Documentation about using Python to crawl twitter data based on twitter API

Section 1 Python and related packages installation

1. Python3.x installation

Please refer this website <https://docs.python.org/3/using/index.html> for install your Python based on your desktop environment.

2. IDE for Python

(1) There are many IDEs for Python, such as PyCharm, Wing, Spyder.

(2) You could choose one for your environment:

1> Pycharm: <https://www.jetbrains.com/pycharm/download/#section=windows>

2> Wing: <https://wingware.com/doc/install/installing>

3> Spyder: <https://pythonhosted.org/spyder/installation.html>

4> Othes

(3) Related packages installation

1> This experiment utilizes tweepy, datetime, csv packages in Python

2> Packages installation could refer this website: <https://packaging.python.org/tutorials/installing-packages/>

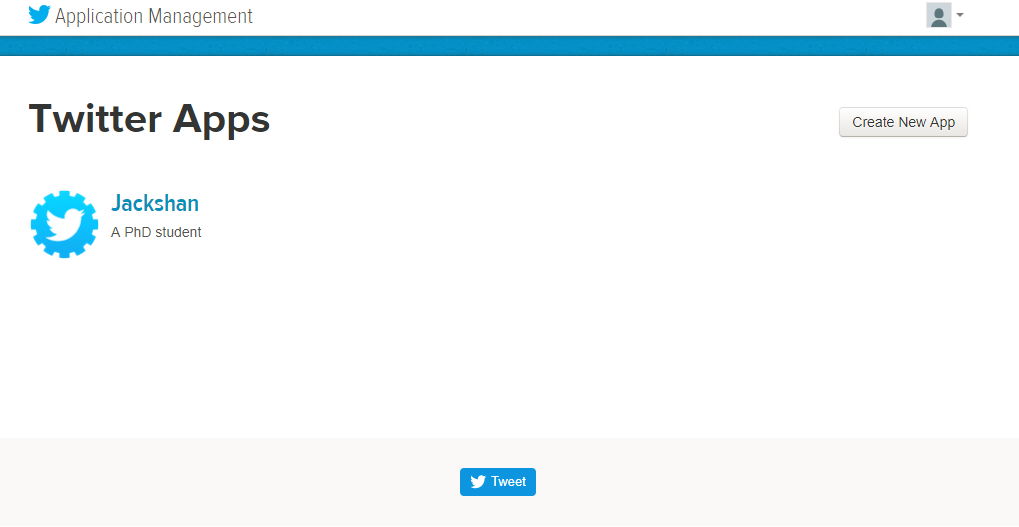
I recommendation to use pip for installing packages

Section 2 Instructions for constructing Python codes to access Twitter API

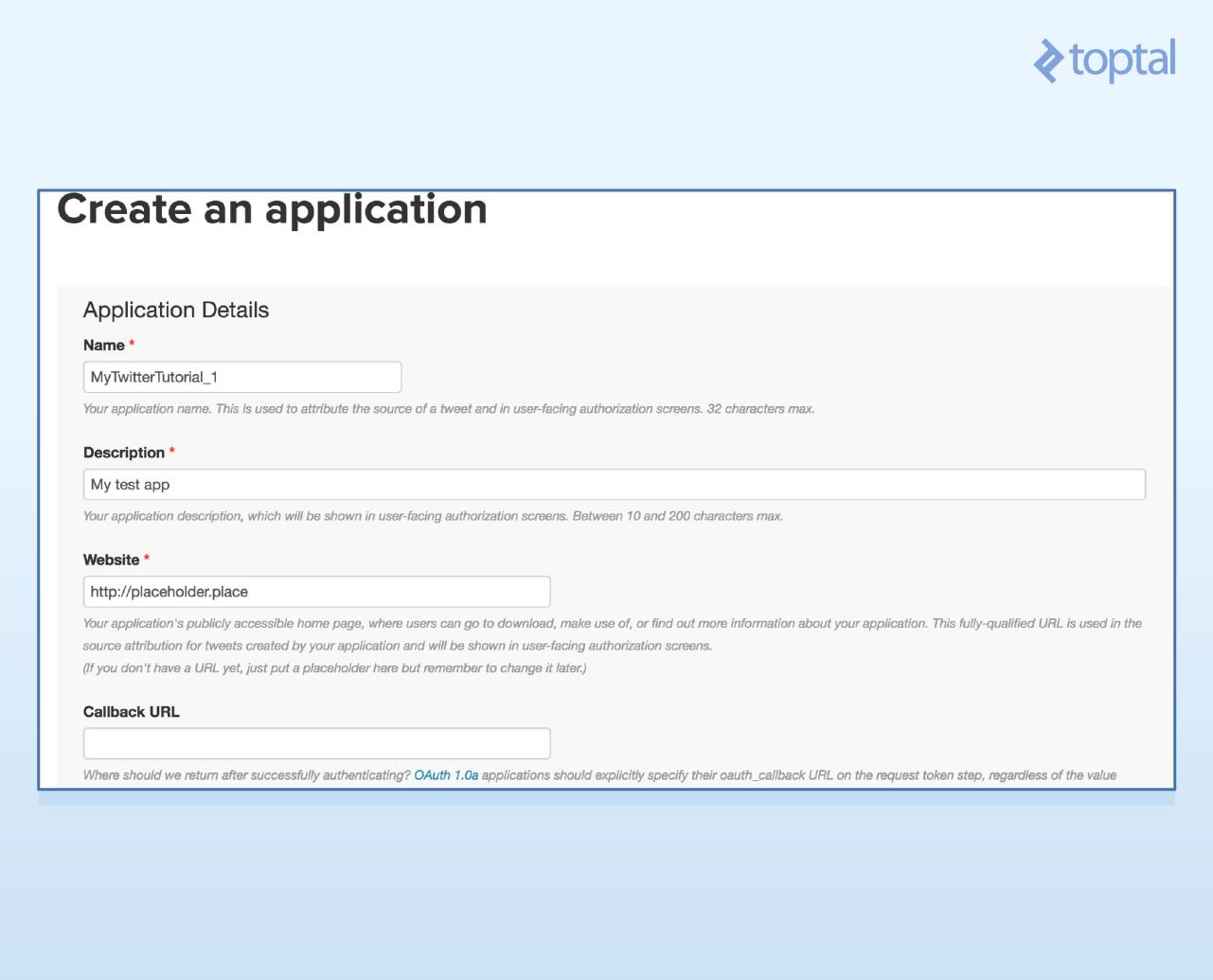
1. Create developer account in Twitter to get the Auth towards Twitter API

1> Open up <https://apps.twitter.com/>

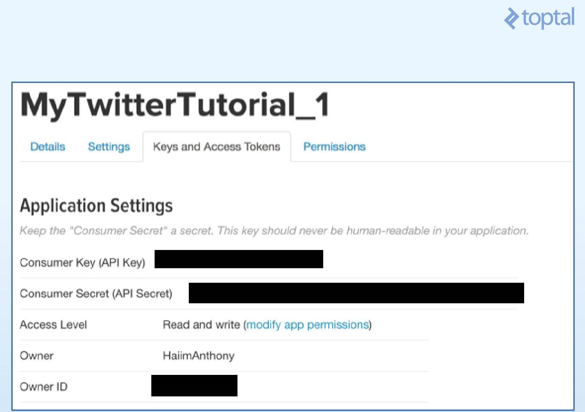
2> Click ‘Create New App’ in the top right of the page.



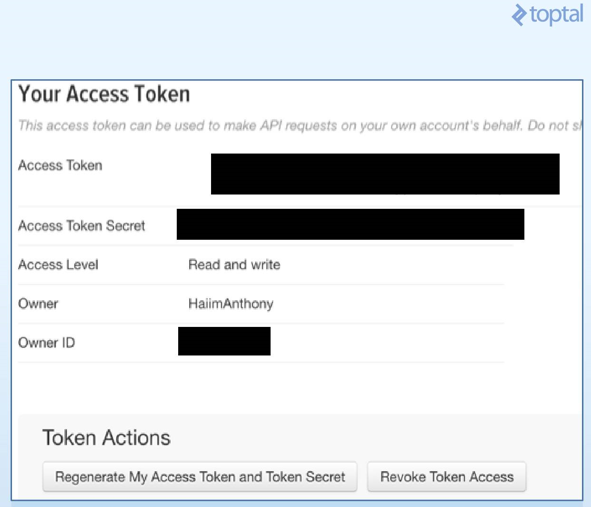
3> Fill out the required blanks. Name is the unique identity for your app. Website can be filled by the default setting.



(4) Click Keys and Access Tokens to get your API key and API secret.



(5) Get your access token and access token secret shown in the following page.



2. Create your Twitter API object

1> Create auth object

auth = tweepy.OAuthHandler(consumer\_key, consumer\_secret)

2> Set your access toke and access secret

auth.set\_access\_token(access\_token, access\_token\_secret)

3> Get the API object

api = tweepy.API(auth)

4> Then you could utilize your api object to access Twitter data.

Section 3 Examples

1. Data crawled:

- Text - the content of tweets.

- created- The time when the tweet was created.

- ScreenName- Screen name of the Tweeter user who posted the tweet.

- replyToUID- Nullable If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet’s author ID. This will not necessarily always be the user directly mentioned in the Tweet.

- replyToSID- Nullable If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet’s ID.

- id-The integer representation of the unique identifier for this Tweet.

- truncated -Indicates whether the value of the text parameter was truncated. For example, as a result of a retweet exceeding the 140 character Tweet length. Truncated text will end in ellipsis, like this ...

- retweeted- whether the tweet has retweeted or not.

- Longitude-The longitude of the Tweet’s location

- Latitude-The latitude of the Tweet’s location

- Favorited-Nullable Perspectival Indicates whether this Tweet has been liked by the authenticating user.

- replytoSN -Screen name of the user whom the tweet was replied to.

- FavoriteCount - the number of Tweet users who like the tweet.

- statusSource- The device from which the tweet was sent.

- retweetCount- the number of retweet of the tweet.

- isRetweet - whether the current tweet is a retweet or a new tweet.

2. Search query is ‘happy’ and the result is stored in CSV files

1> Related code is in the Github: <https://github.com/Jackustc/Crawl_Twitter_data/blob/master/code/twitter_crawler_csv>

2> You could change the required authentication and search query based on the code.

Good luck and best wishes.