ENGINEERING SCRAPING TEST - Idio

Individual Report

Shreevathsaa Mahadevan
shreevathsaa@gmail.com

1. INTRODUCTION

This report describes how we extracted accurate information in a reliable way and technologies, libraries that are used for the implementation.

2. TOOLS AND TECHNOLOGIES USED

The programming language that we have used here is python and we have executed the python code through Jupyter Notebook.
The steps for installing Jupyter Notebook is given in the below link:
`https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/install.html`

There are several libraries used here for scraping the data from the URL. The packages used are:

- ✓ urllib
- ✓ BeautifulSoup, bs4
- ✓ requests, collections
- ✓ PyPDF2

3. DISCUSSION

I have chosen the Examples 1,2,3,5 to solve and as I couldn't login and register to the website, I couldn't scrap the Example 4.
All the five problems were really interesting and had to work on it so deeply to get the desired output. Here we have used beautifulsoup, bs4 and urllib which is mainly used to read and extract the information from a specific url. The soup.find function is mainly used to retrieve the class file under which the contents were stored in. To know the class file name, we have tried looking up the website by inspect the page source. We have also used here PyPDF2 which is a python PDF library which can be used for splitting, merging, cropping, and transforming the pages of PDF files.

3.1. Accomplishments

I have tried to get the contents for the given URL as mentioned in the examples.

- ✓ I was able to get the Title of the page as mentioned
- ✓ I was able to retrieve the body content of the page.

- ✓ With respect to the images, I was able to retrieve the images from a content item as image reference URLs
- ✓ Handled the HTTP error for the URL given in the Example 1
- ✓ Able to retrieve the pdf file contents.

3.2. Running the code

I have uploaded the code as a Jupyter file(.ipynb) in the github where you can open it directly to see the output. The another way is to download the file and open it in the Jupyter notebook installed. For installing instructions, please follow the above link provided in the tools and technologies used section. Please make sure to install all the required libraries mentioned to run the code.

GITHUB LINK: https://github.com/shreevathsaa/Idio-Test

4. CONCLUSION

Based on the discussions made, I would like to infer that this test was really interesting. It has motivated me to a certain extent of how to scrap data in a very efficient way. I strongly believe that the learning outcomes will be definitely useful in future when performing similar work.