Freie Universität Berlin

Prof. Dr. A. Voisard, N. Lehmann

Datenbanksysteme, SoSo 20

Übung 05

TutorIn: Gröling, Marc
Tutorium 04

David Ly & Thore Brehmer

30. Mai 2020

---

4 Aufgabe: Web-Scraper (20 Punkte)

1) *Schreiben Sie einen Web-Scraper in Python für das Scrapen der Webseite "heise.de". Verwenden Sie für Ihren "heise.de"-Scraper das Python Framework "Beautiful Soup" und das Python-Modul "Requests". Der "heise.de"-Scraper soll die Überschriften aller Artikel mit einem Bezug zum Thema "https" (https://www.heise.de/thema/https) zuerst in einer Datenstruktur speichern.* *(10 P.)*

```python
from bs4 import BeautifulSoup
import requests
from collections import Counter

url = 'https://www.heise.de/thema/https/seite-'
pages = 5

# returns the html code of a given url
def getPage(url):
    return BeautifulSoup(requests.get(url).text,'lxml')

# returns a list of all titles  (in "Datenstruktur", jeder Eintrag is prim_key)
def get_titles():
    all_titles = []
    # goes through each page
    for seite in range(1,pages+1):
        # getPage
        page = getPage(url+str(seite))
        # fing all titles
        titles = page.findAll('span',class_='a-article-teaser__title-text')
        #goes through each title and appends to a list, also strips unnecessary chars
        for title in titles:
            all_titles.append(title.text.strip())
    return all_titles
```

src/main.py

2) *Wie lauten die Top-3 Wörter in den Überschriften aller zum Thema "https" veröffentlichten Artikel auf "heise.de"?* *(10 P.)*
Hinweis: Schauen Sie sich als Beispiel einen Web-Scraper in Python für die Webseite "greyhound-data.com" auf GitHub (https://github.com/xconnect/fub.bsc.dbs.scraper.greyhound-data.com) und die gescrapte Webseite "greyhound-data.com" an. Versuchen Sie zuerst das Programm zu verstehen bevor Sie mit der Programmierung beginnen.

```
26 # returns top n used words of all titles
27 def top_words(n, all_titles):
28     aio_list = ""
29     # goes through all titles and combines them to one list
30     for title in all_titles:
31         aio_list +=title + ' '
32     # splits each word in a new list
33     words = aio_list.split()
34     # counts all wourds and returns top n
35     top_n_words = Counter(words).most_common(n)
36     return top_n_words

38 print(top_words(3,get_titles()))
```

src/main.py

```
thore@ubuntu:~/Desktop$ python3 main.py clear
[('HTTPS', 29), ('und', 18), ('für', 16)]
thore@ubuntu:~/Desktop$
```