1.4 Web Scraping

1.4.1 requests

Den HTML Code einer Webseite herunterladen.

Modul einbinden

```
In [19]: import requests
```

Anwendung

```
In [20]: url = "http://python.beispiel.programmierenlernen.io"

         r = requests.get(url)

         # um nicht das Dokument zu sprengen, geben wir hier nur den HTML-Head aus
         print(r.text.split("<body>")[0])
```

```
<!DOCTYPE html>
<html lang="de">
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
    <meta name="description" content="">
    <meta name="author" content="">

    <title>Crawler-Auflistung</title>

    <!-- Bootstrap core CSS -->
    <link href="./lib/bootstrap/css/bootstrap.min.css" rel="stylesheet">
    <link href="./css/narrow-jumbotron.css" rel="stylesheet">
  </head>
```

Weitere Infos: http://docs.python-requests.org/en/master/user/quickstart/

1.4.2 beautifulsoup4

Kann HTML Code zerlegen und weiterverarbeiten.

Modul einbinden

```
In [21]: from bs4 import BeautifulSoup
```

Anwendung

```
In [22]: # BeautifulSoup kommt ins Spiel nachdem dem Webseite heruntergeladen wurde
         # (z.B. mit dem Requests-Modul)

         import requests
         url = "http://python.beispiel.programmierenlernen.io/index.php"
         r = requests.get(url)
```

```
In [23]: doc = BeautifulSoup(r.text, "html.parser")

         # mit bs4 können wir auf bestimmte Bereiche innerhalb der HTML zugreifen
         # z.B. auf die Inhalte der Tags mit der Klasse card-text
         content = doc.select_one(".card-text").text

         print(content.replace(". ", ". \n"))
```

```
Optio numquam ut accusantium laborum unde assumenda.
Ea et totam asperiores fugiat voluptatem vitae.
Et provident nam et mollitia.
```

Weitere Infos: https://www.crummy.com/software/BeautifulSoup/bs4/doc/#

1.4.3 urllib

Vereinfacht die Arbeit mit URLs

Modul einbinden

```
In [24]: import urllib
```

Anwendung

```
In [25]: from urllib.parse import urljoin

         url = "http://python.beispiel.programmierenlernen.io/index.php"
         # häufig sind Quellen als solche abgekürzten URLs angegeben
         src = "./img/1.jpg"

         image_url = urljoin(url, src)

         print(image_url)
```

```
http://python.beispiel.programmierenlernen.io/img/1.jpg
```

Mehr Details: https://docs.python.org/3/library/urllib.parse.html

---