UiA Universitetet i Agder

DAT234

Mandatory Python Scripting

11.october 2019

Elias Kløverød Brynestad

Kristoffer Slettebakken

Phuong Ha Thi Pham

Contents

1 Introduction

The mandatory delivery consisted of six tasks within Python Scripting. For each task, there were only allowed to use the packages *requests* and *re*. The following tasks were:

1. Visit the site https://links.datapor.no (Lenker til en ekstern side.) with the use of requests and print out the source code.
2. How many links are using http vs https , with the use of regex.
3. Print out all the unique TLD.
4. print out all the unique hostname for each url.
5. Find all the unique html tags, with the use of regex
6. Visit each of the links found on https://links.datapor.no (Lenker til en ekstern side.), and visit each of the links and print out the title of the site *hint: look for <title> or <meta> tags*

2 Tools

The tasks were solved using the Python 3.7.3 programming language (www.python.org). The code was written using the Pycharm version 2019.2.1 IDE by JetBrains (www.jetbrains.com). All students in the group used the same software.

3 Task

3.1 How the tasks were solved

To get started with the tasks, it was necessary to research on the requests and re packages in addition to study the source code of the website to manage to harvest it. For the research of requests, the website Real Python was used (Ronquillo, A. 2019). A blog from Medium (Ahmad, Z., 2018) and a YouTube video from sentdex (sentdex, 2014) were used for studying the re package.

The source code of datapor's website was frequently used to ensure that the outputs of the tasks were correct.

3.2 Problems that were encountered

When solving the tasks, there occured a problem regarding task 6, where uia.instructure.com could not be extracted. When trying to harvest the source code, the following error message occured:

```
{"status":"uautentisert","errors":[{"message":"brukergodkjenning kreves"}]}
```

The problem was most likely that the website demands a login with Feide, thereby the status was unauthorized and user authentication was required.

Other problems that was encountered was that not all group members had the opportunity to meet while solving the tasks because we take different courses in our study. This, combined with a lack of communication, lead to unequal distribution of work. To prevent this from happening in the future the intent is to plan early on in the week when everyone is able to come together to work and what needs to be done, so that the workload can be properly distributed depending on schedules.

4 Summary

Because of good research and use of information available on different websites, we were well prepared for the tasks. What became a difficulty was the fact that our schedules were not compatible. A result of this was an uneven share of workload between the group members as when we were nearing the deadline some of the group members became unavailable due to other courses in the study. As a group we have come up with a plan to prevent this from happening in the future. To sum up, solving the tasks were not too difficult and other tasks further down the line will be properly distributed between the group.

5 References

Ahmad, Z. (2018, April 4). *Regular Expressions - The Last Guide*. Retrieved from https://medium.com/tech-tajawal/regular-expressions-the-last-guide-6800283ac034

Python 3.7.3 Programming Language
https://www.python.org/downloads/

PyCharm 2019.2.1 IDE
https://www.jetbrains.com/pycharm/download/?_ga=2.40414884.760240466.1570708841-886158132.1567419535#section=windows

Ronquillo, A. (2019, January 21). *Python's Requests Library (Guide)*. Retrieved from https://realpython.com/python-requests/.

Sentdex, (2014, July 20).
Python 3 *Programming tutorial - Regular expressions / regex with re*
Retrieved from https://www.youtube.com/watch?v=sZyAn2TW7GY