Dev Club Winter Assignments

Assignment 3: Web scraping in Python

9th December 2019

Next Assignment Release : **12 December 2019**

P.S If stuck, join this Slack channel and ask us questions directly here

1 Introduction

You love Cyanide and happiness comics, but are about to go vacationing on a remote place (with no internet) for a month. You need a month's stock of comics for your comic relief, but manually downloading such a huge number can be exhausting and time-consuming.
Your task is to download comics from explosm.net/comics/archive using web scraping in Python. The libraries you are going to use are *requests* and *BeautifulSoup*. Though you can any text editor (like VS Code) for writing the Python script, do check out **PyCharm** and **Jupyter notebook**.

2 Resources

Before starting with this assignment, you should learn about the basics of Python and web scraping. Here are some tutorials to get you started.

1. Python basics
   - `http://learnpython.org/`
   - `http://kaggle.com/learn/python`
   - `http://tutorialspoint.com/python/index.htm`
2. Requests library
   `http://realpython.com/python-requests/`
3. Web scraping in Python using BeautifulSoup
   - `http://geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/`
   - `http://medium.com/@pknerd/write-your-first-web-scraper-in-python-with-beautifulsoup-564dddd8693c`
   - `http://medium.com/python-pandemonium/6-things-to-develop-an-efficient-web-scraper-in-python-1dffa688793c`

Also make sure you know how to inspect elements in a website using Developer tools (`developers.google.com/web/tools/chrome-devtools/dom/`)

3 Specifications

You will be provided with an input file *input.txt*. Assume this file to be present in the same directory as your python (.py) file.

input.txt
start_month start_year
end_month end_year
author1 author2 ... authorN

Line 1 contains a string start_month followed by a space, and an integer start_year. Similarly Line 2 contains end_month followed by a space, and end_year. Line 3 contains N space separated strings ($N \geq 1$) where each string denotes the firstname of an author. You have to download only those comics which lie in the given time range (start month, start year to end month, end year ; both inclusive) and which are written by the authors specified.

These comics are to be downloaded in a hierarchical folder structure, with a comic of a specific month and year should be kept in the directory year/month/. Create these directories alongside your python file, i.e. inside the same directory as your python file. The name of each comic strip should be of the form date-authorFirstName.png. e.g. For this comic , create a directory *2019* alongside your python file, inside it a subdirectory *december* and save the comic as 2019.12.06-Rob.png, so the final structure looks like *2019/december/2019.12.06-Rob.png*

Final output structure:
year1/month1/dateA-authorAFname.png
year1/month1/dateB-authorBFname.png
year1/month2/dateC-authorCFname.png
...

Here's an example

Sample input.txt
december 2018
december 2018
Dave

Sample output
./2018/december/2018.12.01-Dave.png
./2018/december/2018.12.04-Dave.png
./2018/december/2018.12.08-Dave.png
./2018/december/2018.12.11-Dave.png
./2018/december/2018.12.15-Dave.png
./2018/december/2018.12.18-Dave.png
./2018/december/2018.12.22-Dave.png
./2018/december/2018.12.25-Dave.png
./2018/december/2018.12.29-Dave.png

./2018/december/2018.12.31-Dave.png

BONUS (OPTIONAL) :

1. Download a random comic strip using the site's random comic generator (explosm.net/rcg). Since the random comic appears as 3 separate image files, save 3 seaparate image files in a directory *random*.

   input.txt
   Random

   Output structure:
   ./random/frame1.png
   ./random/frame2.png
   ./random/frame3.png

2. Download the latest N comics from the site ($N \geq 1$). Save them in a directory *latest*

   input.txt
   latest N

   Sample input.txt
   latest 2

   Sample Output structure:
   ./latest/2019.12.06-Rob.png
   ./latest/2019.12.04-Kris.png

4 Conclusion

The purpose of this assignment is to familiarise you with Python and basics of web scraping. Don't worry if you are unable to complete some of the tasks, and don't worry too much about the strict input/ouput guidelines either. What's important is that you LEARN and enjoy the essence of coding and development! Feel free to modify the input and output formats, but do mention them as comments in your code.
As usual, submit the link to your git repository (containing only the .py file) here