

Python

Data Crawling (Twitter)

Outline

- Pengantar
- Persiapan
- Mengirimkan data ke twitter
- Mengambil data dari twitter
- Visualisasi

Data Crawling dari Twitter (2)

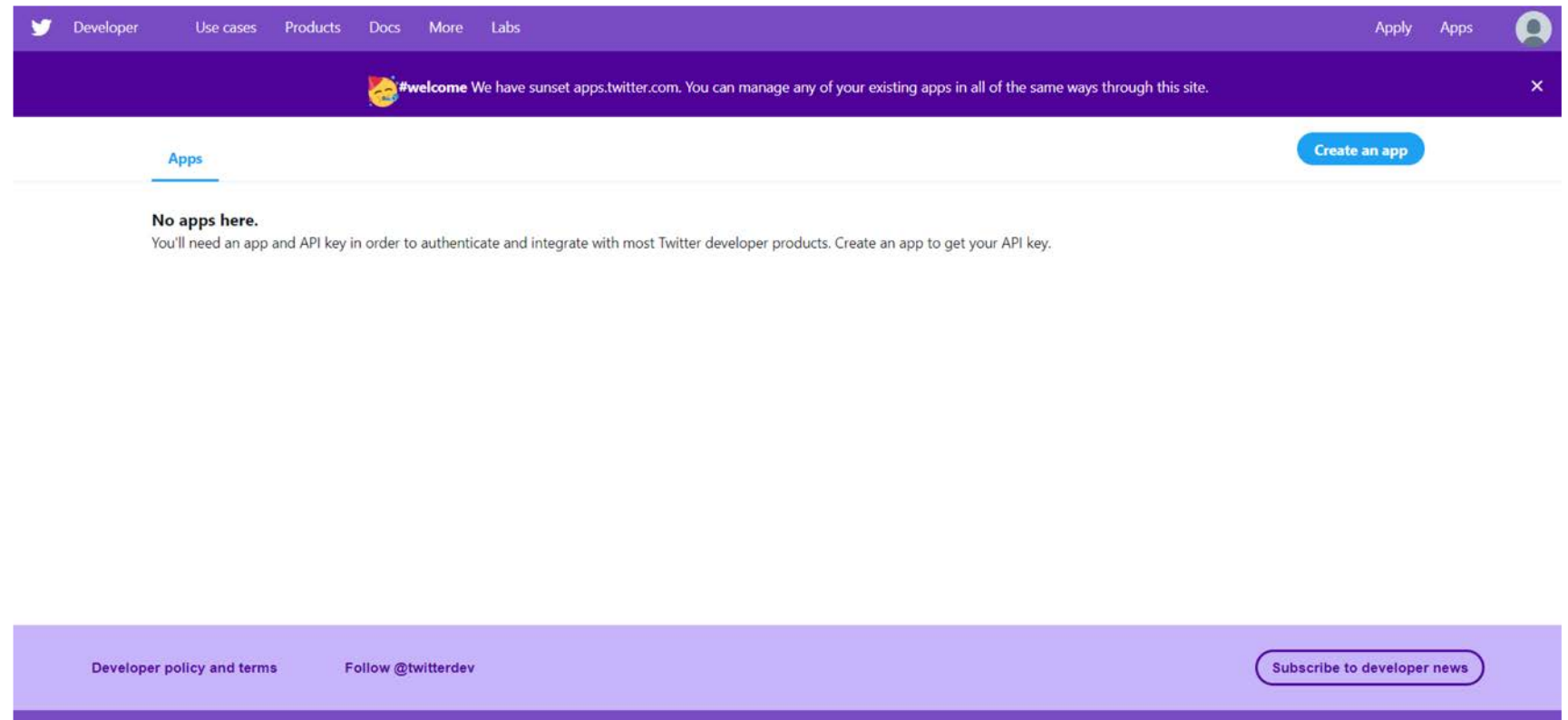
- Cara mengolah data dari twitter tidak berbeda dari tahapan pada *data crawler*.
- Tahap pertama, seorang data analis harus memiliki beberapa *key* yang memungkinkan koneksi ke API yang disediakan oleh Twitter.
- Data analis menggunakan *key* tersebut pada bahasa pemrograman yang dipilih untuk menganalisa data-data twitter.

Data Crawling dari Twitter (3)

- Untuk mendapatkan *access key* ke twitter, data analis harus membuat aplikasi yang berinteraksi dengan Twitter API.
- Langkah pertama adalah mendaftarkan aplikasi ke Twitter. Linknya adalah: <https://developer.twitter.com/en/apps>
- Login menggunakan akun twitter atau buat akun twitter jika belum memilikinya.

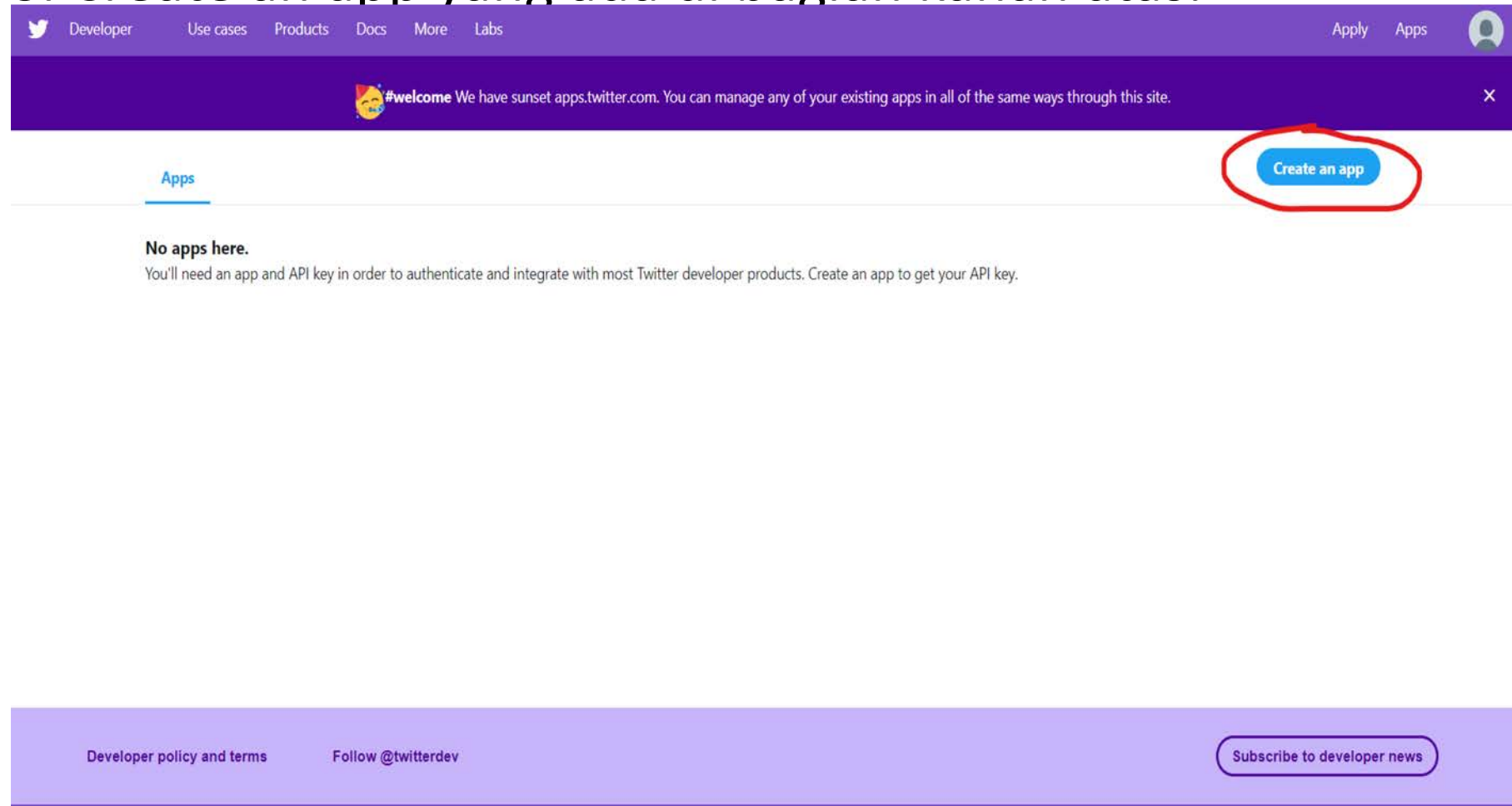
Python – Twitter Step-by-Step (1)

- Registrasi aplikasi di twitter.
 - Di browser, masukkan url: <https://developer.twitter.com/en/apps>
 - Tampilannya:



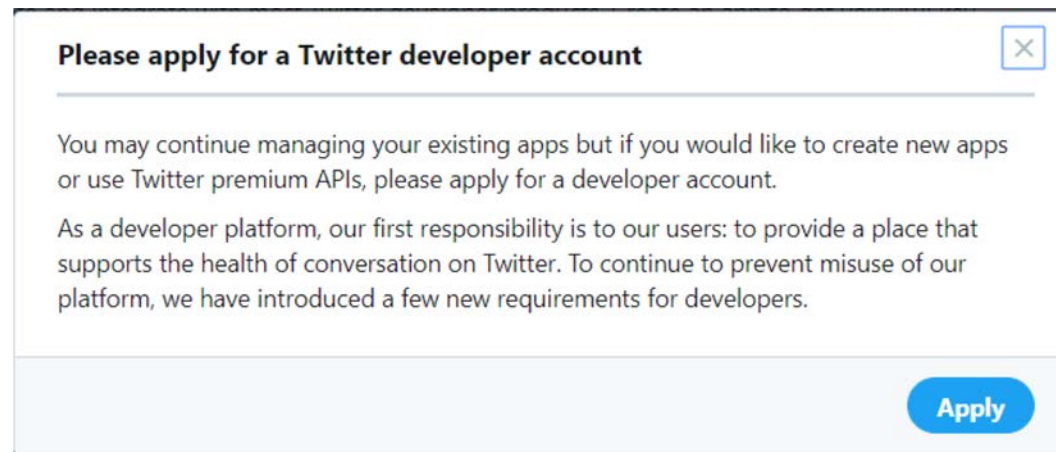
Python – Twitter Step-by-Step (2)

- Klik tombol Create an app yang ada di bagian kanan atas.



Python – Twitter Step-by-Step (3)

- Saat di klik, akan muncul sebuah *window* seperti berikut:



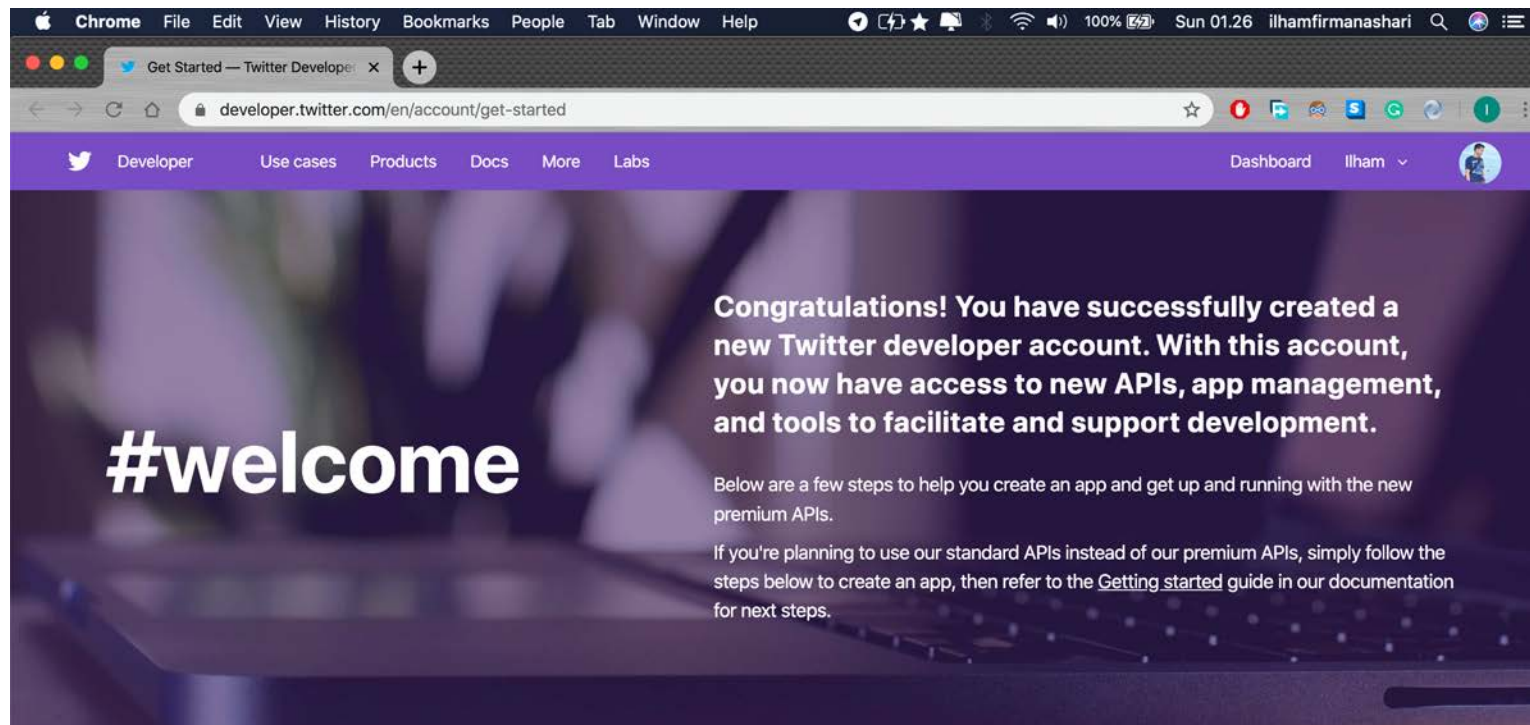
- *Window* ini akan mengkonfirmasi akun twitter yang baru pertama kali ingin membuat aplikasi.
- Klik tombol Apply untuk mengkonfirmasi pembuatan aplikasi yang pertama kali.

Python – Twitter Step-by-Step (4)

- Ikuti instruksi pada langkah-langkah berikutnya untuk mendapatkan akses developer di twitter.
- Bila pengisian survey untuk mendapatkan akses developer berhasil, Twitter akan mengirimkan email konfirmasi ke email yang terdaftar di twitter.
- Klik link yang terdapat di dalam email tersebut untuk menyelesaikan proses pendaftaran sebagai akun developer di twitter.

Python – Twitter Step-by-Step (5)

- Bila berhasil, akan tampil halaman sebagai berikut:



Python – Twitter Step-by-Step (6)

- *Scroll down* pada halaman tersebut untuk menemukan link pembuatan aplikasi.
- Klik link pembuatan aplikasi untuk mendapat twitter key yang akan digunakan dalam pemrograman twitter.

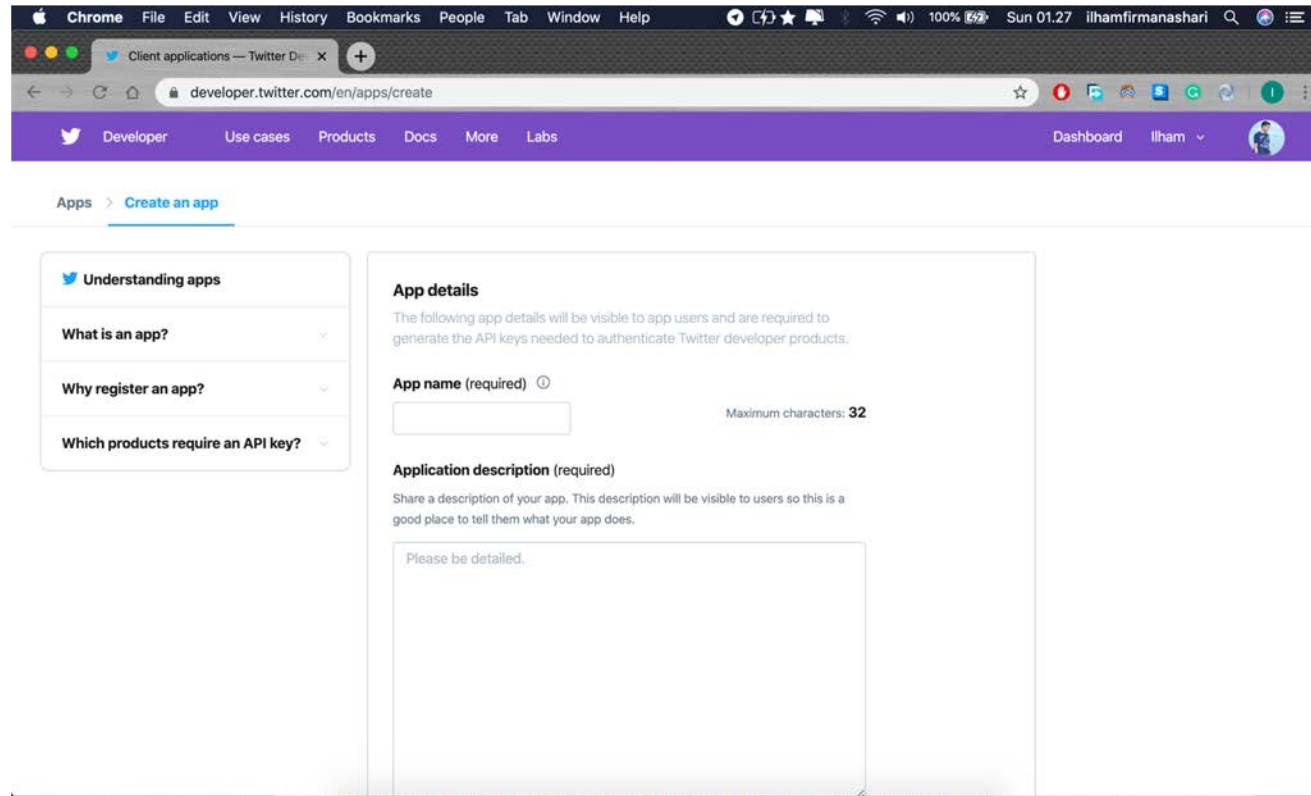
Get started

✓ [Create an app](#)

To use an API, we require you create an app as part of our OAuth authorization scheme. Visit the [Apps](#) page of this developer portal to create one. Then, return to this page to complete the next step.

Python – Twitter Step-by-Step (7)

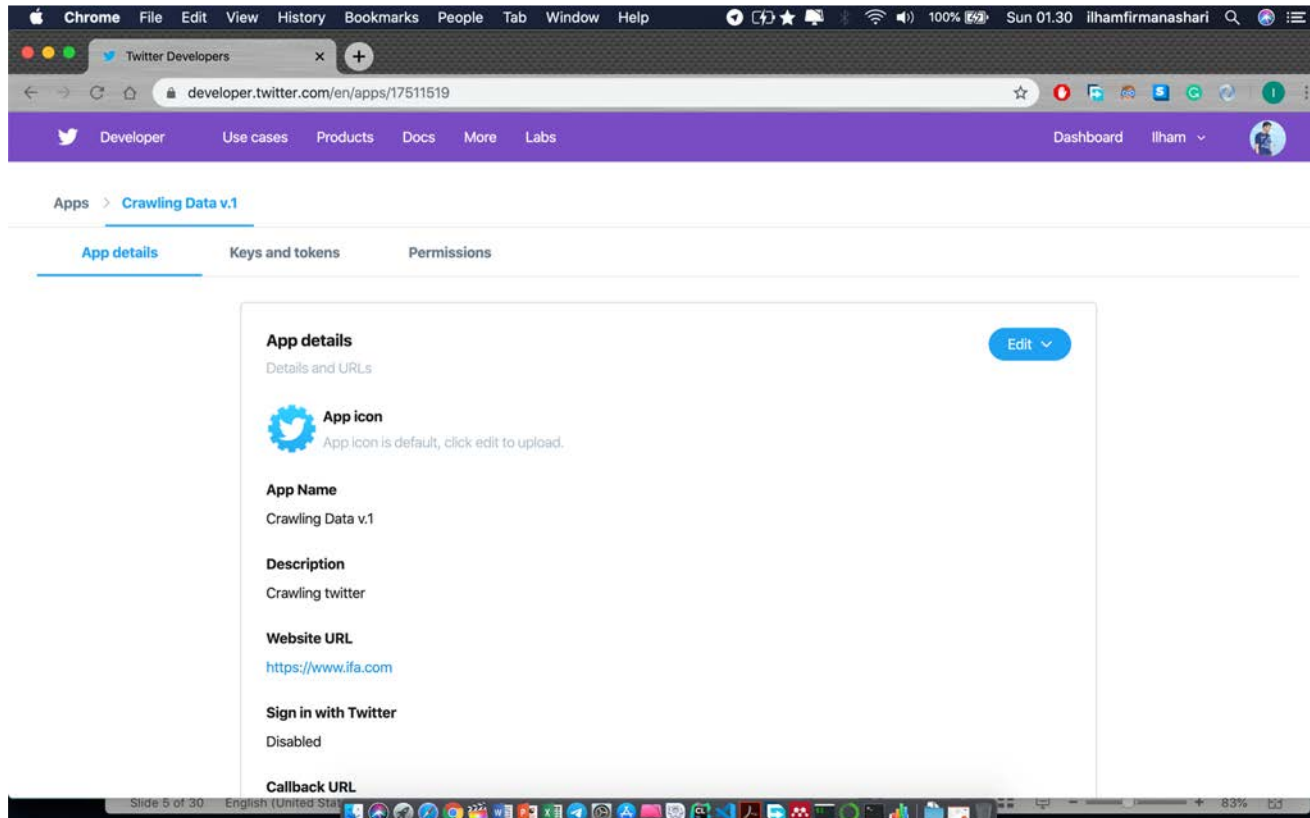
- Klik link *Create an app* untuk memulai pembuatan aplikasi di twitter.
- Isi form untuk menyelesaikan pembuatan aplikasi di twitter.



The screenshot shows the 'Create an app' page on the Twitter Developer Portal. The browser is Chrome, and the URL is `developer.twitter.com/en/apps/create`. The page has a purple header with navigation links: Developer, Use cases, Products, Docs, More, Labs, Dashboard, and a user profile for 'Ilham'. Below the header, the breadcrumb 'Apps > Create an app' is visible. On the left, there is a sidebar with the title 'Understanding apps' and four expandable sections: 'What is an app?', 'Why register an app?', and 'Which products require an API key?'. The main content area is titled 'App details' and contains a note: 'The following app details will be visible to app users and are required to generate the API keys needed to authenticate Twitter developer products.' Below this note, there are two required fields: 'App name (required)' with a text input and a character limit of 32, and 'Application description (required)' with a larger text area. The description field has a placeholder text 'Please be detailed.'

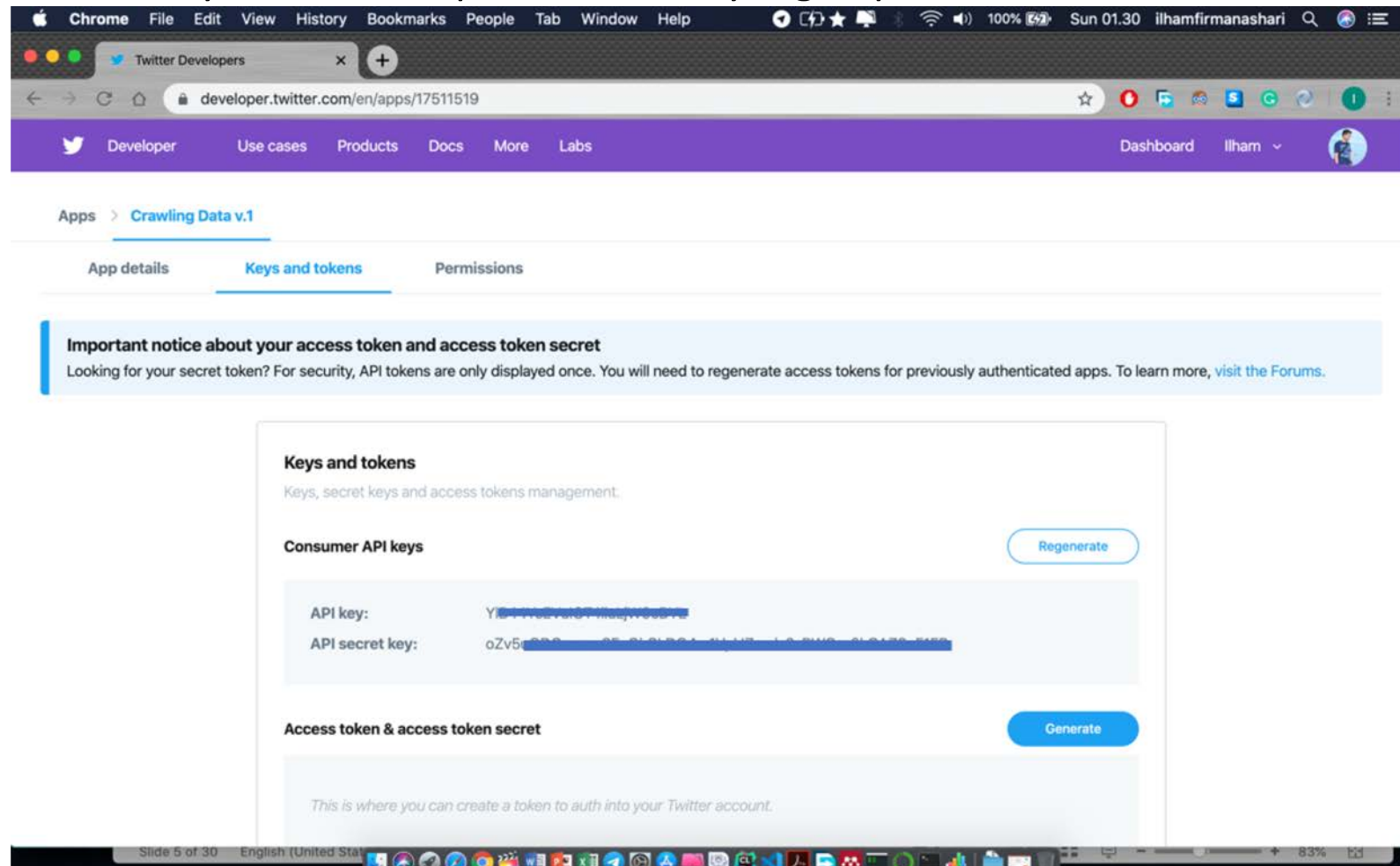
Python – Twitter Step-by-Step (8)

- Jika berhasil mengisi form pembuatan aplikasi di twitter, akan muncul halaman berikut:

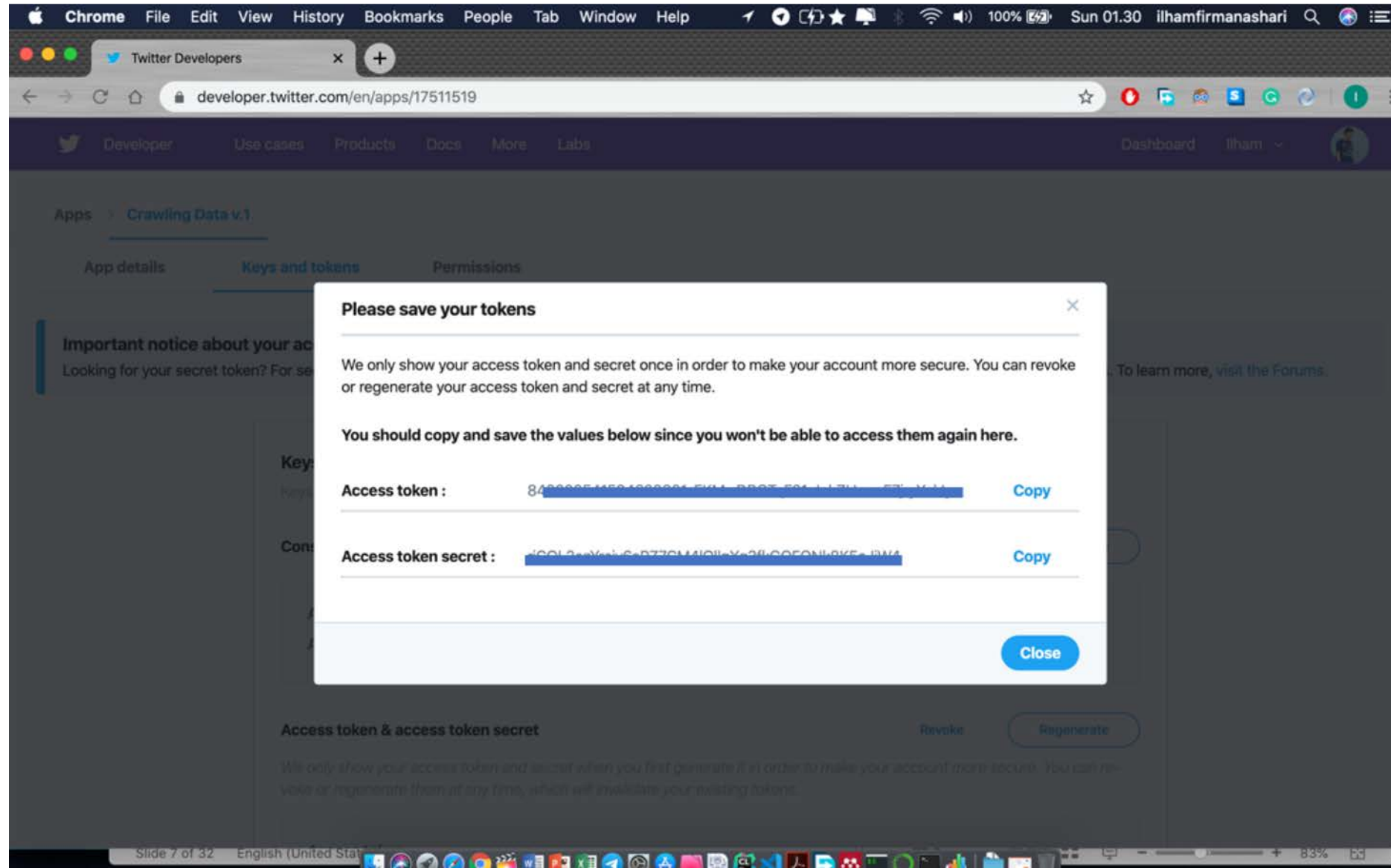


Python – Twitter Step-by-Step (8)

- Klik link Keys and Tokens untuk melihat key dan token aplikasi twitter yang dapat dimanfaatkan pada pemrograman python.
- Tampilannya seperti berikut:



Python – Twitter Step-by-Step (9)



Get Started

Python – Preparation (1)

- Setelah mendapatkan key dan token dari twitter, langkah selanjutnya adalah menginstall library tweepy untuk python.
- Instruksinya:

Install tweepy

```
pip install tweepy
```

Collecting tweepy

Downloading <https://files.pythonhosted.org/packages/36/1b/2bd38043d22ade352fc3d3902cf30ce0e2f4bf285be3b304a2782a767aec/tweepy-3.8.0-py2.py3-none-any.whl>

Collecting requests-oauthlib>=0.7.0 (from tweepy)

Downloading https://files.pythonhosted.org/packages/a3/12/b92740d845ab62ea4edf04d2f4164d82532b5a0b03836d4d4e71c6f3d379/requests_oauthlib-1.3.0-py2.py3-none-any.whl

Requirement already satisfied: PySocks>=1.5.7 in ./opt/anaconda3/lib/python3.7/site-packages (from tweepy) (1.7.1)

Requirement already satisfied: six>=1.10.0 in ./opt/anaconda3/lib/python3.7/site-packages (from tweepy) (1.12.0)

Requirement already satisfied: requests>=2.11.1 in ./opt/anaconda3/lib/python3.7/site-packages (from tweepy) (2.22.0)

Collecting oauthlib>=3.0.0 (from requests-oauthlib>=0.7.0->tweepy)

Downloading <https://files.pythonhosted.org/packages/05/57/ce2e7a8fa7c0afb54a0581b14a65b56e62b5759dbc98e80627142b8a3704/oauthlib-3.1.0-py2.py3-none-any.whl> (147kB)

153kB 141kB/s eta 0:00:01

Python – Connecting (2)

- Setelah berhasil meng-*install* library tweepy, gunakan kode berikut untuk mengatur koneksi program python ke aplikasi twitter.

Credentials

```
# Credentials  
  
consumer_key = "  
consumer_secret = "P  
access_token = "040000541504000001-  
access_token_secret = "0-
```

Python – Connecting (3)

Authorization

```
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)
```

Python – Send Data Tweets

- You can send tweets using your API access. Note that your tweet needs to be 280 characters or less.

```
# Post a tweet from Python
api.update_status("Look, I'm tweeting from #Python in my #earthanalytics class! @EarthLabCU")
# Your tweet has been posted!
```

Python – Get Data Tweets (1)

- Gunakan kode berikut untuk mendapatkan *tweets* dari aplikasi yang sudah dibuat.

```
1  # tweets from my stream
2  public_tweets = api.home_timeline()
3  for tweet in public_tweets:
4      print(tweet.text)
```

- Bila berhasil, output akan menampilkan semua *tweets* yang bisa dibaca oleh akun twitter milik kita sendiri.

Python – Get Data Tweets (2)

- Keluarannya seperti berikut:

```
#SobatBKN pagi ini demi keamanan nyai dan akang Peserta SKD [ASLI] mimin lagi periksa statusm  
u loh 🙏 semoga kamyu s... https://t.co/uiq6o6Lnir  
#SobatBKN mimin dikunjungi juga loh sama Wamen @Kemenag_RI dan Dir. PSIK BKN Pusat memberikan  
motivasi di siang har... https://t.co/fGR4cAlAKk  
Peserta terdaftar ujian SKD (total) 3.361.802
```

```
Data Per 29 Feb 2020, pukul 17.21 wib  
Peserta login SKD 2.944.279
```

```
K... https://t.co/dluHBueuwa  
#SobatBKN mimin rela menunda, demi kesuksesan kamu 😊 dan kita tetap bersamamu dilain waktu c  
ihuy 🥳🥳 Kemenag tilok g... https://t.co/hA5feKzzlJ  
Hari yang cerah ini Kepala BKN beserta Ka.Kanreg Bandung mengunjungi tilok SKD Kemenag Graha  
Batununggal Indah Band... https://t.co/rcvqelgfZz  
#SobatBKN, Sukses yang kita raih dalam setiap hal tak lepas dr peran serta dan dukungan kelua  
rga.
```

Python – Get Data Tweets (3)

- Keluaran dari kode sebelumnya, juga dapat dimodifikasi sehingga dapat tampil dalam format json.

- Kode pythonnya:

```
import json
for status in tweepy.Cursor(api.home_timeline).items(10):
    # Process a single status
    print(json.dumps(status._json))
```

- Keluarannya adalah semua informasi *tweets* status yang tersusun dalam format json.

```
{
  "created_at": "Sun Jun 30 05:40:49 +0000 2019",
  "id": 1145205527300857856,
  "id_str": "1145205527300857856",
  "text": "Percobaan tweets pertama",
  "truncated": false,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [],
    "urls": []
  },
  "source": "<a href='\"https://mobile.twitter.com/\"' rel='\"nofollow\"'>Twitter Web App</a>",
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 1145151860644311041,
    "id_str": "1145151860644311041",
    "name": "digitest",
    "screen_name": "digitest13",
    "location": "",
    "description": "",
    "url": null,
    "entities": {
      "description": {
        "urls": []
      }
    },
    "protected": false,
    "followers_count": 0,
    "friends_count": 1,
    "listed_count": 0,
    "created_at": "Sun Jun 30 02:07:34 +0000 2019",
    "favourites_count": 0,
    "utc_offset": null,
    "time_zone": null,
    "geo_enabled": false,
    "verified": false,
    "statuses_count": 1,
    "lang": null,
    "contributors_enabled": false,
    "is_translator": false,
    "is_translation_enabled": false,
    "profile_background_color": "F5F8FA",
    "profile_background_image_url": null,
    "profile_background_image_url_https": null,
    "profile_background_tile": false,
    "profile_image_url": "http://abs.twimg.com/sticky/default_profile_images/default_profile_normal.png",
    "profile_image_url_https": "https://abs.twimg.com/sticky/default_profile_images/default_profile_normal.png",
    "profile_link_color": "1DA1F2",
    "profile_sidebar_border_color": "C0DEED",
    "profile_sidebar_fill_color": "DDEEFF",
    "profile_text_color": "333333",
    "profile_use_background_image": true,
    "has_extended_profile": false,
    "default_profile": true,
    "default_profile_image": true,
    "following": false,
    "follow_request_sent": false,
    "notifications": false,
    "translator_type": "none"
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "is_quote_status": false,
  "retweet_count": 0,
  "favorite_count": 0,
  "favorited": false,
  "retweeted": false,
  "lang": "in"
},
{
  "created_at": "Sat Jun 29 11:04:04 +0000 2019",
  "id": 1144924489370628097,
  "id_str": "1144924489370628097",
  "text": "Sekjen Kominfo Paparkan Pelindungan Data Pribadi dan Beasiswa Talenta Digital https://t.co/raGLLS9NlD",
  "truncated": false,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [],
    "urls": [
      {
        "url": "https://t.co/raGLLS9NlD",
        "expanded_url": "http://dlvr.it/B7VQ35",
        "display_url": "dlvr.it/B7VQ35",
        "indices": [
          78,
          101
        ]
      }
    ]
  },
  "source": "<a href='\"https://dlvr.it/B7VQ35\"' rel='\"nofollow\"'>dlvr.it/B7VQ35</a>"
}
```

Python – Get Data Tweets (4)

- *Tweets* dari sebuah *taggar* bersifat real-time yang disebut *streaming*.
- Data *streaming* dari twitter akan menampilkan semua *tweets* baru dari sebuah taggar.
- Pemrograman python juga dapat dimodifikasi untuk menangkap *tweets* baru dari sebuah taggar.

Python – Get Data Tweets (5)

- Kode python untuk mengambil data *streaming* twitter

```
# Streaming Tweets
#override tweepy.StreamListener to add logic to on_status
class MyStreamListener(tweepy.StreamListener):
    def on_status(self, status):
        print(status.text)

# Create you Stream object with authentication
stream = tweepy.Stream(auth, MyStreamListener())

# Filter Twitter Streams to capture data by the keywords
stream.filter(track = ['corona'])
```

Wait until the end for a special surprise

Narration: @akarM13_

Production: @bbbakooo , @B4r3z and...

RT @briasnewaccount: Y'all want doctors to come up with a cure for the corona virus but think routine vaccines cause autism and the flu sho...

Kübra Par'ı çok da bilmem. Fakat buradaki niyetinin kötü olmadığından eminim. Bilimsel, veri bağlamında bir oranlam... <https://t.co/zBLRrXJAcY>

Not your third eye open

RT @AdrianNormanDC: So far in 2020:

```
description = status.user.description
loc = status.user.location
text = status.text
coords = status.coordinates
name = status.user.screen_name
user_created = status.user.created_at
followers = status.user.followers_count
id_str = status.id_str
created = status.created_at
retweets = status.retweet_count
bg_color = status.user.profile_background_color
```


Python – Get Data Tweets (6)

- Mengambil data berdasarkan ID dan mengubah ke dalam bentuk dataframe

```
tweets = []  
for tweet in api.user_timeline(id='jokowi', count=10):  
    tweets.append((tweet.created_at, tweet.id, tweet.text))  
  
tweetsdf = pd.DataFrame(tweets)
```

Python – Get Data Tweets (7)

```
tweets = tweepy.Cursor(api.search, q="#climate+change -filter:retweets", lang="en", since="2018-11-01").items(5)
all_tweets = [tweet.text for tweet in tweets]
all_tweets[:5]
```

Python – Cleaning Data Tweets (8)

```
def remove_url(txt):  
    return " ".join(re.sub("([0-9A-Za-z \t])|(\w+:\/\/\S+)", "", txt).split())
```

```
def remove_url(txt):  
    """Replace URLs found in a text string with nothing  
    (i.e. it will remove the URL from the string).  
  
    Parameters  
    -----  
    txt : string  
        A text string that you want to parse and remove urls.  
  
    Returns  
    -----  
    The same txt string with url's removed.  
    """  
    |  
    return " ".join(re.sub("([0-9A-Za-z \t])|(\w+:\/\/\S+)", "", txt).split())
```

```
all_tweets_no_urls = [remove_url(tweet) for tweet in all_tweets]  
all_tweets_no_urls[:5]
```

Python – The result (9)

```
words_in_tweet = [tweet.lower().split() for tweet in all_tweets_no_urls]
words_in_tweet[:1]
```

Split text based
on word

```
[['hope',  
  'is',  
  'found',  
  'in',  
  'the',  
  'streets',  
  'its',  
  'walking',  
  'with',  
  'millions',  
  'of',  
  'young',  
  'people',  
  'all',  
  'over',  
  'the',  
  'world',  
  'join',  
  'them',  
  'and',  
  'lets',  
  'ch']]
```

Python – Word Frequencies Analysis (10)

```
import itertools
import collections
# List of all words across tweets
all_words_no_urls = list(itertools.chain(*words_in_tweet))

# Create counter
counts_no_urls = collections.Counter(all_words_no_urls)

counts_no_urls.most_common(15)
```

```
[('climate', 4),
 ('change', 4),
 ('the', 3),
 ('with', 3),
 ('young', 2),
 ('and', 2),
 ('when', 2),
 ('to', 2),
 ('science', 2),
 ('hope', 1),
 ('is', 1),
 ('found', 1),
 ('in', 1),
 ('streets', 1),
 ('its', 1)]
```

Python – Convert to DataFrame(11)

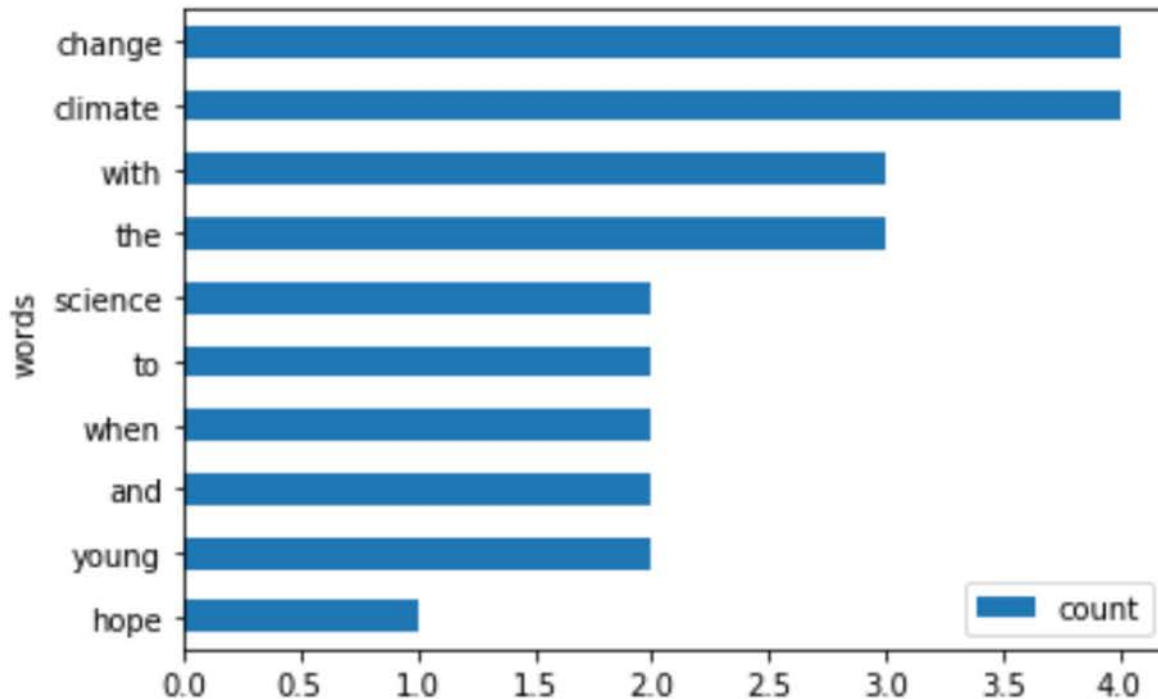
```
clean_tweets_no_urls = pd.DataFrame(counts_no_urls.most_common(10),  
                                     columns=['words', 'count'])  
  
clean_tweets_no_urls.head()
```

	words	count
0	climate	4
1	change	4
2	the	3
3	with	3
4	young	2

Python – Visualization (12)

```
import matplotlib.pyplot as plt
clean_tweets_no_urls.sort_values(by='count').plot.barh(x='words',
                                                         y='count')

plt.show()
```



THANK YOU