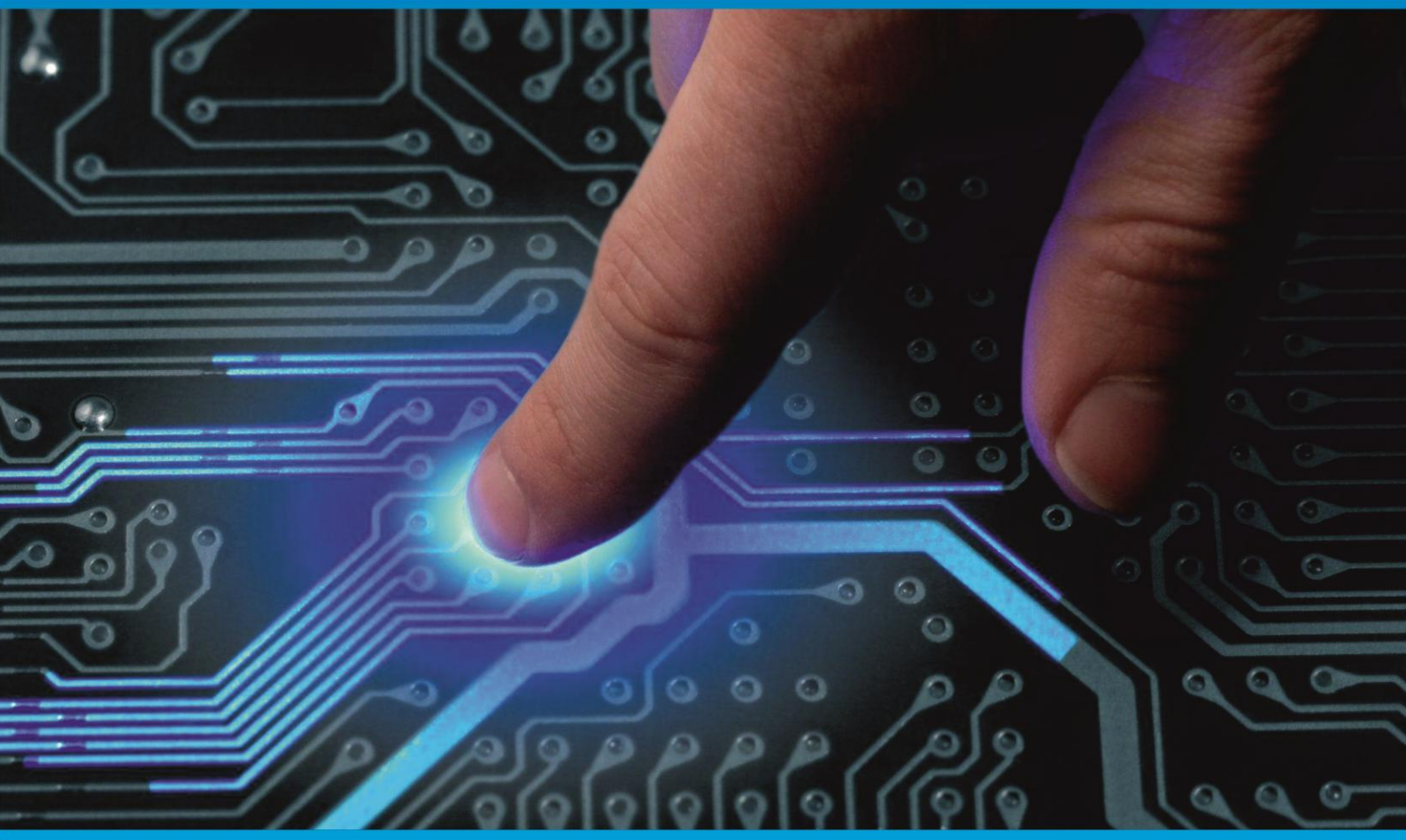
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 8, Issue 10, October 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 10, October 2020 ||

Web Scraping of Educational Websites using Beautiful Soup

Prof. Kshirsagar Sarika

Assistant Professor, Dept. of MCA, JSPMs Jayawantrao Sawant College of Engineering, Hadapsar, Pune, Maharashtra, India

ABSTRACT: Today data is very important for changing the world.It helps for cure the disease or disorder, improvement of company profit, decision making of organisation.We got lots of data from various sources like websites,institutes. Hospitalsetc. data from sources are unstructured. How to get data in structured format? Web scarping gives directed structured data with excel,csv,tsv or any required format.so in this paper we study the web scraping and how scrape the educational websites using web scraping tools and how education institutes boost the ranking.It will help to take decision by comparing education details.

KEYWORDS:Web scraping. Education, website, python.

I. INTRODUCTION

Web scraping is an automated tool for finding and extracting data from on-line sources. It utilizes computer programming software and customized software code to mine data or other information from on-line sources in order to remove a copy of the data and store it in an external database for analysis. Typically, the data harvested through web scraping is analysed to answer questions that could not be answered, or answered efficiently, using the data as it was originally presented on-line. Essentially, web scraping is a way to pull information from particular web pages and re-purpose it for customized analysis[1]. Web scrapping is a process of automatic data and information collection from the internet, commonly in website pages using mark-up languages such as HTML or XHTML whose data analysed for certain needs and purposes[2]. Web scraping data is used for personal work or organizational work. InIndia there are near about 31000 educational institutes in India for various courses. We see that more competitions in private institutesfor admission and their ranking.Thus, for marketing of institution is very important in this era.Online marketing analyst use web scraping methods to grab some information from other competitors such as emails, targeted keywords and links and also traffic source.[3]

II. WEB SCRAPING SOFTWARE

There are lot of software available for scraping the website. But due to huge amount of data available on websites some software's not properly working. So there are common software programming languages like R and Python are typically used to write the software code for both the crawler and the scraper. Hence, software programming skills are essential for building and deploying a web scraper. The software code, however, is constructed based on specific search and data extraction criteria established by the researcher based on his/her understanding of the on-line data source(s) of interest and the research questions the analysis will attempt to answer[1].

III. RELATED WORK

saurkar et al.[4]Focused on the overview on the information extraction technique i.e. web scraping, different techniques of web scraping and some of the recent tools used for a web scraping. They discussed the basic of web mining and focused on the techniques used for web scraping.Crystal Pereira [3]discussed the tools and techniques used in scraping and its impact on the social networks.they gave thelist of tools which is useful for web scraping.Vargiu et al. [5] They proposed a collaborative filtering-based Web advertising system aimed at finding the most relevant ads for a generic Web page by exploiting Web scraping. To illustrate how the system works in practice, a case study is presented.gaikwad et al.[6] proposed an Education Search Engine in two-stage technique, namely Smart Crawler, for efficient gathering deep web interfaces. To achieve more accurate results for a focused crawl, Smart Crawler ranks websites links to prioritize highly relevant result in websites link rankings. The results showed that the smart crawler and scarper can realize the high-efficient and flexible data collection function, and laid the foundation for Web data

mining. This efficiently retrieves web data mining interface from large-scale sites and achieves higher.Gupta et al.[7] applied the proposed algorithm on live dynamic web pages of patent portal to compare the result against existing web extraction algorithms and found to be more efficient in terms of throughput.Ashiwal[8]developed method for retrieving web information using BeautifulSoup and python script. BeautifulSoup is tool for web information retrieval. Most of the web information presents in unstructured format. The proposed system retrieves the unstructured data in user's pattern and makes it useful.

IV. METHODS FOR WEB SCRAPING

For web scraping we use education websites', g, shiksha.com website used for scraping. We have filtered the MCA collegesfrom all colleges. Select one college from filtered colleges.e.g. we selected Savitribai phulepune university. When we inspect the college we got HTML elements file.Which included all HTML tags which as follows.

Fig 1:HTML tag file

above fig contains all required tags**.**from this we identified all the elements of any college.now we implementing the python language for exucution and xtracting the information.

To extract the HTML code, extractor relies on Web scraping through two specialized libraries: HTMLParser and Beautiful Soup. HTMLParser defines a class HTMLParser that serves as the basis for parsing text files formatted in HTML and XHTML. The class is instantiated without arguments and its instance is fed by HTML data and calls handler functions when tags begin and end. The class is meant to be overridden by the user to provide a desired behavior. BeautifulSoup is a Python library that parses broken HTML. BeautifulSoup is not a real HTML parser but uses regular expressions to dive through tag soup. The main features of BeautifulSoup are: it yields a parse tree that makes approximately as much sense as the original document, in case of the programmer gives it bad markup; it provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree; and it automatically converts incoming documents to Unicode and outgoing documents to UTF-8[5].

Follwing are the steps for web scraping

1.Install the BeautifulSoupin jupyter notebool.use

!pip install BeautifulSoup

2. Get the URL using request object

3. use BeautifulSoup object for extracting the content of URL.

4.soup object will find the all education URL elements.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 10, October 2020 ||

5.form that education elements.we only extract the following elements

a.Title of educational institutes

b.Course name

c..rating of the course

follwing screen shows the coding of the extraction of web scraping.

```
In [1]: import requests
        from bs4 import BeautifulSoup

        URL = "https://www.shiksha.com/it-software/colleges/mca-colleges-
        r = requests.get(URL)

        soup = BeautifulSoup(r.content, 'html5lib') # If this line cause
        print(soup.prettify())
```

Follwing screen shoes the output after scraping

```
In [1]: import requests
        from bs4 import BeautifulSoup
        import csv

        URL = "https://www.shiksha.com/it-software/colleges/mca-colleges-
        r = requests.get(URL)

        soup = BeautifulSoup(r.content, 'html5lib')
        edu_elements = soup.find_all(id='main-wrapper')
        for edu_element in edu_elements:

            title_elem = edu_element.find('div', class_='elipsysBox')
            # company_elem = edu_element.find('div', class_='company')
            block_elem = edu_element.find('label', class_='blockLabel')
            value_elem = edu_element.find('div', class_='valueTxt')
            rating_elem = edu_element.find('span', class_='ctpv2-rating')
            if None in (title_elem,block_elem,value_elem,rating_elem):
                continue
            print(title_elem.text.strip())
            print(block_elem.text.strip())
            print(value_elem.text.strip())
            print(rating_elem.text.strip())
            print()

UNIPUNE - Savitribai Phule Pune UniversityGaneshkhind, Pune
MCA Courses
1 Course
3.6
```

```
In [ ]:
```

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 10, October 2020 ||

V. CONCLUSION

There are manyeducational websites available but data is in unsrucrured format .In this paper we extracted the data using web scraping.It is very useful for researcher;academician and education institute. Institute also increase study all remaing institute data and increase their rating .It gives the data in structured format.the results shows that web sraping method gives the accurate result and it is very useful for enhacing the rating of institutes.

REFERENCES

[1] E. J. Farley and L. Pierotte, "An Emerging Data Collection Method for Criminal Justice Researchers," pp. 1–9, 2017.

[2] C. Slamet, R. Andrian, D. S. Maylawati, Suhendar, W. Darmalaksana, and M. A. Ramdhani, "Web Scraping and Naïve Bayes Classification for Job Search Engine," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 288, no. 1, 2018, doi: 10.1088/1757-899X/288/1/012038.

[3] R. Crystal Pereira and T. Vanitha, "Web Scraping of Social Networks," *Int. J. Innov. Res. Comput. Commun. Eng. (An ISO*, vol. 3297, no. 7, pp. 237–240, 2015, [Online]. Available: www.ijircce.com.

[4] A. V Saurkar and S. A. Gode, "An Overview On Web Scraping Techniques And Tools," *Int. J. Futur. Revolut. Comput. Sci. Commun. Eng.*, pp. 363–367, 2018, [Online]. Available: http://www.ijfrcsce.org.

[5] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering- based approach to web advertising," *Artif. Intell. Res.*, vol. 2, no. 1, pp. 44–54, 2012, doi: 10.5430/air.v2n1p44.

[6] R. R. Gaikwad and M. Bhonsle, "Enhanced Indexing and Scraping for Educational Search Engine using Web Usage Mining," vol. 5, no. 3, pp. 404–411, 2018.

[7] G. Gupta and I. Chhabra, "Optimized Template Detection and Extraction Algorithm for Web Scraping of Dynamic Web Pages," *Glob. J. Pure Appl. Math.*, vol. 13, no. 2, pp. 973–1768, 2017, doi: 10.1016/j.stem.2013.12.005.Human.

[8] P. Ashiwal, P. Tripathi, and R. Miri, "Web Information Retrieval Using Python and BeautifulSoup," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 4, no. VI, pp. 335–339, 2016.

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details