Penumpang di Kampung Rambutan Menurun', '5 Wisatawan Diminta Pulang Usai Jalani Rapid Test
di Stasiun Bogor', 'Indonesia's Maluku Islands Ready to Become a Plastic-Free Tourism Destination', 'Politisi PAN Minta Pemerin
tah Tak Sepihak Putuskan UMP 2021', 'Libur Panjang, 83.257 Volume Kendaraan Melintas di Tol Pemalang-Batang', 'Over Half Millio
n Vehicles Leave Jakarta for Long Weekend ', 'Masuk Riau Diperketat, Penumpang Wajib Cek Suhu Tubuh di Pos Jaga', 'Pandangan Al
isson soal Tekel Horor Jordan Pickford', 'Update 30 Oktober: Covid-19 di Kota Tangerang Kini 2.159, Bertambah 19 Kasus', 'Qatar
Akan Tuntut Pegawai yang Periksa Wanita Telanjang di Bandara', 'Max Kilman Hobi Berduel dengan "Serigala" Kekar ', 'BNPB Minta
Depok Segera Isolasi OTG Covid-19 di Lokasi Khusus', 'Gudang Rosok Dibongkar Pencuri, Beberapa Kuintal Barang Bekas Hilang ',
'Elf Berpenumpang 14 Orang Masuk Jurang, Bocah 9 Tahun Tewas, Belasan Lain Luka', 'Lirik dan Chord Lagu I Want You Around - Sno
h Aalegra', 'Ibu Pelaku Teror Penyerangan Pisau di Perancis Menangis dan Terkejut atas Perbuatan Anaknya', 'Soal Pembubaran Dek
larasi KAMI di Jambi, Ini Penjelasan Gugus Tugas', 'Andrea Pirlo Ciptakan Versi Terburuk Juventus dalam 1 Dekade Terakhir', 'KP
U Tetapkan Jadwal Debat Kandidat Pilkada Tangsel: 22 November dan 3 Desember 2020', 'Libur Panjang, 509.140 kendaraan Tinggalka
n Jakarta', 'Lirik dan Chord Lagu Partition dari Beyonce', 'Kelly Tandiono Soroti Fasilitas Bersepeda di Jalanan Dalam Kota',
'Penjualan Sepeda Turun 30 Persen pada Agustus-Oktober 2020', 'Bulan November, KSPI Lakukan Rentetan Aksi Unjuk Rasa', 'Demo An
ti-Perancis Menjalar ke Bangladesh, Pakistan, dan Afghanistan', 'Kebiasaan Khabib Kala Bertarung Dirasakan Langsung oleh Conor
McGregor', 'Makan Buah Kiwi Efektif Atasi Sembelit, Sudah Tahu?', '[UPDATE] Grafik Covid-19 30 Oktober: Total 7.116 Kasus di De
pok', 'Bersepeda di Masa Pandemi, Kelly Tandiono Batasi Rombongan Tak Lebih dari 10 Orang', 'Karyawati SPBU Tewas dengan Tubuh
Terluka di Jalan, Keluarga Duga Dibunuh', 'Bicara tentang Menghargai Pasangan, Bagas HP Lepas Singel Stay Here Love', 'Mengapa
Emas Jadi Cara Berinvestasi Terbaik Saat Pandemi? ', 'Klaster Pilkada Purbalingga Meluas, dari Paslon, Tim Sukses, hingga Petug
as KPU Positif Covid-19', 'Pria yang Merusak Motornya dengan Batu Besar Akhirnya Ditilang', 'TREASURE Bakal Tampil di MAMA 202
0?', 'Lirik dan Chord Lagu All Bad dari Justin Bieber', 'Polemik Lokasi Khusus Isolasi OTG Covid-19 di Depok, Kini BNPB Sebut W
isma Makara UI Boleh Dipakai', 'BPBD Catat 31 Desa dan 10 Kecamatan Terdampak Banjir Kebumen', 'Ingin Naik Gunung Saat Libut Pa
njang, Simak Info 6 Gunung Ini', 'Jelang Laga Alaves Vs Barcelona, Ter Stegen dan Umtiti Sudah Kembali Berlatih']
```

6. Menyimpan output kedalam file csv bernama scrap_judul.csv

```
judul_berita.to_csv('scrap_judul.csv',index=False)
```

7. Ketika ditampilkan isi dari judul_berita

```
judul_berita
```

	judul	link	tanggal	kategori
0	Palestina Desak PBB Bahas Perdamaian di Timur ...	https://www.kompas.com/global/read/2020/10/29/...	29/10/2020, 23:36 WIB	GLOBAL
1	Program Garuda Select Kembali Bergulir dan Sud...	https://bola.kompas.com/read/2020/10/29/233034...	29/10/2020, 23:30 WIB	BOLA
2	Indonesia Highlights: US Secretary of State Po...	https://go.kompas.com/read/2020/10/29/23275907...	29/10/2020, 23:27 WIB	GO
3	Foreign Ships Continue Illegal Fishing in Indo...	https://go.kompas.com/read/2020/10/29/23023587...	29/10/2020, 23:02 WIB	GO
4	Indonesia's Third Wealthiest Person Tan Siok T...	https://go.kompas.com/read/2020/10/29/22545807...	29/10/2020, 22:54 WIB	GO
...	...	...	...	...
75	Lirik dan Chord Lagu All Bad dari Justin Bieber	https://www.kompas.com/hype/read/2020/10/30/21...	30/10/2020, 21:50 WIB	HYPE
76	Polemik Lokasi Khusus Isolasi OTG Covid-19 di ...	https://megapolitan.kompas.com/read/2020/10/30...	30/10/2020, 21:42 WIB	NEWS
77	BPBD Catat 31 Desa dan 10 Kecamatan Terdampak ...	https://regional.kompas.com/read/2020/10/30/21...	30/10/2020, 21:42 WIB	NEWS
78	Ingin Naik Gunung Saat Libut Panjang, Simak In...	https://travel.kompas.com/read/2020/10/30/2140...	30/10/2020, 21:40 WIB	TRAVEL
79	Jelang Laga Alaves Vs Barcelona, Ter Stegen da...	https://bola.kompas.com/read/2020/10/30/214000...	30/10/2020, 21:40 WIB	BOLA

80 rows × 4 columns

3. Newspaper3k

Library ini dapat dipergunakan untuk mengekstrak data berita dari berbagai sumber, tanpa perlu memeriksa elemen html dari masing-masing sumber data, dan menyesuaikan elemen apa saja yang akan diambil, kemudian membangun script berbasis BeautifoulSoup. Banyak waktu yang bisa dipangkas, dibandingkan jika harus membuat satu script yang berbeda untuk masing-masing situs berita. Newspaper3k dapat menscrap artikel berita dari manapun, mengekstrak informasi, dan juga meringkas informasi tersebut. Terdapat juga opsi untuk mengakses teks lengkap dari suatu artikel.

Instalasi

Instalasi dapat dilakukan melalui pip, untuk mengakses pip, dapat melalui anaconda-navigator. Klik base root lalu open terminal. Kemudian masukkan perintah berikut

```
pip install newspaper3k
```

Langkah-langkah untuk scrapping kompas dengan mempergunakan Newspaper3k adalah

1. Import library pandas dan Article dari newspaper

```
from newspaper import Article
import pandas as pd
```

2. Buat variabel bernama data_berita yang bertipe dataframe untuk menampung hasil scrap

```
data_berita = pd.DataFrame(columns=['judul', 'penulis','tanggal','isi'])
```

3. Mulai lakukan scrapping

```
# url dari situs yang akan discrap dimasukkan dalam variabel alamat
alamat='https://news.detik.com/berita-jawa-timur/d-5207658/pdip-resmi-pecat-yusuf-widyatmoko-dari-keanggotaan-partai'

# inisialisasi object Article dari satu alamat yang kemudian akan disimpan dalam variabel berita
berita=Article(alamat)

# perintah untuk memparsing artikel
berita.download()
berita.parse()

# perintah untuk mengekstrak informasi  mengekstrak informasi.
# informasi yang akan diambil adalah judul, penulis, tanggal publikasi, teks berita
print('Judul:', berita.title)
print('Penulis:', berita.authors)
print('Tanggal publikasi:', berita.publish_date)
print('Teks berita:', berita.text)
isi_berita=[[berita.title, berita.authors, berita.publish_date, berita.text]]
data_berita = data_berita.append(pd.DataFrame(isi_berita,
            columns=['judul', 'penulis','tanggal','isi']),ignore_index=True)
```

Hasil dari script diatas adalah sebagai berikut (hasil bisa bervariasi, tergantung url yang dipergunakan untuk scrapping)

```
Judul: PDIP Resmi Pecat Yusuf Widyatmoko dari Keanggotaan Partai
Penulis: ['Ardian Fanani', 'Https', 'Www.Facebook.Com Detikcom', 'Ardian Fanani - Detiknews']
Tanggal publikasi: 2020-10-09 22:34:43
Teks berita: PDIP resmi memecat Yusuf Widyatmoko dari keanggotaan partai. Melalui Surat Keputusan DPP PDIP Nomor 63/KPTS/DPP/X/
2020, yang ditandatangani Ketua Umum PDIP Megawati Soekarnoputri dan Sekjen PDIP Hasto Kristiyanto, per 1 Oktober 2020 Yusuf ti
dak lagi berstatus anggota.

Surat pemecatan ini dibacakan langsung oleh Sekretaris DPD PDIP Jawa Timur Untari Bisowarno, di Kantor DPC PDIP Banyuwangi, Jal
an Jaksa Agung Suprapto.

"PDI Perjuangan memberikan sanksi organisasi berupa pemecatan kepada Yusuf Widyatmoko dari keanggotaan partai," kata Untari mem
bacakan surat keputusan tertanggal 1 Oktober 2020, Jumat (9/10/2020).

PDIP menilai, tindakan Yusuf tidak mengindahkan instruksi partai terkait rekomendasi calon Bupati dan Wakil Bupati Banyuwangi p
ada Pilkada Serentak tahun 2020. Di mana Yusuf telah mencalonkan diri dari partai PKB dan Demokrat.
```

Berikut adalah potongan berita asli dari detik

PDIP Resmi Pecat Yusuf Widyatmoko dari Keanggotaan Partai

Ardian Fanani - detikNews

Jumat, 09 Okt 2020 22:34 WIB

2 komentar

SHARE

Surat pemecatan Yusuf dibacakan langsung oleh Sekretaris DPD PDIP Jawa Timur Untari Bisowarno/Foto: Ardian Fanani/detikcom

Banyuwangi - PDIP resmi memecat Yusuf Widyatmoko dari keanggotaan partai. Melalui Surat Keputusan DPP PDIP Nomor 63/KPTS/DPP/X/2020, yang ditandatangani Ketua Umum PDIP Megawati Soekarnoputri dan Sekjen PDIP Hasto Kristiyanto, per 1 Oktober 2020 Yusuf tidak lagi berstatus anggota.

Surat pemecatan ini dibacakan langsung oleh Sekretaris DPD PDIP Jawa Timur Untari Bisowarno, di Kantor DPC PDIP Banyuwangi, Jalan Jaksa Agung Suprapto

Hasil scrapping dengan mempergunakan Newspaper3k cukup bagus, tidak ada data yang salah dan tidak memerlukan script sepanjang BeautifulSoup, hanya perlu beberapa baris saja.

4. Menyimpan hasil scrapping ke dalam file csv

```
data_berita.to_csv('scrap_berita.csv',index=False)
```

Ekstraksi Data Menggunakan Penerapan API Twitter

Twitter adalah layanan bagi teman, keluarga, dan teman sekerja untuk berkomunikasi dan tetap terhubung melalui pertukaran pesan yang cepat dan sering. Pengguna memposting *Tweet*, yang dapat berisi foto, video, tautan, dan teks. Pesan ini diposting ke profil pengguna, terkirim ke pengikut, dan dapat dicari di pencarian *tweet* (*What Is Twitter and Why Should You Use It? - Economic and Social Research Council*, n.d.). Dengan menggunakan *Twitter*, penggunanya bisa menuliskan *tweet* tentang apa yang sedang dipikirkan dan bisa direspon oleh pengguna lainnya. Selain itu, pengguna *Twitter* bisa saling berkirim pesan melalui *Direct Message* kepada pengguna lain. Fitur yang paling menarik dari *Twitter* adalah *Trending Topic* yakni hal yang sedang ramai dibicarakan oleh para pengguna berdasarkan lokasi sehingga kita bisa mengetahui hal apa yang sedang ramai dibicarakan di berbagai lokasi.

Awalnya pengguna *Twitter* hanya bisa menuliskan *tweet* sepanjang 140 karakter disesuaikan dengan batasan karakter SMS (*short message service*) yang hanya 160 karakter(*What Is Twitter and Why Should You Use It? - Economic and Social Research Council*, n.d.). Batasan 140 karakter tersebut diputuskan karena *Twitter* memiliki fitur SMS *Twitter* sehingga bisa mempermudah penggunanya untuk menuliskan *tweet* melalui SMS. Namun mulai tahun 2017, *Twitter* resmi mengumumkan penambahan batasan karakter untuk *tweet* sebanyak 280 karakter. Hal itu dilakukan karena banyaknya keluhan pengguna yang harus menyunting *tweet* agar dapat muat dalam 140 karakter sehingga para pengguna *Twitter* kurang leluasa dalam menuliskan *tweet* (What Is Twitter and Why Should You Use It? - Economic and Social Research Council, n.d.).

Tweet dapat ditemukan dengan mencari di mesin pencarian dengan menggunakan keyword atau hashtag tertentu. Informasi yang tersebar di Twitter pun sangat cepat dan efektif, karena pengguna nya yang sangat banyak. Banyak manfaat yang didapatkan dengan memanfaatkan Twitter seperti memberikan informasi dari pemerintah untuk masyarakat, melihat trend yang sedang terjadi, menambah jaringan sosial pertemanan karena Twitter tidak memiliki batasan pertemanan. Dengan semua fasilitas yang disediakan oleh twitter ini, banyak pihak yang dapat menemukan informasi lebih cepat yang dapat dianalisis lebih lanjut. Metode ini cukup tepat dipergunakan untuk mendapatkan opini/pendapat dari masyarakat terkait dengan suatu hal, secara cepat, sekaligus juga dapat menjangkau responden lebih banyak, jika dibandingkan dengan metode seperti kuesioner atau penyebaran angket (Mas'udah et al., 2020). Secara garis besar, data dari twitter dapat diperoleh dengan mempergunakan API twitter yang sudah disediakan.

Twitter telah menyediakan API REST yang dapat digunakan oleh developer/programmer untuk mengakses dan membaca data Twitter. Mereka juga menyediakan Streaming API yang dapat digunakan untuk mengakses Data Twitter secara real-time. Sebagian besar perangkat lunak yang dibangun untuk mengakses data Twitter menyediakan library yang memanfaatkan API Twitters Search dan Streaming. Oleh karena itu, batasan dari API ini otomatis juga berlaku pada perangkat lunak tersebut. batasan dari API Twitters Search adalah kita hanya dapat mengirim 180 Permintaan setiap 15 menit. Dengan jumlah maksimum 100 tweet per Permintaan, ini berarti kita dapat mengambil 4 x 180 x 100 = 72.000 tweet per jam. Salah satu kelemahan yang lebih besar dari API Twitters Search adalah kita hanya dapat mengakses Tweet yang ditulis dalam 7 hari terakhir. Ini adalah hambatan utama bagi siapa saja yang mencari data masa lalu (beberapa bulan atau tahun terakhir) untuk membuat model. Selain API Twitters Search dan Streaming, beberapa API paling terkenal yang disediakan oleh Twitter antara lain :

a. Tweets: pencarian, posting, pemfilteran, keterlibatan, streaming dll.

b. Ads: untuk mengelola kampanye dan audiens, analitik.

c. Pesan langsung (masih dalam Beta): mengirim dan menerima, balasan langsung, pesan sambutan dll.

d. Akun dan pengguna (Beta): manajemen akun, interaksi pengguna.

e. Media: mengunggah dan mengakses foto, video, dan GIF animasi.

f. Tren: trending topik di lokasi tertentu.

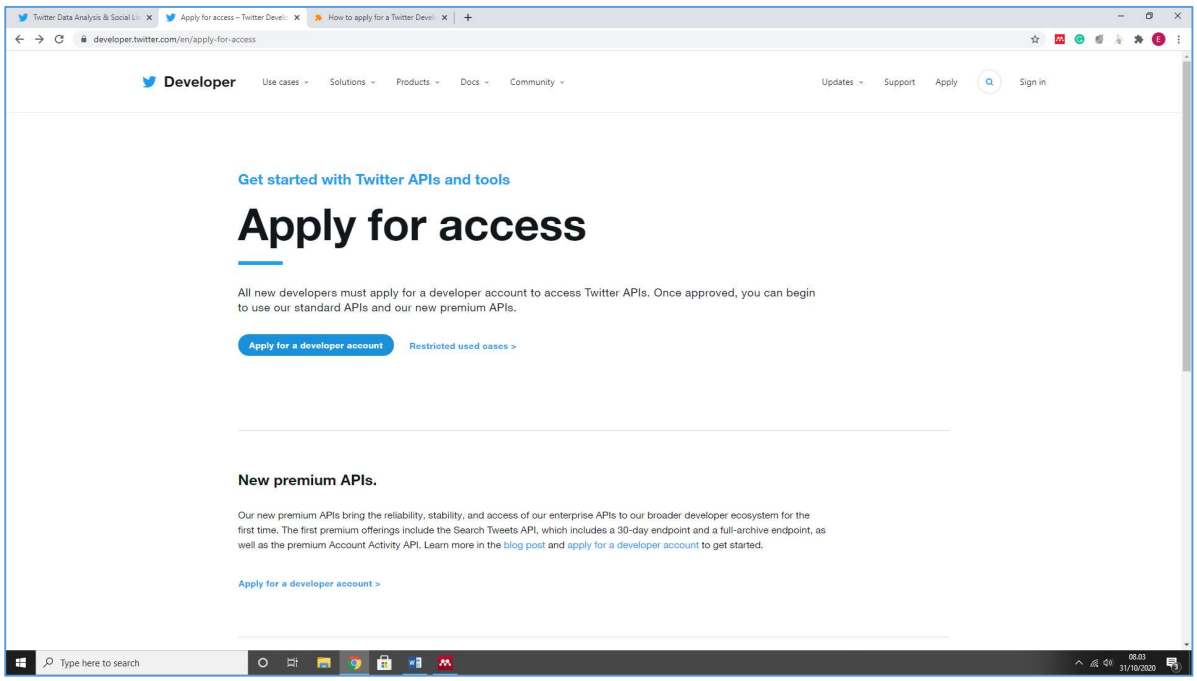
g. Geo: informasi tentang tempat yang dikenal atau tempat di dekat lokasi.

Untuk dapat menggunakan twitter API, syaratnya adalah mendaftar sebagai pengembang twitter melalui tautan https://developer.twitter.com/en/apply-for-access. Hal ini dilakukan untuk mendapatkan Twitter API Keys yang digunakan untuk mengakses twitter API.

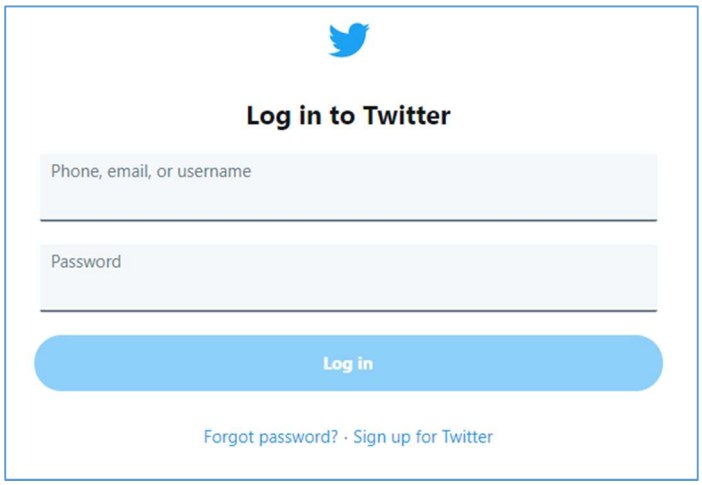
1. Membuat kredensial developer twitter

Berikut tahapan untuk Mendapatkan Kredensial dari developer twitter supaya bisa menggunakan twitter API.

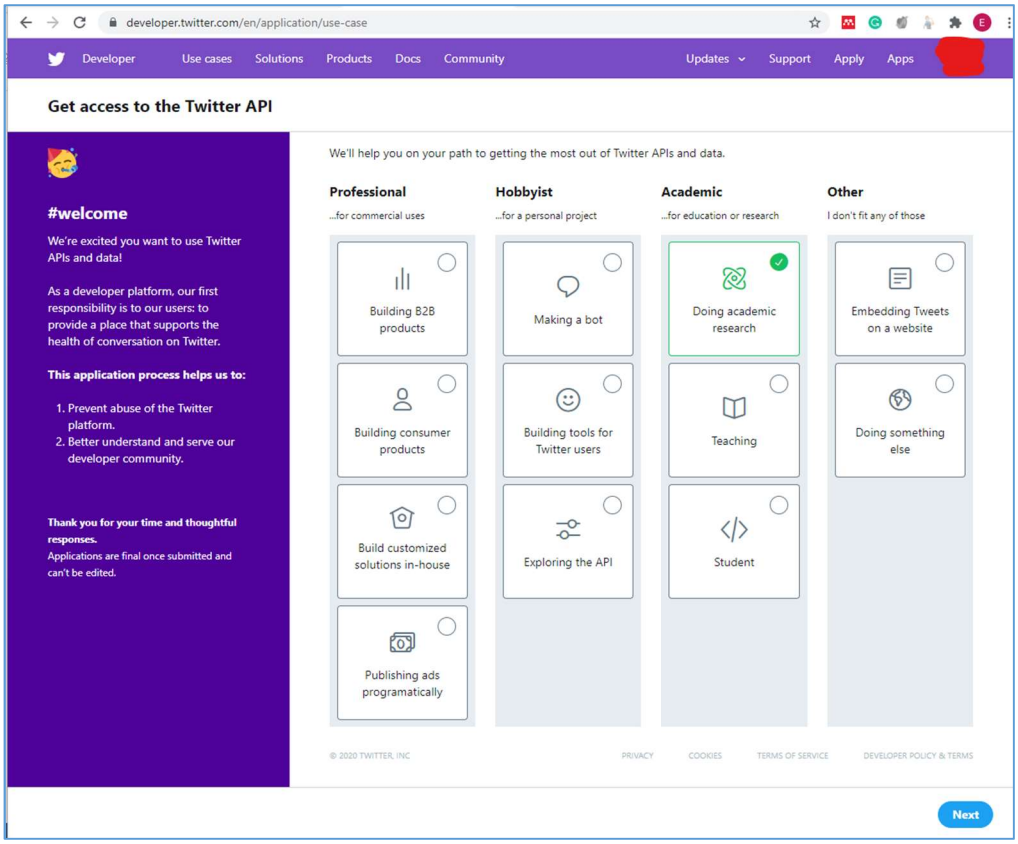
1. klik tombol “apply for a developer account”



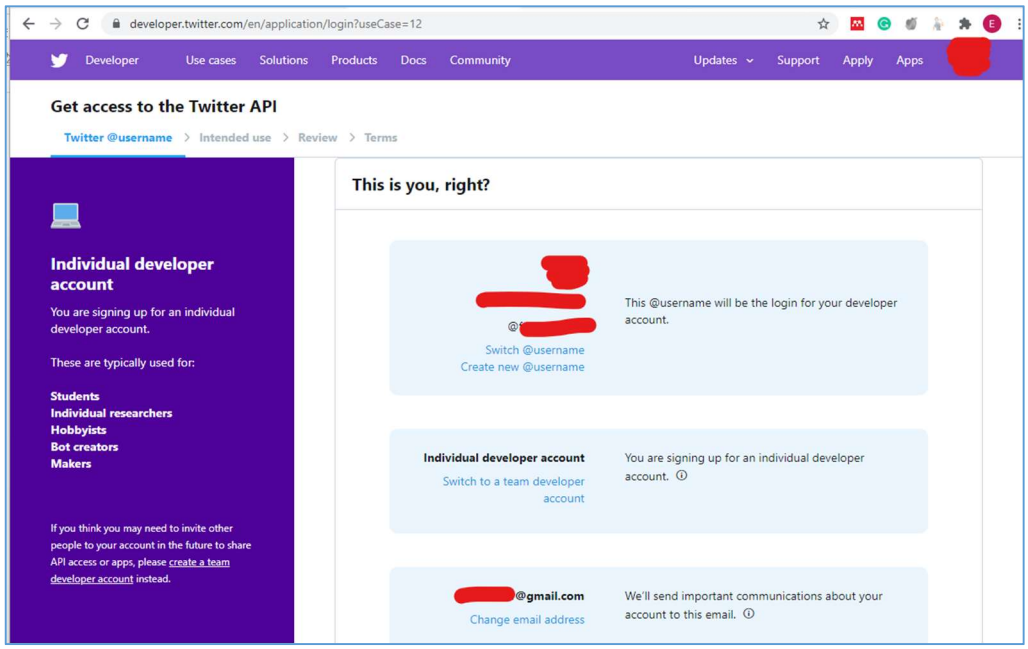
2. agar dapat memperoleh akun developer, dsebelumnya disyaratkan harus memiliki akun di twtitter. Login terlebih dahulu dengan mempergunakan akun anda masing-masing



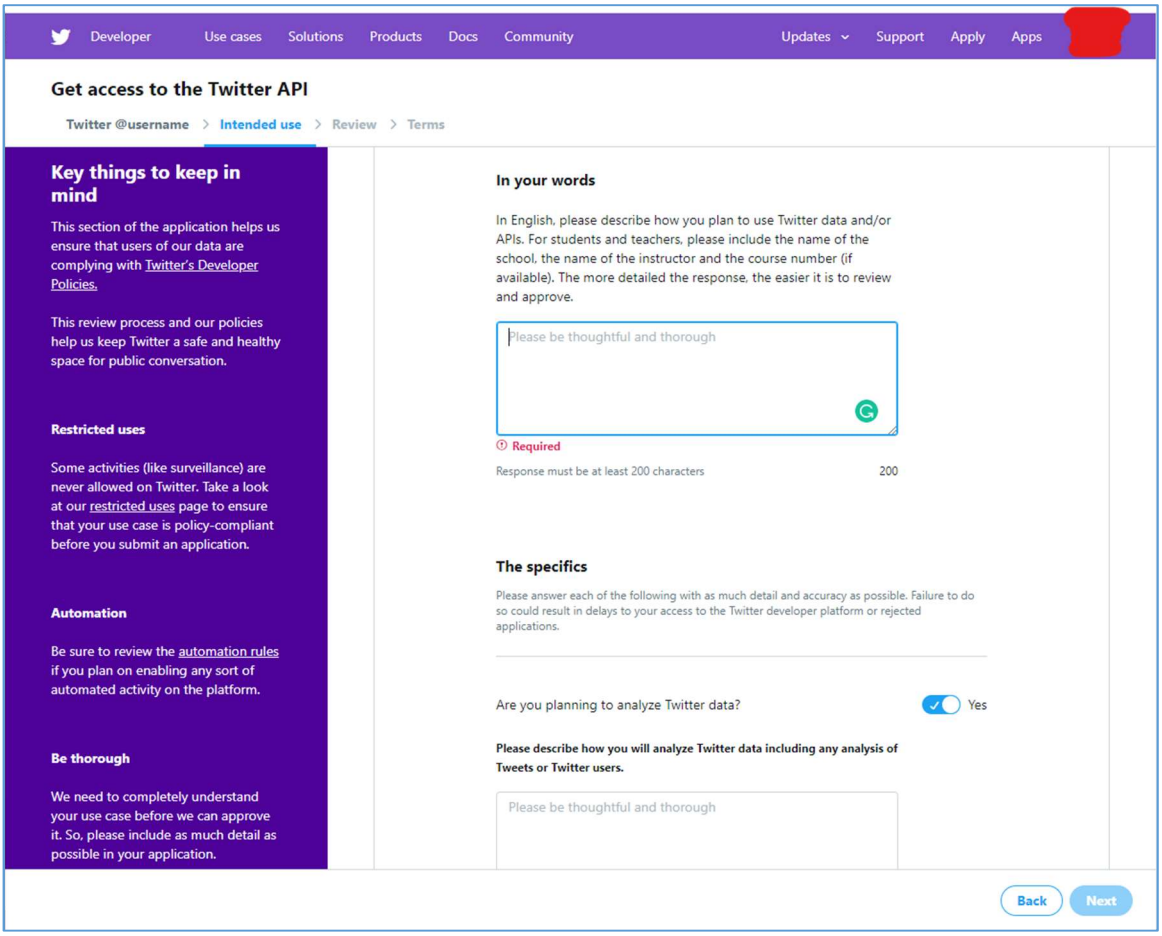
3. Pilih alasan utama menggunakan Twitter Developer Tools



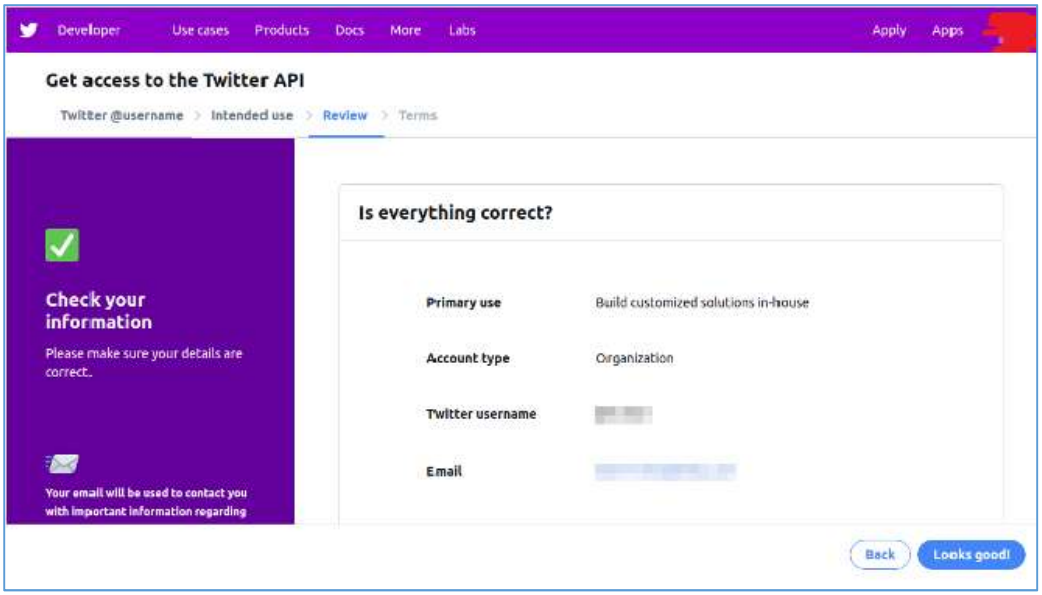
4. Verifikasi detail Nama Pengguna Twitter yang terkait dengan akun pengembang



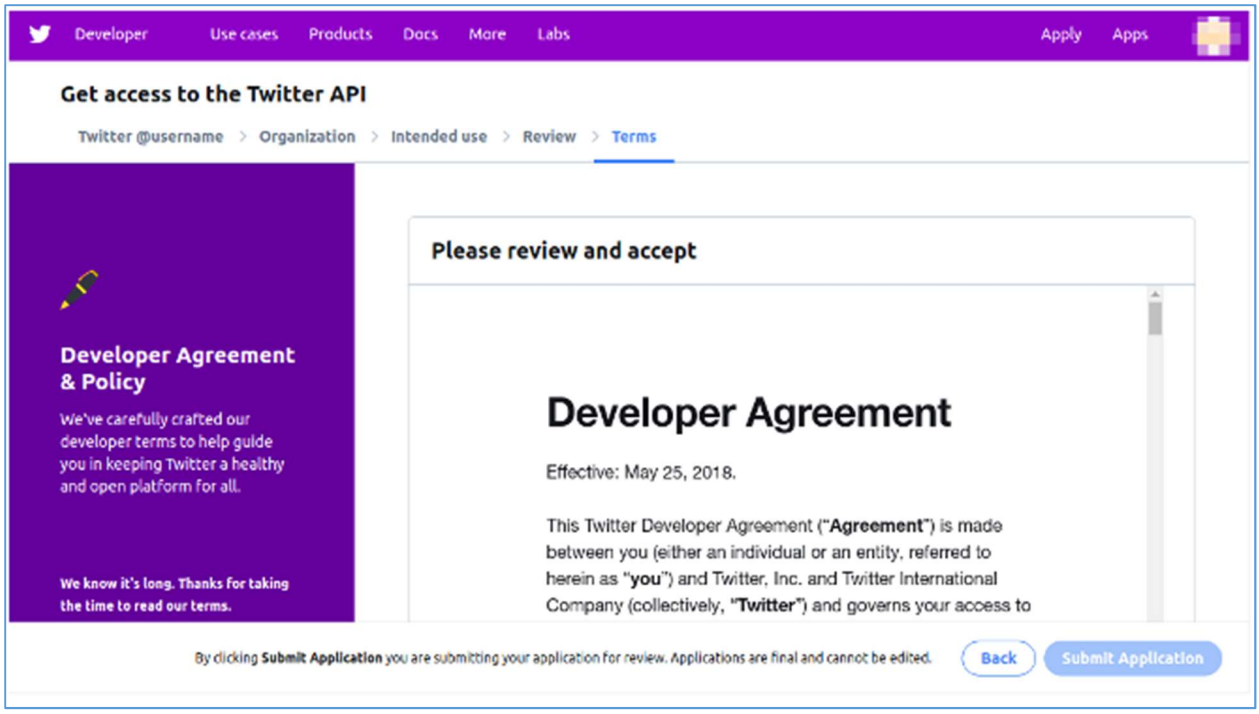
5. Jelaskan apa tujuan penggunaan API Twitter yang akan diajukan



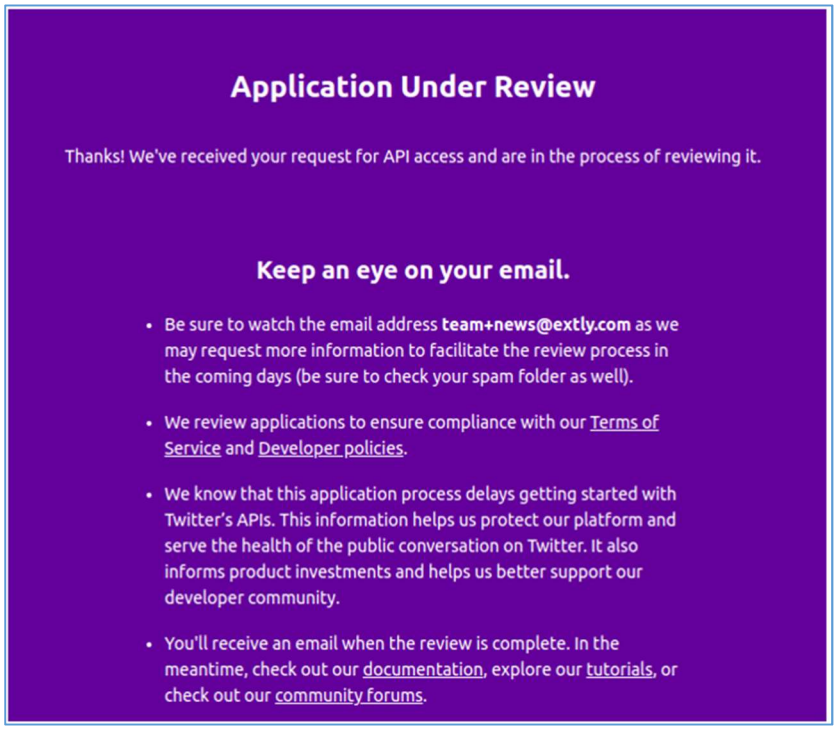
6. Periksa informasi yang telah diisikan



7. Setujui Developer Agreement dan verifikasi akun email yang telah dimasukkan



8. Ajuan sedang ditinjau, dan akan segera menerima pemberitahuan dengan hasilnya.



Python adalah salah satu bahasa pemrograman dengan jumlah library terbanyak yang dikembangkan untuk Twitter API. Daftar library yang memanfaatkan Twitter API dilihat dari jumlah kontributor, jumlah bintang yang diterima, jumlah pengamat, "umur library" sejak rilis

pertama dll. Karena tweepy menempati posisi teratas, buku ini akan memberikan contoh bagaimana cara menggunakan tweepy untuk melakukan ekstraksi data twitter.

Tweepy merupakan library python yang digunakan untuk mengakses Twitter API yang salah satunya adalah untuk ekstraksi tweets. Tweepy memiliki beberapa keterbatasan diantaranya hanya mampu melakukan ekstraksi tweet maksimal 7 hari ke belakang dan terbatas melakukan ekstraksi 18000 tweets per 15 menit. Selain itu, Tweepy hanya mampu mendapatkan 3200 tweet terbaru dari seorang pengguna.

Instalasi tweepy

Instalasi dapat dilakukan melalui pip, untuk mengakses pip, dapat melalui anaconda-navigator. Klik base root lalu open terminal. Kemudian masukkan perintah berikut

```
pip install tweepy
```

2. Ekstraksi twitter dengan API Search

API ini akan mengambil data dari twitter melalui pencarian atau nama pengguna. API ini akan memberikan akses ke kumpulan data yang **sudah ada** dari **tweet yang sudah terjadi**. Melalui API Search, pengguna meminta data tweet yang cocok dengan suatu kriteria "pencarian". Kriteria dapat berupa kata kunci, nama pengguna, lokasi, nama tempat, dll. Ilustrasi dari API Search ini adalah pengguna akan melakukan pencarian secara langsung di Twitter (menavigasi ke search.twitter.com dan memasukkan kata kunci).

Dengan API Search, pengembang meminta data tweet yang **telah terjadi** dan dibatasi oleh Twitter. Untuk pengguna individu, jumlah maksimum tweet yang dapat kita terima adalah 3,200 tweet paling baru, terlepas dari kriteria kueri. Berikut langkah-langkah untuk mengambil data mempergunakan tweepy

1. Import library yang diperlukan

```
import pandas as pd
import tweepy
import json
```

2. Masukkan kredensial developer twitter

```
consumer_key=
consumer_secret=
access_token =
access_token_secret =

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth,wait_on_rate_limit=True,wait_on_rate_limit_notify=True)
```

3. Tentukan keyword yang ingin dicari dan jalankan twitter api.search

```
hashtag_yang_dicari='#belajardarirumah'

crawl_tweet=api.search(q=hashtag_yang_dicari,count=100,include_entities=False)
```

4. Ambil tweet dalam bentuk json

```
json_data = [r._json for r in crawl_tweet]
df = pd.json_normalize(json_data)
```

5. Cek apakah variabel df sudah terisi

```
df.head()
```

	created_at	id	id_str	text	truncated	
0	Sat Oct 31 00:10:50 +0000 2020	1322330150491443200	1322330150491443200	RT @TVRINasional: Jadwal Siaran dan Materi Pem...	False	href="http://twitter.com/dc
1	Fri Oct 30 23:45:22 +0000 2020	1322323743297867776	1322323743297867776	RT @Kemdikbud_RI: Selamat pagi, #SahabatDikbud...	False	href="http://twitter.com/dc
2	Fri Oct 30 23:43:10 +0000 2020	1322323187963604992	1322323187963604992	RT @Kemdikbud_RI: Selamat pagi, #SahabatDikbud...	False	href="http://twitter.com/dc
3	Fri Oct 30 23:35:08 +0000 2020	1322321167328243712	1322321167328243712	Selamat pagi, #SahabatDikbud. Di hari terakhir...	True	href="http://twitter.com/dc
4	Fri Oct 30 22:52:56 +0000 2020	1322310547459641344	1322310547459641344	Jadwal Siaran dan Materi Pembelajaran untuk Pr...	True	href="http://twitter.com/dc

5 rows × 212 columns

6. Simpan data dalam bentuk csv supaya mudah diakses kembali

```
#menyimpan menjadi csv
df.to_csv("data_tweeet.csv",index=False)
```

Latihan

1. Lakukan ekstraksi data tweet, dan dokumentasikan langkah-langkah yang anda lakukan dalam word
2. Lakukan ekstraksi data dari artikel (sumber artikel bebas, bisa dari medium, surat kabar dll)