

## Data Scraping mit Python – Vorbereitungen

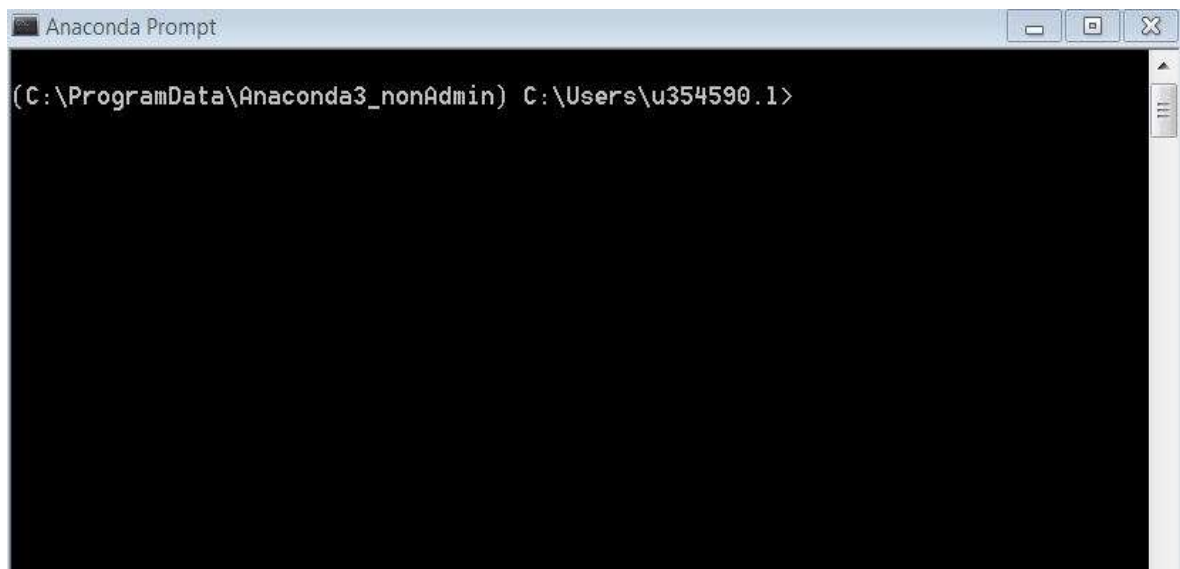
Die nachfolgende Anleitung beschreibt die Installation von Software sowie die Beschaffung von Zugangsberechtigungen für Programmierschnittstellen.

### Continuum Anaconda für Python 3 (64-Bit version)

Wir verwenden Anaconda sowie mehrere Zusatzpakete für die Programmiersprache Python. Anaconda kann unter folgendem Link heruntergeladen werden:

<https://www.anaconda.com/download/>

Nach der Installation (empfohlen sind Standardeinstellungen), kann ein Terminal (Mac / Linux) - oder für Windows Nutzer das Programm "[Anaconda Prompt](#)" - gestartet werden.



Im Anschluss sollten die folgenden Befehle einzeln ausgeführt werden:

```
conda install jupyter notebook
conda install seaborn
conda install html5lib
conda install selenium
pip install pdfminer.six
```

Warnungen können bei der Installation zunächst ignoriert und alle Fragen mit "yes" beantwortet werden. Um zu überprüfen, ob die Installation funktioniert hat, sollte folgender Befehl im Terminal / Anaconda Prompt ausgeführt werden:

```
jupyter notebook
```

Der Befehl sollte eine Jupyter Programmierungsumgebung im Standardbrowser öffnen.



## Chrome Browser (aktuelle Version)

Wir werden einige Zeit im Browser verbringen. Grundsätzlich sind die meisten aktuellen Browser geeignet, für den Kurs ist aber die Verwendung eines Chrome Browsers dringend empfohlen, da wir einige spezielle Features von Chrome verwenden. Eine aktuelle Version kann hier installiert werden:

<https://www.google.com/chrome/>

Zusätzlich wird noch eine aktuelle Version des Chrome Drivers benötigt, um den Browser fernsteuern zu können:

<https://sites.google.com/a/chromium.org/chromedriver/downloads>

Hierbei ist zu beachten, dass die Versionsnummer des Chrome Drivers zu der Versionsnummer des installierten Chrome Browser passen sollte. Nach herunterladen der korrekten Version für das eigene Betriebssystem sollte der Chrome Driver in einem Ordner abgelegt werden.

## API's: Guardian & Twitter

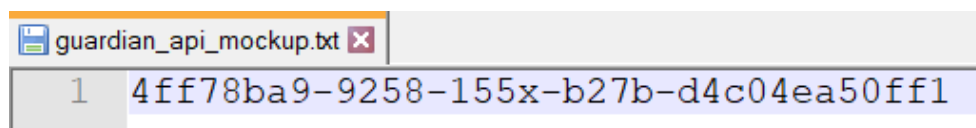
Wir werden im Verlauf des Kurses Daten aus Programmierschnittstellen (API) abgreifen, für die wir uns zunächst Zugangsrechte beschaffen müssen.

### Guardian

Für die Guardian API wird ein „developer key“ benötigt, der über folgenden Link registriert werden kann:

<https://open-platform.theguardian.com/access/>

Als Begründung („reason for key“) kann hier z.B. „Education“ angegeben werden. Der per E-Mail versendete Key kann anschließend in einer Textdatei abgelegt werden.



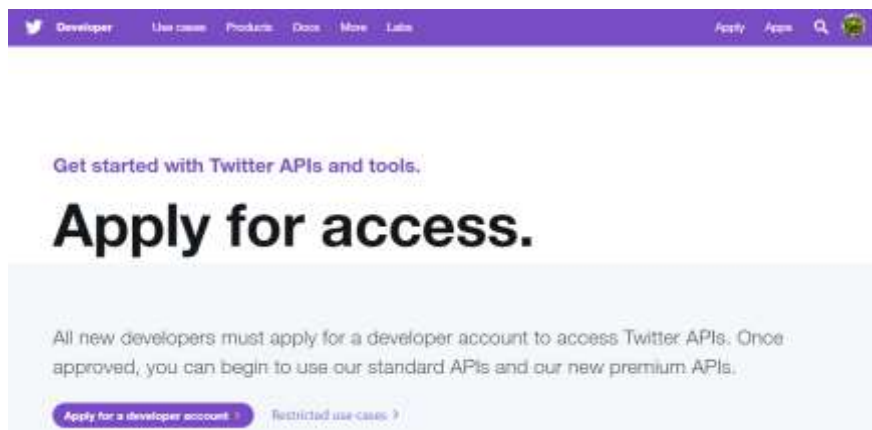
## Twitter

Für Twitter ist der Vorgang leider etwas aufwändiger. Zunächst muss (falls nicht bereits vorhanden) ein Twitter Account erstellt werden:

<https://twitter.com/i/flow/signup>

Es empfiehlt sich hier dringend, bei der Anmeldung eine gültige Mobiltelefonnummer anzugeben. Ansonsten könnten spätere Schritte für die Beschaffung der Zugangsrechte möglicherweise nicht ausgeführt werden.

Im nächsten Schritt muss auf <https://apps.twitter.com> ein „Developer“ Zugang freigeschalten werden.



Dabei sind einige Angaben zu machen, z.B. zum geplanten Verwendungszweck des Zugangs.

What is your primary reason for using Twitter developer tools?  
We'll help you on your path to getting the most out of Twitter APIs and data.

Professional <small>for commercial uses</small>	Hobbyist <small>for a personal project</small>	Academic <small>for education or research</small>	Other <small>if it fits any of these</small>
<input type="radio"/> Building iOS products	<input type="radio"/> Making a tool	<input checked="" type="radio"/> Doing academic research	<input type="radio"/> Unfolding Tweets on a website
<input type="radio"/> Building consumer products	<input type="radio"/> Building tools for Twitter users	<input type="radio"/> Teaching	<input type="radio"/> Doing something else

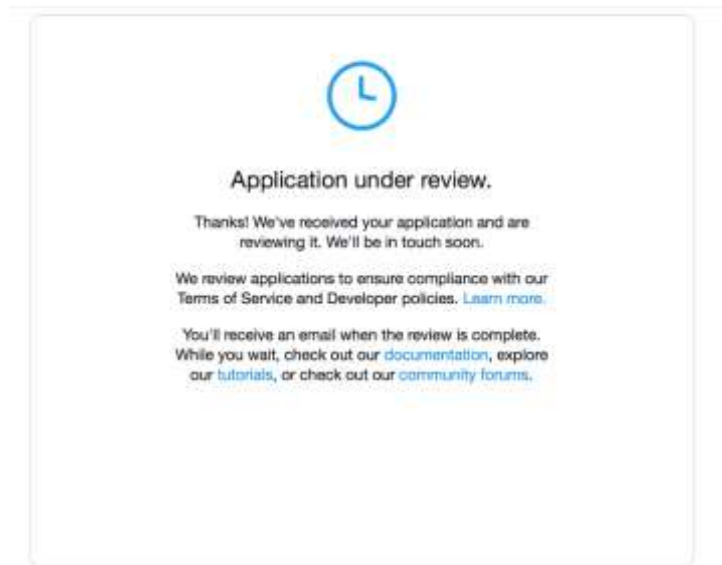
Are you planning to analyze Twitter data? ☒ Yes

Please describe how you will analyze Twitter data including any analysis of Tweets or Twitter users.

I would like to analyze likes and retweets of twitter users. I also would like to analyze texts of tweets.

Response must be at least 100 characters ☒

Nach absenden der Informationen kann es etwas dauern, bis Twitter die Zugangsbestätigung versendet. In Einzelfällen kann es vorkommen, dass Twitter per E-Mail weitere Informationen vor der Freischaltung verlangt.



Nachdem der Developer Zugang freigeschaltet ist, kann schließlich über <https://apps.twitter.com> eine Twitter App erstellt werden. Hierbei genügt es nur die Pflichtfelder auszufüllen. Es kann eine beliebige Webseite (z.B. <https://google.com>) eingetragen werden.

Terms of Service URL ⓘ

Privacy policy URL ⓘ

Organization name ⓘ

Organization website URL

**Tell us how this app will be used (required)**

This field is only visible to Twitter employees. Help us understand how your app will be used. What will it enable you and your customers to do?

This app is for collecting Twitter data to conduct social science research. It will not be used for commercial purposes.

Cancel Create

Nach Erstellung kann die App über <https://apps.twitter.com> aufgerufen werden. Über den Reiter „Keys and Access Tokens“ werden in einem letzten Schritt die benötigten Zugangsdaten (*Consumer Key, Consumer Secret, Access Token, Access Token Secret*) angelegt.

Details Settings **Keys and Access Tokens** Permissions

### Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	YOUR CODE WILL APPEAR HERE
Consumer Secret (API Secret)	YOUR CODE WILL APPEAR HERE
Access Level	Read-only (modify app permissions)
Owner	chris_bail
Owner ID	

**Application Actions**

Regenerate Consumer Key and Secret Change App Permissions

### Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	YOUR CODE WILL APPEAR HERE
Access Token Secret	YOUR CODE WILL APPEAR HERE
Access Level	Read-only
Owner	chris_bail
Owner ID	

Es empfiehlt sich, die Zugangsdaten in einer Textdatei abzuspeichern.

```
1 lhW9T481fgzQv9CJhHQjyjS75
2 oDmgDMzFzyZKKKMJ3FtD2LH0Pfd1eS7RPZFjdn487Fh0gfNDI7
3 416910461-jZNz0hSoEZ3o45712SxkjNtsAxZLrUwxAmTv4EHw
4 GLNAzfpFI61UbFgjH71ExHHZOIFBX2KB8KGvUbg2xhj9
```

## Zusammenfassung

1. Anaconda (Python 3, 64Bit) installieren
2. Python Pakete über Kommandozeile installieren
3. Chrome Browser & Chrome Driver installieren
4. Zugangsberechtigung für Guardian API beschaffen
5. Zugangsberechtigung für Twitter API beschaffen