BANGLADESH UNIVERSITY OF BUSINESS AND TECHNOLOGY (BUBT)

Project Report On-

"Development of a Web Scraper to get news in oneclick"

COURSE CODE: CSE 352

COURSE TITLE: Artificial Intelligence and Expert System Lab

SUBMITTED BY

Name	ID	Intake
Aminur Rahman	18191203013	31
Shezan Mahmud	17183203041	30
Anisur Rahman Anas	17182203031	29
Md.Topu Raihan Robin	18192203036	32

SUBMITTED TO

Mr. Shamim Ahmed

Assistant Professor,

Department of Computer Science & Engineering

BUBT

Submission Date: 28th, July 2022

ABSTRACT

With the world digitizing every day, the importance of collecting and structuring data is increasing. The manual of collecting information is not advanced enough to go through thousands if not millions of news sources every day. So, the use of the web scraping technology eases the difficulty by building a web news scraper that allows easy accessto different news in a single click.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION **3**

1.1 Web scraping 3

1.2 Scraping news using Web scraping technology 4

1.2 Technical Concepts **4**

1.2.1 Django 4

1.2.2 HTML 4

1.2.3 CSS 5

1.2.4 JavaScript 5

1.2.5 Beautiful Soup 5

1.3 Objective **5**

1.4 Motivation **6**

Chapter 2 - Background Study **7**

2.1 An Intelligent survey of personalized Information Retrieval using web scraper [2] 7

2.2 Socially Smart - an Aggregator System for social media using Web Scraping [3] 8

2.3 Exploiting web scraping in a collaborative filtering based approach to web advertising [4] 9

Chapter 3 - Methodology **11**

3.1 System Diagram 11

3.2 Steps of solving the problem 12

Chapter 4 - References **13**

CHAPTER 1: INTRODUCTION

1.1 Web scraping

Web scraping typically extracts massive amounts of data from websites. This technology may directly access the World Wide Web using the Hypertext Transfer

Protocol or a web browser. This is an automated process of gathering data in order to synchronize them for future usage. It is a form of extracting data and putting them in alocal database so that further analysis and manipulation can be done on the data that was extracted from the websites. Web scraping is used for various reasons like aggregating financial data, monitoring consumer behavior and price, analyzing the market, tracking news etc.

Web scraping a web page means fetching the data from the web page and extracting the data from the page. The process starts when a user downloads a page or in other words views a page. When the downloading or fetching the page is done, the extraction

process can start. The fetched data is then copied to a local database and later the datacan be shown in a different form. [1]

Figure 1: Web Scraping process

1.2 Scraping news using Web scraping technology

Among a lot of uses, a very important job done by the web scraping technology is scraping the news websites. It allows a user to have access to a lot more information atthe same time. Using web scraping technology, different data from different websites can be shown in a single webpage. This will save a user a lot of time, energy and otherresources, and will also fasten the process of collecting data.

The process has been done with the help of web scraping, so the method is similar to the web scraping technique. The headlines of three Bangladeshi online news portals are fetched and then extracted so they can be shown in a web app.

1.2 Technical Concepts

1.2.1 Django

According to djangoproject.com, Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. It follows the

model-template-views architectural pattern. The web app is built using this framework.

1.2.2 HTML

HTML or Hypertext Markup Language decides what the content of a web page shouldhave. With the help of HTML, we can insert different things like images, charts, tables,texts etc. into a web page.

1.2.3 CSS

CSS or Cascading Style Sheets decides the visual design of a website. It is used to present and style a web page.

It also lets a user include design, layout and variations in display for different devices and screen sizes.

1.2.4 JavaScript

Using JavaScript, the content and style of the web page can interact with the user.It is used for the client side behaviour of a web page. The dedicated JavaScript engine executesthe code when a user enters the web page.

1.2.5 Beautiful Soup

Beautiful Soup is a package of Python. It scrapes information from different websites.

Beautiful Soup is used to parse the HTML content of the three news portals and identifythe elements with their tags and access them. The headlines of the selected news portals are extracted by Beautiful Soup.

1.3 Objective

The process of gathering information from different sources can be very time consumingand hectic if done manually. The development of a news scraper can fasten the processand save time.

The objectives of the project are:

- To build a news scraper
- To gather information in a more efficient way
- To save time of searching different news sources
- To present data in a more structured way

1.4 Motivation

In this era of Globalization, everyone wants to stay updated with the latest world news.As the world is digitizing, our way of getting the latest information has also been digitized. Most of us follow different news sites for national and international news. So as the demand is increasing, the number of different news portals is also increasing.

There are millions of news portals all over the Internet. When someone decides to go through the contemporary news, the easiest way for the reader is to pick the device andgo to a news portal. Nowadays even the researchers also depend on the news sites to keep an eye on the concurrent information and attain the data.

This is where the problem arises. It is impossible for a person to type every single

website address and access each one of them. It requires a lot of time, so this process of data collection is not cost-efficient. It slows down the pace and decreases efficiency. Scraping the news is a very useful solution to this problem. The use of web scraper cansave a lot of time and increase efficiency as one can easily access different news sites without having the trouble of typing every single address.

Chapter 2 - Background Study

2.1 An Intelligent survey of personalized

Information Retrievalusing web scraper [2]

In this paper the authors tell us about personal information retrieval using web scraper and discusses the background and other attributes associated

with personal information retrieval. In this paper the author explains the evolution of information retrieval from

and how web scrapers is the most advanced technique. According to the authors the data on net is every increasing and in the vast sea of data it takes a lot of time and effort to get the desired data and web scraper can solve this problem by scraping the desired data from the html web pages and represent in the desired format For example for person it is very difficult to search the topic he/she interested by using google because therewill a lot of links to go through but if that person uses a web scraper it becomes to get the information without any work and that's how the whole process of personalized information retrieval is done in a jiffy by the web scraper and the work done by the user is reduced significantly

In the paper authors also talk about the legality problems one can face using web scraper as well as how web scraper can be used for illegal goals

for example, an example of web scraping in the travel industry is the case

of Hipmunk. Hipmunk was used to scrape data, collect the price informationand other vital statistics. The information was obtained before it could be received by the online travel agents and suppliers with an actual partnership agreement. However, this activity was stopped immediately

when they were asked to.

The authors also talk about some of the popular web scraper tools in today's world one of them is FMiner. IT is very important software which extracts information from the weband displays information in a readable format for the users. The software's main

implementation idea is to obtain the information from the URL of the websites. The webscraping software is designed to scan through set sites, and scrape set data at your

requirement. The FMiner software extracts information and stores the data in various formats so that the users can access them easily. The formats used by FMiner are .csv, access, Excel or even an SQL server.

2.2 Socially Smart - an Aggregator System for social media usingWeb Scraping [3]

In this paper the authors tell us about a web application that they have created namedSocially Smart which aggregates all the latest social media posts from multiple social media web-based platforms such as famous reddit blog, Onion News and GitHub andsummarizes them to display in short and crisp words. It will fetch the top and trending posts without less priority posts.

Socially Smart fetches the individual posts from the social media websites and providesthem in a single flexible platform. It reduces the time being spent on social media and the efficiency of the useful info is also increased since the posts fetched the most commented and trending posts on respective websites. This Social Smart uses two different methods one is Web Scraping / Web Crawling and the other one is fetching thedata using websites api to fetch data, SQL database to store the data and Django framework to create a web app and for UI basic html templates

BeautifulSoup library to use the web scraper. According to the author The Social Smart work principle are as follows:

First it visits the social media websites specified and scrapes the data. Once done it willstore the data in the SQLite database deleting the previous data present in it to update the top and trending posts. For this the user needs to give the event triggering click on the home page. Once the data is stored in the database the logic will analyse the data and keeps only the high commented and trending posts by deleting the remaining. Nowthe posts in the database are finally displayed to user in simple UI/UX.

2.1 Exploiting web scraping in a collaborative filtering-based approach to web advertising [4]

In this paper, authors focus on techniques that extract the content of a Web page. Theyadopted scraping techniques in the Web advertising field. To this end, They propose a collaborative filtering-based Web advertising system aimed at finding the most relevantads for a generic Web page by exploiting Web scraping. Collaborative filtering consistsof automatically making predictions (filtering) about the interests of a user by collecting

preferences or tastes from similar users (collaboration); the underlying idea is that thosewho agreed in the past tend to agree again in the future. Collaborative filtering systems try to predict the utility of items for a particular user based on the items previously

rated by other users. The authors used collaborative filtering algorithms to automate recommendation of ads.

The authors in this paper exploit Web scraping to suggest suitable ads to a given Webpage. The authors propose a Web advertising system that relies on collaborative filtering and exploits scraping techniques to analyze the page content.

The proposed system has three modules

Inlink extractor :- The inlink extractor gives as output the list of the 10 extracted peer pages that will be scrapped by the Ad extractor in order to identify the most related ads

Ad Extractor :- This module is aimed at extracting banner ads from the peer pages.

To this end, we rely on Web scraping, i.e., a set of techniques used to automatically getsome information from a website instead of manually copying it. Scraping is performed by using Python libraries provided by HTMLParser and Beautiful Soup.

The ad extractor module gives as output an ordered set of the extracted banners, together with the corresponding url and their descriptions. This set will be analyzed bythe ad selector that selects the three banners to be inserted in the original web page.

Ad selector: This module is aimed at selecting suitable banner ads from the set extracted by the ad extractor. The ad extractor module takes as input all the extracted

in links and analyzes them to extract the information related to all the embedded ads. Inparticular, in this work the authors are interested in extracting banner ads.Thus, the module looks for HTML anchor tag and selects those that refer to an image. Finally, the extracted ads are collected in an ads repository and once selected by the adselector, they can be put in the original Web page in two ways: (a) as banners, by simply presenting the retrieved images, or (b) as textual ads.

Chapter 3 - Methodology

3.1 System Diagram

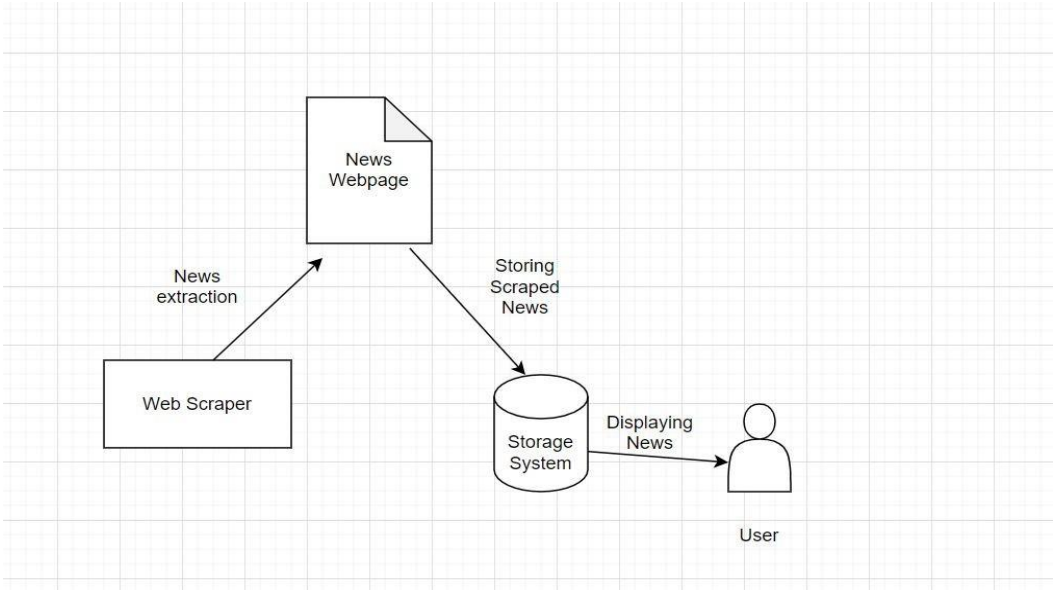

Figure 2: System diagram of the problem

3.2 Steps of solving the problem

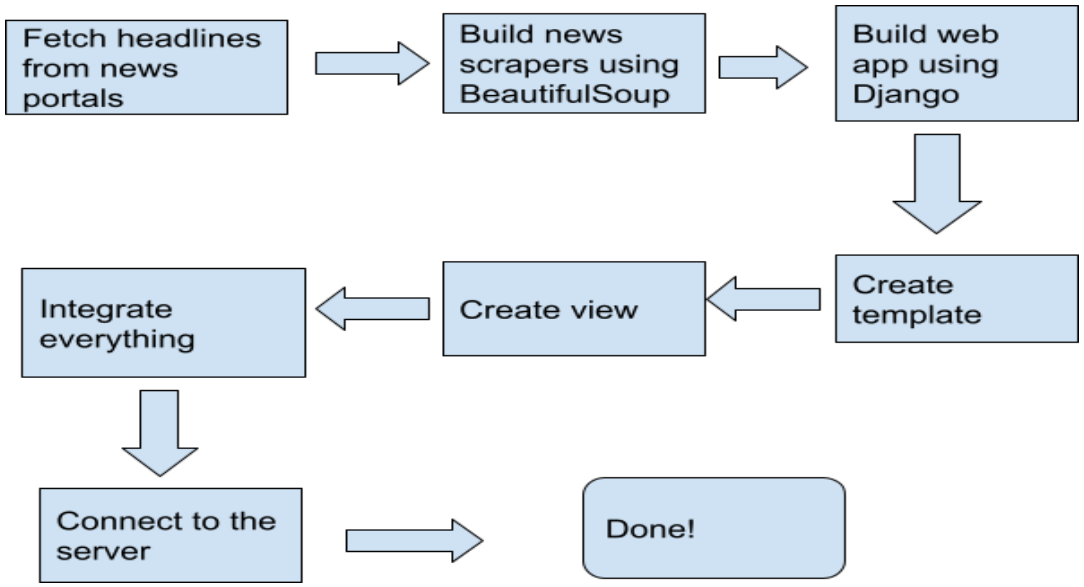

Figure 3: Steps of building the app

Chapter 4 - References

[1] "Web Scraping" , Wikipedia, Available: https://en.wikipedia.org/wiki/Web_scraping.

[2] Bhaskar Ghosh Dastidar , Devanjan Banerjee , Subhabrata Sengupta, "An Intelligent Survey of Personalized Information Retrieval using Web Scraper" , I.J. Education and Management Engineering, 2016, 5, 24-31, Available: http://www.mecs-press.net/ijeme

[3] Namburu Srikanth, Vennapusa Tejaswini, Chethala Harish, D. Praveen Kumar, "Socially Smart an Aggregation System for Social Media using Web Scraping" , International ResearchJournal of Engineering and Technology (IRJET) Volume: 06 Issue: 04 | Apr 2019, Available https://www.irjet.net/archives/V6/i4/IRJET-V6I4165.pdf

[4] Eloisa Vargiu, Mirko Urru, "Exploiting web scraping in a collaborative filtering based approach to web advertising", Artificial Intelligence Research, 2013, Vol. 2, No. 1 , Online Published: December 5, 2012 , Available: http://www.sciedu.ca/journal/index.php/air/article/view/1390

_____The End_____