Data Scraping mit Python – Vorbereitungen

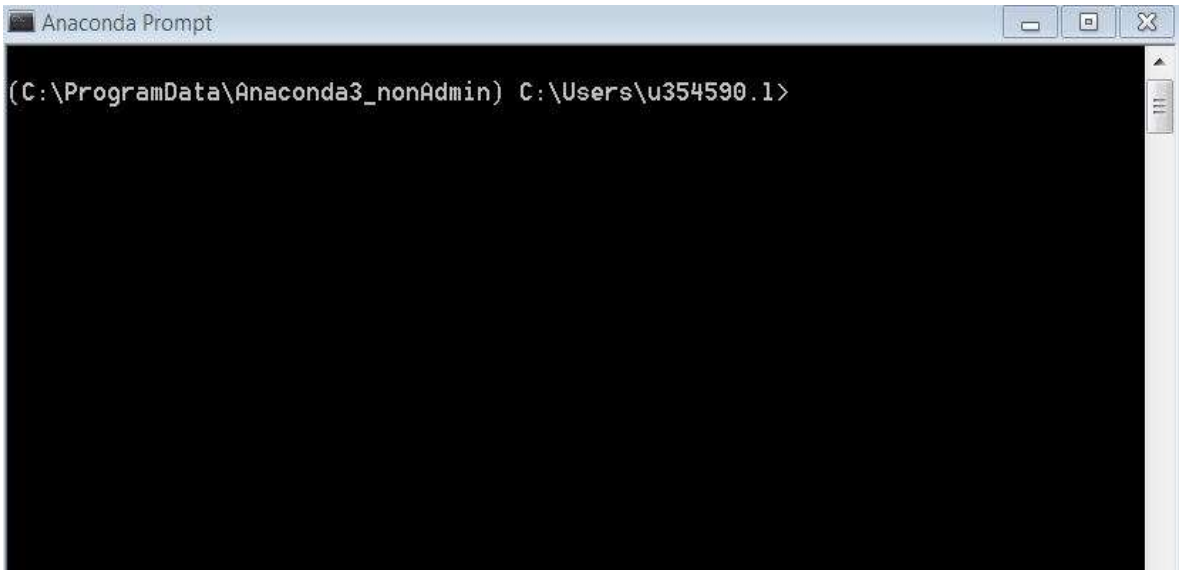
Die nachfolgende Anleitung beschreibt die Installation von Software sowie die Beschaffung von Zugangsberechtigungen für Programmierschnittstellen.

Continuum Anaconda für Python 3 (64-Bit version)

Wir verwenden Anconda 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 Programmierungumgebung 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 Browers 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.

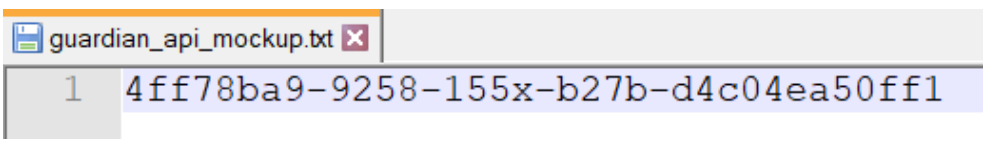
API's: Guardian & Twitter

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.

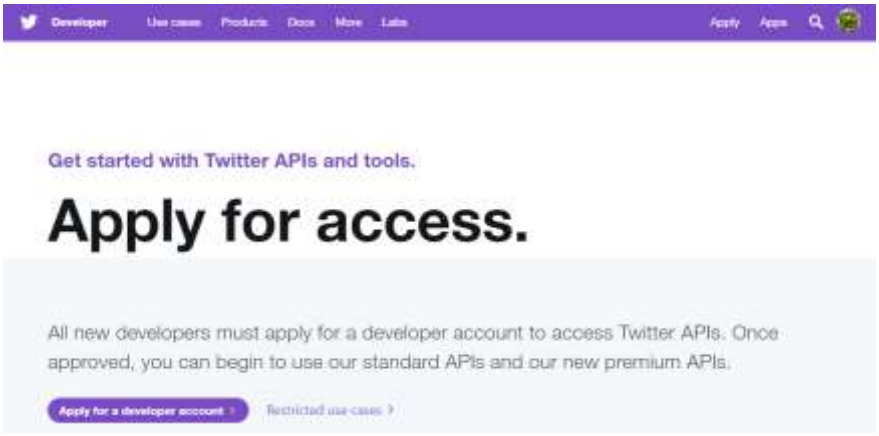


Twitter

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

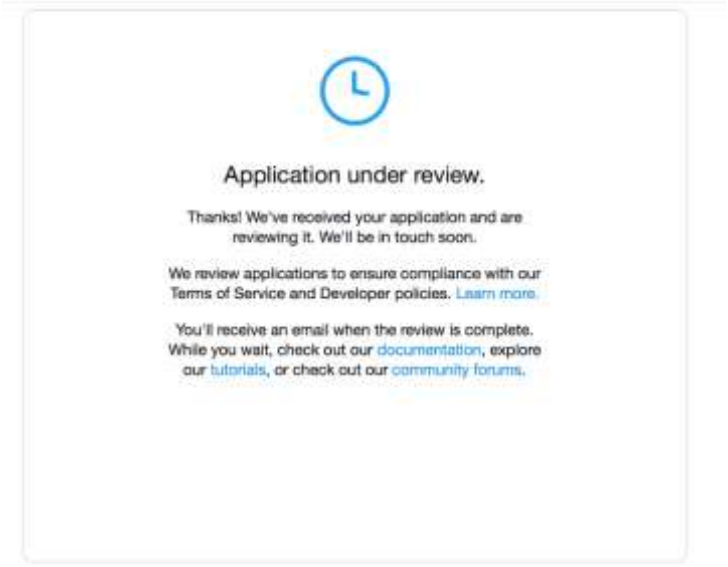
Es empfiehlt sich hier dringend, bei der Anmeldung eine gültige Mobiltelefonnummer anzugeben. Ansonsten könnten spätere Schritte für die Beschaffung der Zugangsrechte möglicherweise nicht ausgeführt werden.

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



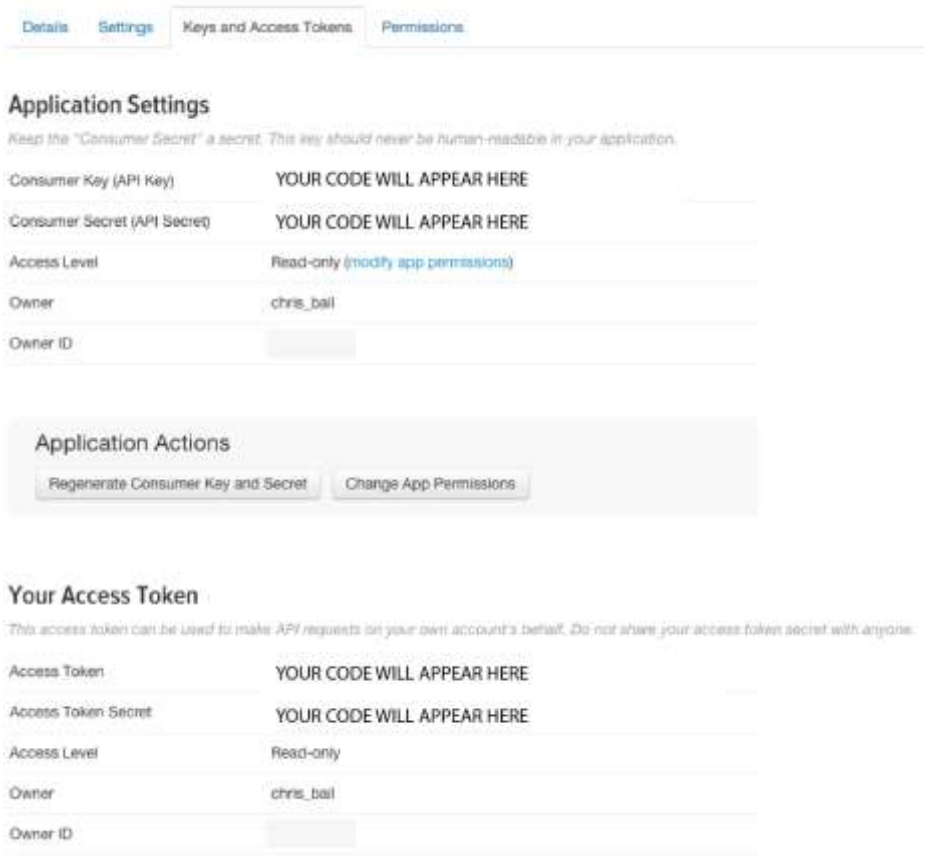
Dabei sind einige Angaben zu machen, z.B. zum geplanten Verwendungszweck des Zugangs.

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 freigeschalten 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.

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.



Es empfiehlt sich, die Zugangsdaten in einer Textdatei abzuspeichern.

twitter_api_mockup.txt

```
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