Data Scrapping with Python

By

Data Scrapper

Contents

[Code structure 2](#_Toc3898276)

[Troubleshooting 4](#_Toc3898277)

[References 5](#_Toc3898278)

# Code structure

Under this heading you will find certain sample of our current code with the explanation.

Without further delay:

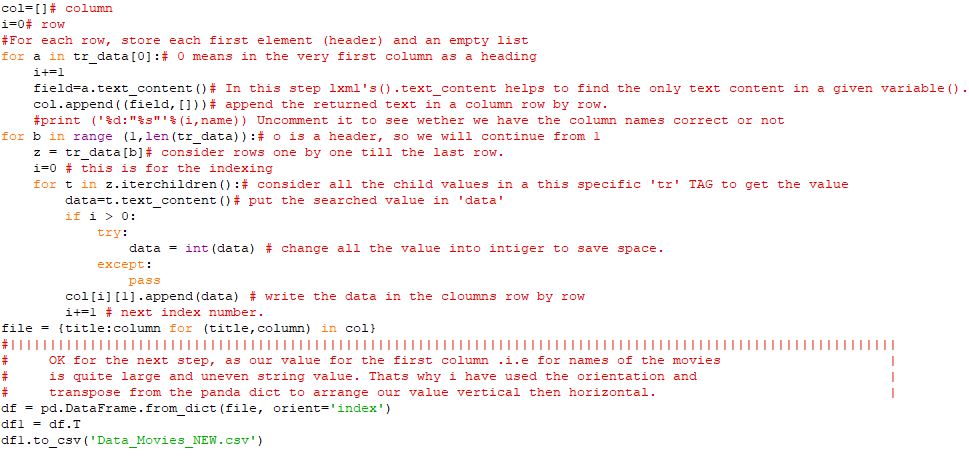
* How does our code works:
  + our code is solely written in python language.
  + We have used several libraries like numpy, beautiful soup, urllib, lxml, requests which will be discussed in broad further in this document.
  + Our code works in following steps:
    - First, it requests the page which we want to scrape.



* + - This page is them stored in a variable in the form of html document.



* + - This document is then searched for the specific html tags which will retrieve the desired table for scarping.
    - This table is then iterated thoroughly row by row to scrape the data into a csv file.



* What important library we have used:
  + Numpy as np:
    - Import the library numpy as reference np.
    - Numpy is a library used for scientific calc.
    - WHY: to put the parsed value from lxml into columns and rows of our table.
  + Pandas as pd:
    - Import library pandas as alias pd.
    - This library is responsible for data manipulation and analysis.
    - To read or write a data to an output file.
  + Urllib.requests:
    - From urllib library use the request sub-lib to import urlopen.
    - We are using this lib only to get the page in a html format variable.
  + Beautiful Soup:
    - From beautiful soup import Beautiful Soup.
    - This lib is used for the manipulation and to search the specific tags in html page from web.
  + Lxml.html:
    - This lib is used to convert the html page to a string format.
  + Requests:
    - This lib is used to request the web page from internet.

# Troubleshooting

* What problems we have faced during my scraping:
  + It is illegal to scrape data on most of the websites in Canada. So, we have to carefully inspect the site rules for scraping. At the end we chose the Wikipedia for our scarping because it is an open source project.
  + Data requirement for this project was to have about 1500 data entry to operate a machine learning algorithm. So, to have a data with that mush entry is nearly impossible but we successes to have approx. 1300 entry.
  + Now, came the challenging part. Making the prediction based on the dataset which follows the upper written rules. We must find a specific data set in which we can predict uncertainty.
    - For example, in our data set we have successfully predicted the number of awards based on the nomination.
* Problems we faced with library were:
  + First of all, we did not know which library to use.
  + When we came across different libraries for the scraping, we did not know how to use them, what were their purpose and to what extant we can use these libraries.
  + Then after watching YouTube videos and reading about them on the internet and trying out the examples given on the internet, we were finally able to use them according to our needs.
* Data conversion issues:
  + After getting the data scraped into a variable, we did not get the desired output.
  + The data we are getting was very disoriented. We are getting the data in horizontal manner in the csv file.
  + Then we used the orientation option of the panda library to orient it vertically.

(Wikipedia, the free encyclopedia, n.d.)

(Nazrul, 2018)

(Beautiful Soup Documentation, n.d.)

(Bhanot, 2018)

# References

*Beautiful Soup Documentation*. (n.d.). Retrieved from crumm: https://www.crummy.com/software/BeautifulSoup/bs4/doc/#installing-beautiful-soup

Bhanot, K. (2018, 10 1). *Dataset creation and cleaning: Web Scraping using Python — Part 1*. Retrieved from towardsdatascience: https://towardsdatascience.com/dataset-creation-and-cleaning-web-scraping-using-python-part-1-33afbf360b6b

Nazrul, S. S. (2018, July 25). *Web Scraping HTML Tables with Python*. Retrieved from towardsdatascience: https://towardsdatascience.com/web-scraping-html-tables-with-python-c9baba21059

*Wikipedia, the free encyclopedia*. (n.d.). Retrieved 3 19, 2019, from Wikipedia, the free encyclopedia: https://en.wikipedia.org/wiki/List\_of\_highest\_paid\_film\_actors

https://www.youtube.com/watch?v=0\_VZ7NpVw1Y