

Introduction to Web Science

Assignment 3

Jun Sun

junsun@uni-koblenz.de

Iryna Dubrovskaya

idubrovskaya@uni-koblenz.de

Institute of Web Science and Technologies

Department of Computer Science

University of Koblenz-Landau

Submission until: November 23, 2021, 24:00 CET

Tutorial on: November 25, 2021, 16:00 CEST

This assignment focuses on the concepts of 1) World Wide Web, 2) HTTP and 3) Programming in Python. Some of the tasks may require you to do additional research extending the lecture. Please keep the citation rules in mind. For all the assignment questions that require you to write a code, make sure to include the code in the answer sheet, along with a separate python file. Where screen shots are required, please add them in the answers directly and not as separate files.

Date: 23/11/2021

Team Name: Boehm

Abhinav Ralhan (abhinavr8@uni-koblenz.de)

Fatima Akram (fatimaakram9396@uni-koblenz.de)

Hammad Ahmed (hammadahmed@uni-koblenz.de)

Vishal Vidhani (vvidhani@uni-koblenz.de)

1 Python Programming. GET and POST Requests (18 points)

```
#!/usr/bin/env python
# coding: utf-8

# In[31]:

import requests

URL = 'https://en.wikipedia.org/w/rest.php/v1/page/boehm'

getRequest = requests.get(url = URL)

jsonFormat = getRequest.json()

articleId = jsonFormat['id']
articleTitle = jsonFormat['title']
articleTimeStamp = jsonFormat['latest']['timestamp']

API_ENDPOINT = "https://pastebin.com/api/api_post.php"

API_KEY = "F5aT5BB8R7QV0cFg22_ZzIWBGnSBMRxi"

# your source code here
source_code = str(articleId) + '\n' + articleTitle + '\n'+
articleTimeStamp

# data to be sent to api
data = {'api_dev_key':API_KEY,
        'api_option':'paste',
        'api_paste_code':source_code,
        'api_paste_format':'python'}

# sending post request and saving response as response object
r = requests.post(url = API_ENDPOINT, data = data)

# extracting response text
pastebin_url = r.text
print("The pastebin URL is:%s"%pastebin_url)

httpHeader=requests.get(pastebin_url)
HeaderValue = httpHeader.headers

file = open("boehm.txt", "w")
file.write(str(HeaderValue))
file.close()
```

2 Command Line (10 points)

curl https://en.wikipedia.org/w/rest.php/v1/page/boehm

```
admin -- zsh -- 191x46
Last login: Sun Nov 21 12:39:26 on ttys000
(base) admin@users-MacBook-Air ~ % curl https://en.wikipedia.org/w/rest.php/v1/page/boehm
{"id":2574858,"key":"Boehm","title":"Boehm","latest":{"id":988396756,"timestamp":"2020-11-12T22:41:47Z"},"content_model":"wikitext","license":{"url":"https://creativecommons.org/licenses/by-sa/3.0/","title":"Creative Commons Attribution-Share Alike 3.0"},"source":{"other uses":"Boehm"}}
```

To get in json format

curl -s 'curl https://en.wikipedia.org/w/rest.php/v1/page/boehm ' | jq '

```
(base) admin@users-MacBook-Air ~ % curl -s https://en.wikipedia.org/w/rest.php/v1/page/boehm | jq '.
{
  "id": 2574858,
  "key": "Boehm",
  "title": "Boehm",
  "latest": {
    "id": 988396756,
    "timestamp": "2020-11-12T22:41:47Z"
  },
  "content_model": "wikitext",
  "license": {
    "url": "https://creativecommons.org/licenses/by-sa/3.0/",
    "title": "Creative Commons Attribution-Share Alike 3.0"
  },
  "source": "{other uses}{{IPAC-en|us|b|er|m}} is a {{Germany|German}} surname, transliterated from B hm (literally: Bohemian, from {{Bohemia}}) or reflective of a spelling adopted by a given family before the introduction of the {{Diaeresis (diacritic)}#umlaut|umlaut diacritic}}. It may refer to:\n* [[Aleksandra Zi kowska-Boehm]] (born 1949), American-Polish author\n* [[Barry Boehm]] (born 1935), American software engineer\n* [[Christopher Boehm]] (b. 1931) American Anthropologist, Primatologist\n* [[David Boehm]] (1893-1962), American screenwriter\n* [[Doug Boehm]] (born 1969), American record producer and sound engineer\n* [[Edward Marshall Boehm]] (1913-1969), American sculptor\n* [[Elisabet Boehm]] (1859-1943), German feminist and writer\n* [[Erhard F. Boehm]] (1911-1994), Australian farmer and amateur ornithologist\n* [[Gero von Boehm]] (born 1954), German journalist\n* [[Gottfried Boehm]] (born 1942), German art historian and philosopher\n* [[Hanns-Peter Boehm]] (born 1928), German chemist and professor emeritus\n* [[Henry Boehm]] (1775-1875), American clergyman and pastor\n* [[Joseph Boehm]] (Sir (Joseph) Edgar Boehm, 1834-1898), Austrian sculptor\n* [[Martin Boehm]] (1725-1812), American clergyman and pastor\n* [[Mary Louise Boehm]] (1928-2002), American pianist and painter\n* [[Paul Boehm]] (born 1974), Canadian skeleton racer\n* [[Peter Boehm]], Canadian diplomat\n* [[Peter M. Boehm]] (1845-1914), soldier in the American Civil War, Medal of Honor recipient\n* [[Robert Boehm]] (1914-2006), American political activist\n* [[Ron Boehm]] (born 1943), retired ice hockey winger\n* [[Roy Boehm]] (1924-2008), known as the "First SEAL", established the U.S. Navy's first [[Navy seals|SEAL Team]].\n* [[Sydney Boehm]] (1908-1998), American screenwriter and producer\n* [[Theobald Boehm]] (1794-1881), Bavarian inventor and musician\n* [[Boehm system]] of flute fingering\n* [[Boehm system (clarinet)]], a similar system for the clarinet\n* [[Theodore R. Boehm]] (born 1938), Justice of the Indiana Supreme Court\n* [[Traugott Wilhelm Boehm]] (1836-1917) founder of Hahndorf Academy in South Australia\n\nSee also ==\n* [[Boehm system]]\n* [[Boehm system (clarinet)]]\n* [[Boehm garbage collector]]\n* [[Behm]]\n* [[Bohm (surname)]]\n* [[B hm (disambiguation)]]\n* [[B hmer]]\n\n{{Category:German-language surnames}}\n{{Category:Surnames of Czech origin}}"
```

To extract specific data from JSON using the jq command

curl -s https://en.wikipedia.org/w/rest.php/v1/page/boehm | jq '.id, .title, .latest.timestamp'

```
(base) admin@users-MacBook-Air ~ % curl -s https://en.wikipedia.org/w/rest.php/v1/page/boehm | jq '.id, .title, .latest.timestamp'
2574858
"Boehm"
"2020-11-12T22:41:47Z"
(base) admin@users-MacBook-Air ~ %
```

To create a new paste using -d to post data to the paste

curl -X POST -d 'api_dev_key=jbnPnEo90_UXc1IJbh1DJccT09IQxZ2H' -d

'api_paste_code=66286378

"Bohem"

"2021-10-25T01:41:35Z" -d 'api_option=paste' "https://pastebin.com/api/api_post.php" -d

'api_paste_name=boehm'


```

(base) admin@users-MacBook-Air ~ % curl -X POST -d 'api_dev_key=jbnPnEo90_UXc1IJbh1DJccT09IQxZ2H' -d 'api_paste_code=66286378
"Bohem"
"2021-10-25T01:41:35Z" -d 'api_option=paste' "https://pastebin.com/api/api_post.php" -d 'api_paste_name=boehm'
https://pastebin.com/dbEGM9rh
(base) admin@users-MacBook-Air ~ %

```

Post in the browser

https://pastebin.com/dbEGM9rh



boehm

A GUEST

NOV 21ST, 2021

1

NEVER

SHARE

TWEET

text

0.04 KB

raw

download

clone

embed

print

report

1.

66286378

2.

"Bohem"

3.

"2021-10-25T01:41:35Z"

To extract only the header write -I
curl -I <https://pastebin.com/dbEGM9rh>

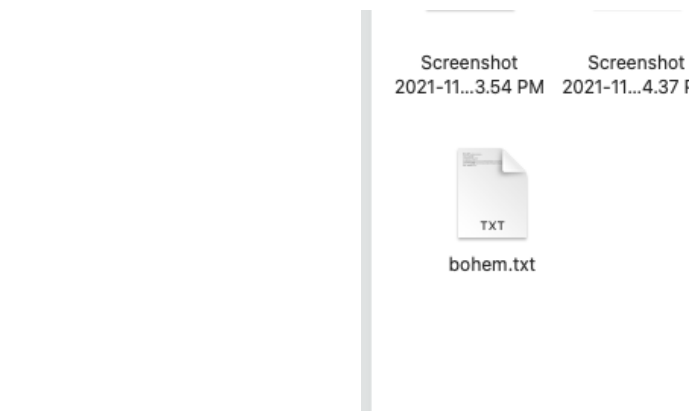
```
(base) admin@users-MacBook-Air: ~ % curl -I https://pastebin.com/dbEGM9rh
HTTP/1.1 200 OK
Date: Sun, 21 Nov 2021 11:47:49 GMT
Content-Type: text/html; charset=UTF-8
Connection: keep-alive
x-frame-options: DENY
x-content-type-options: nosniff
x-ssr-protection: 1; mode=block
set-cookie: csrfr-frontend=2d69534c22026155ec7cf53a835531fe4e992af544fb9facac42fb92d706e9%3A2%3A%7B1%3A%0%3B%3A14%3A%22_csrfr-frontend%22%3B1%3A1%3B%3A32%3A%22eB71qqmTp1IvGwfmR720uaQS44PbU
Ziz%22%3B%7D; path=/; HttpOnly
CF-Cache-Status: DYNAMIC
Expect-CT: max-age=604800, report-uri="https://report-uri.cloudflare.com/cdn-cgi/beacon/expect-ct"
Server: cloudflare
CF-RAY: 6b19af78be6cc272-FRA

(base) admin@users-MacBook-Air: ~ %
```

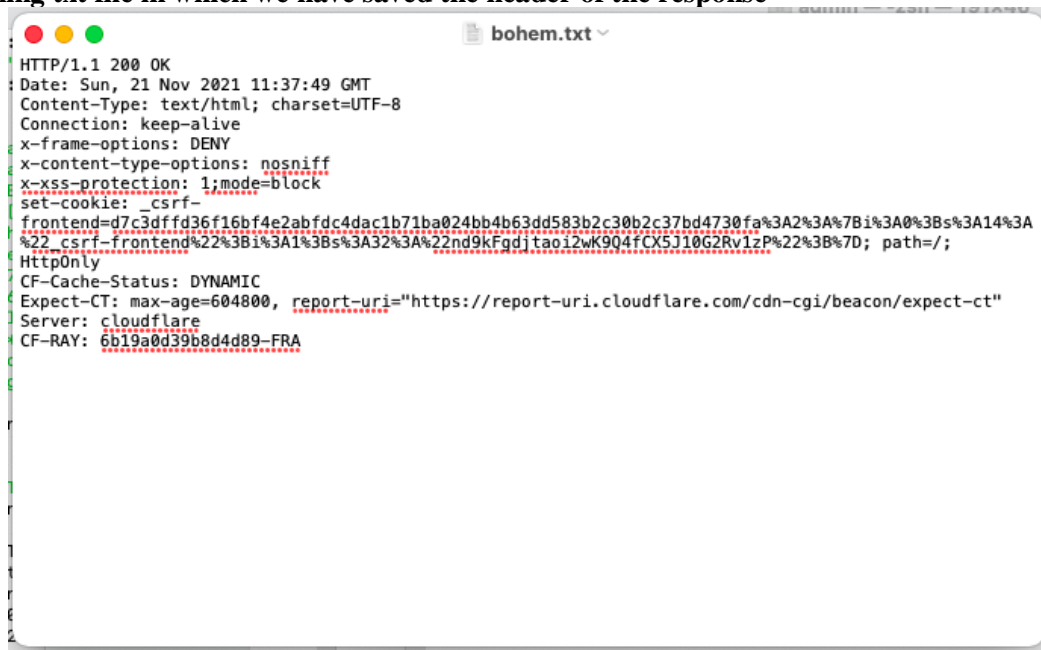
To write to a txt file use -o and specify file name after it
 curl -I https://pastebin.com/dbEGM9rh -o boehm.txt

```
(base) admin@users-MacBook-Air ~ % curl -I https://pastebin.com/dbEGM9rh -o boehm.txt
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           %             %         Dload  Upload  Total   Spent    Left  Speed
  0     0     0     0     0     0      0      0 --:--:-- --:--:-- --:--:--     0
(base) admin@users-MacBook-Air ~ %
```

Saving text file



Opening txt file in which we have saved the header of the response



3 Short questions: (12 points)

3.1. Explain in your own words what is the difference between the Internet and the World Wide Web? (4 points)

1 The Internet is the extensive network of network, in which hosts are connected to each other via LAN or WAN to transfer or communicate with each other and on the other hand, world wide web is the repository of the information which can be access via internet.

2 Internet is governed by the TCP/IP or UDP Protocol, whereas WWW uses the HTTPs protocol to transfer the data/information between the web applications.

3 Internet works on the Link, Internet and Transport layers and WWW works on the application layer.

4 The main components of the Internet are mac address, Ip address, and ports and main components of the world wide web are URL, HTML, and HTTP to access the resources over the internet.

3.2. Explain how the status of the HTTP GET request can be interpreted? (4 points)

The http responses are divided into five categories. Each category has defined the different purposes of the response and it can be interpreted by the first digit of the status code, which defines the classification of the responses. While the other two digits do not have any categorization. Below are the five classes defined:

1. 1xx – It is the information status response, and it will indicate that request was received, but still in progress or in a continuity state.
2. 2xx – It is the successful status response, and it will indicate that request was received, processed, and accepted successfully.
3. 3xx – It is the redirection status response, and it will indicate that further actions need to be taken to complete or process the request.
4. 4xx – It is the client-side failure status response, and it will indicate that request cannot be processed due to the bad request, bad syntax, and resource not found on the server.
5. 5xx – It is the server-side failure status response, and it will indicate that server cannot complete and process the request successfully, due to any technical or server failure.

Below is an example of the http response of the HTTP GET request received from the server.

```
HTTP/1.1 200 OK
Content-Type: text/html
```

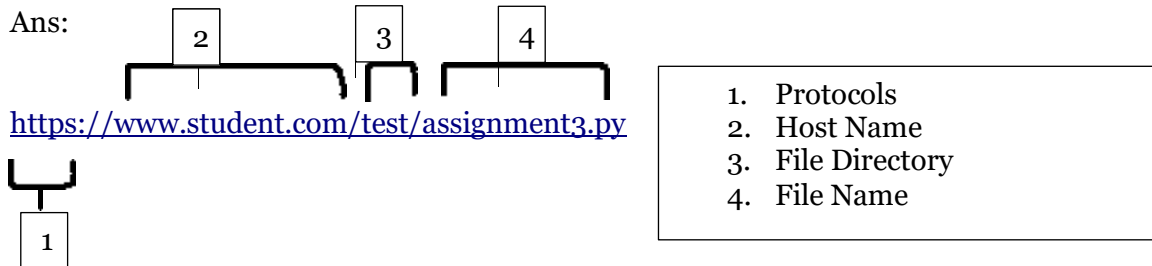
Whereas the first line indicates that the server accepts the protocol that client wants to communicate with the **HTTP/1.1** along with the status code **200 ok** followed by the header.

If the server did not find the requested resource, then it will send the response with the status code of 404(not found).

```
HTTP/1.1 404 NOT FOUND
```

3.3. Explain which parts of an URL are processed by the client, the server and the Internet? (4 points)

Ans:



Protocols, Ports, and Fragments are processed by the client web browser.

- a. Protocols defines how the web browser can send or request the data from the server. For example, **http(s)://**
- b. Ports: It is usually not visible on the URL when the Protocols contains the http or https request it is predefined to the 80 and 443 port numbers.
- c. Fragments: It is a internal part of the web pages, it refers the section of the web page to redirect. The fragments are not sent in a http request. They are the part of the web browser and process by the client only.

Resource path and parameter queries are processed by the server to find the resource in a host on a specific path.

- a. Resource path: It is referred to the path or directory of the file in a web server. So, that's why it is process by the server to find the specific file in a request directory.
- b. Query Parameters: It is an optional part of the URL. Which is used to search or get some specific information from the web server along the directory path. Therefore, it will process by the server.
- c.

Complete hierarchical host name is processed by the internet.

- a. Host Name/Domain Name: Certain host contains the files on an internet. Host can be a single machine connected to the internet. So, it is important to provide such name to those host on an internet. Host name is associated with an Ip address of a host on an internet by the DNS system. Therefore, it is process by the internet.