UNIVERSITY OF WARSAW
2400-DS1WSMS /WEB SCRAPING AND SOCIAL MEDIA SCRAPING
2020-2021 SUMMER / PROJECT DESCRIPTION

WATERSTONES.COM
BESTSELLERS BOOKS SCRAPING

Submitted by:


Mustafa Sanli


Muhammed Burhan Kurt

Principal Investigator: Anna Lewczuk

09.05.2021

1. Overview

A data scientist is the one who understands data best. And he should know all aspects of the data. It should also be able to generate as much data as it can process the data. In this project, we extract data from waterstones.com and analyze the data we obtain. In that project we used BeautifulSoup(bs4), selenium libraries and scrapy framework. In the BeautifulSoup and selenium we just scrapy from the waterstones.com and we use xpath format and also in the scrapy we use css format to extract data. In the all libraries we used after scrapped data we covert them to csv file to make analyze. In the analyze part, we just show the basic data analysis and visualization.

(https://github.com/M-Sanli/WaterStones.com_Scraping)

2. Approach and Methodology

This project includes;

- the libraries and codes used in the scraping waterstones.com website, analysis price of the data set that we scraped in the Python language. Mainly *selenium, BeautifulSoup(bs4), pandas, mathplotlib* libraries
- *Scrapy* framework

3. Objectives

As a data scientist, our goal is to successfully extract data from any website using scraping methods. Data extraction was done using the methods we learned in the Web Scraping and Social Media Scraping course of the University of Warsaw, Data Science and Business Analytics. When performing data scraping, since the data is also saved as a csv file, it can be easily analyzed later.

4. Participants

a. Muhammed Burhan KURT | 428168

   -Scrapy, Selenium, Analysis of the data and Project Description

b. Mustafa SANLI | 436656

   -BeautifulSoup, Selenium, Analysis of the data and Github