Python Programming

Student Name and Surname-Student Number

Osid Alsagir-160316052

Supervised By

Dr. Öğr. Üyesi Hakan Murat KARACA

Web Crawling Homework Report

Web Crawling using python programming language

The program developed by python programming language using Scrapy library this library used for crawling websites in easy way

Introducing the program:

The program crawls two different sites ,gets data and save them in files the files to be saved are three files from two types so for every site the data will be saved it three files the first file is a text file for saving the all URLs that the program visit and the second file is also a text file that saves the important data in lines but the last file is a json file that saves the data and visited URLs the json file can be used to transfer data to another program or application .

In every time we run the program the files will be initialized and naming every file with the time and date

Now there is two spiders the imdb_spider and n11_spider they visit IMDb site and N11 site respectively .

IMDb:

The program gets the top rated movies from the site and gets movie's name, director's name and the rate from 10 list them inside text file and json file **crowler\TopRatedMovies** in this directory the program save the files.

"There is an example files there"

N11:

The program gets the " cok-satanlar" from the site and gets the product's name and price list them inside text file and json file **crowler\n11product** in this directory the program save the files.

"There is an example files there"

How the program works:

IMDb (imdb_spider.py):

The program goes to this URL https://www.imdb.com/chart/top/?ref_=nv_mv_250 and opens the HTML file in this file the program try to find the CSS classes and find the <a> attribute and get the URL inside it, after that this URL will be visited from the parseData method and try to find the CSS classes and that contains the title of the movie ,director and the rate as well, then these data will be saved in three different files that we talked about and it still do that until the main URL has no more <a> attributes to visit

N11(n11_spider.py):

The program goes to this URL https://www.n11.com/cok-satanlar and opens the HTML file in this file the program try to find the CSS classes and find the <a> attribute and get the URL inside every product, after that this URL will be visited from the parseData method and try to find the CSS classes and that contains the title of product and price, then these data will be saved in three different files that we talked about and it still do that until the main URL has no more <a> attributes to visit

Methods and how they work:

IMDb (imdb_spider.py):

`start_requests()` This method will be called when the program starts, and it has other called methods and they are `refactoringJsonFile and refactoringTextFiles` These two files is to initialize the JSON and TXT files and make them as the given names.

`parse` This method will be called after initializing files so this method will goes into the given URL and gets the other URLs that this page contains, after this for every URL found it will be saved by calling `writtingToFile` to URL text file then the URL will be transferred to `parseData` method to extracts the title and other data and save them in TXT and JSON by calling `writtingToJsonFile` and `writtingToFile` threading used in this step to make saving to files at the same time that makes the program faster than usual .

N11(n11_spider.py):

This spider is not different from the IMDb mechanism but because we have tabs in this site se we had to visit every tab there, so `parse` method is responsible now to visit all the tabs and `parseUrls` method to get the products from the single tab .

In conclusion : After installing scrapy we can run this program after getting to the project directory "**crowler>**" and write the commands on the terminal :

For runing IMDb spider : scrapy crawl TopRatedMovies

For runing N11 spider : scrapy crawl n11product