“SMART BUY”

Mini Project Report

Submitted in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE and **ENGINEERING**

B.SIVA KUMAR	19L31A0518
P.VATSALYA	19L31A0512
V.KEERTHI	19L31A0526
PAVAN KALYAN.M	20L35A0502

Under the guidance of

Mrs.P.SANDHYA, M.Tech

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VIGNAN’S INSTITUTE OF INFORMATION TECHNOLOGY

(Autonomous) Affiliated to JNTU Kakinada & Approved by AICTE, New Delhi

Re-Accredited by NBA & NAAC (CGPA of 3.41/ 4.00)

ISO 9001:2008, ISO 14001:2004, OHSAS 18001:2007 Certified Institution

VISAKHAPATNAM – 530 039

2021-2022

VIGNAN'S INSTITUTE OF INFORMATION TECHNOLOGY
Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project report entitled "**SMART BUY**" is the Bonafide record of project work carried out under my supervision by **B.SIVAKUMAR (19L31A0518), P.VATSALYA (19L31A0512), V.KEERTHI (19L31A0526), and PAVAN KALYAN.M (20L35A0502)**, during the academic year 2021-2022, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University, Kakinada. The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree or Diploma.

Head of the Department

Mr.B.DINESH REDDY,MTech

HOD

Signature **of Project Guide**

Mrs.P.SANDHYA,MTech

Assistant Professor

DECLARATION

We hereby declare that the project report entitled "**SMART BUY**" has been written by us and has not been submitted either in part or whole for the award of any degree, diploma or any other similar title to this or any other university.

S.NO.	STUDENT NAME	REG.NO.
1	B.SIVA KUMAR	19L31A0518
2	P.VATSALYA	19L31A0512
3	V.KEERTHI	19L31A0526
4	PAVAN KALYAN .M	20L35A0502

Date:

Place:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to acknowledge the assistance and cooperation we have received from several persons while undertaking this B. Tech. Pre-Final Year Project. We owe a special debt of gratitude to **Mrs.P.SANDHYA,** Assistant Professor Department of Computer Science & Engineering, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us.

We also take the opportunity to acknowledge the contribution of Prof. **B.DINESH REDDY ,** Head, Department of Computer Science & Engineering, for his full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Table of Contents

SECTION-1

ABSTRACT

Every shopper looks for the best deals & discounts before buying any product. Nowadays before purchasing anything the buyers do some online research of the products on the internet. One of the major factors which lead to purchasing of any product is cost or pricing. The buyers tend to compare prices before purchasing any product. But since it is very difficult to visit each & every website for price
In comparison, there needs to be a solution to automate this process. The Price comparison website project proposed here gathers information on product prices from various websites & presents it to the users. The users can then choose to buy from the best options available. Even Ecommerce traders can use this price comparison website to study their competitors and form new strategies accordingly to attract new customers & stay ahead of their competitors.

SECTION-2

System Requirements

Hardware Requirement

Laptop or PC

- I3 processor system or higher
- 4 GB RAM or higher
- 10 GB ROM or higher

Software Requirement

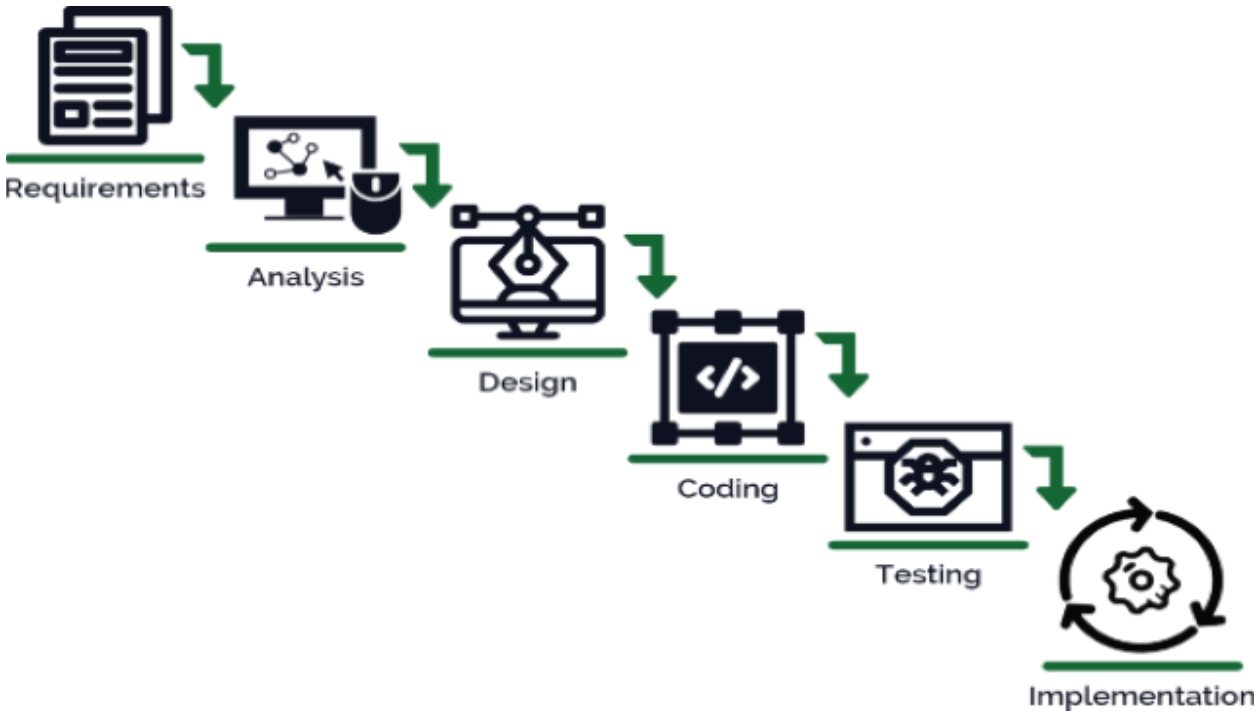
Laptop or PC

- Windows 7 or higher
- XAMPP or WAMP Server
- HTML5 PARSER
- Text Editor (Notepad++ / Sublime Text/VSCODE)
- Active Internet Connection For Fetching Data.
- Any Browser(Google Chrome,Firefox,etc)

SECTION-3

PROJECT LIFECYCLE

The waterfall model is a classical model used in the system development life cycle to create a system with a linear and sequential approach. It is termed a waterfall because the model develops systematically from one phase to another in a downward fashion. The waterfall approach does not define the process to go back to the previous phase to handle changes in requirements. The waterfall approach is the earliest approach that was used for software development.



SECTION-4

WORKING METHODOLOGY

What is web scraping

Web scraping is the process of using bots to extract content and data from a website.

Unlike screen scraping, which only copies pixels displayed on screen, web scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.

Web scraping is used in a variety of digital businesses that rely on data harvesting. Legitimate use cases include:

- Search engine bots crawl a site, analyzing its content and then ranking it.
- Price comparison sites deploying bots to auto-fetch prices and product descriptions for allied seller websites.
- Market research companies using scrapers to pull data from forums and social media (e.g., for sentiment analysis).

Web scraping is also used for illegal purposes, including the undercutting of prices and the theft of copyrighted content. An online entity targeted by a scraper can suffer severe financial losses, especially if it’s a business strongly relying on competitive pricing models or deals in content distribution.

How do Web Scrapers Work?

So, how do web scrapers work? Automated web scrapers work in a rather simple but also complex way. After all, websites are built for humans to understand, not machines.

First, the web scraper will be given one or more URLs to load before scraping. The scraper then loads the entire HTML code for the page in question. More advanced scrapers will render the entire website, including CSS and Javascript elements.

Then the scraper will either extract all the data on the page or specific data selected by the user before the project is run.

Ideally, the user will go through the process of selecting the specific data they want from the page. For example, you might want to scrape an Amazon product page for prices and models but are not necessarily interested in product reviews.

Lastly, the web scraper will output all the data that has been collected into a format that is more useful to the user.

Most web scrapers will output data to a CSV or Excel spreadsheet, while more advanced scrapers will support other formats such as JSON which can be used for an API.

Price scraping

In price scraping, a perpetrator typically uses a botnet from which to launch scraper bots to inspect competing business databases. The goal is to access pricing information, undercut rivals and boost sales.

Attacks frequently occur in industries where products are easily comparable and price plays a major role in purchasing decisions. Victims of price scraping can include travel agencies, ticket sellers and online electronics vendors.

For example, smartphone e-traders, who sell similar products for relatively consistent prices, are frequent targets. To remain competitive, they're motivated to offer the best prices possible, since customers usually go for the lowest cost offering. To gain an edge, a vendor can use a bot to continuously scrape his competitors' websites and instantly update his own prices accordingly.

For perpetrators, a successful price scraping can result in their offers being prominently featured on comparison websites—used by customers for both research and purchasing. Meanwhile, scraped sites often experience customer and revenue losses.

SECTION-5

WORKING PRINCIPLE

Generally, web scraping involves three steps:

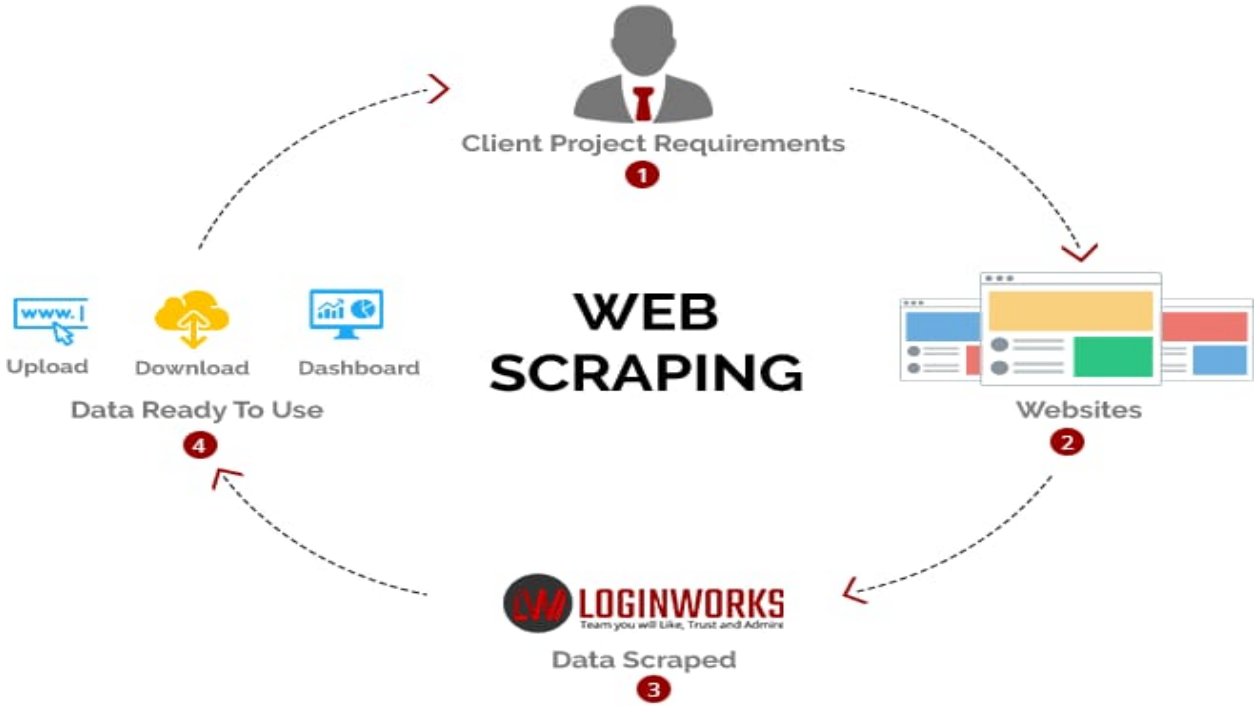
- First, we send a GET request to the server and we will receive a response in the form of web content.
- Next, we parse the HTML code of a web site following a tree structure path.
- Finally, we use the python library to search for the parse tree.

I know what you think -- web scraping looks good on paper but actually more complex in practice. We need coding to get the data we want, which makes it the privilege of someone who's a master of programming. As an alternative, there are web scraping tools automating web data extraction at fingertips.

A web scraping tool will load the URLs given by the users and render the entire website. As a result, you can extract any web data with simple point-and-click and file in a feasible format into your computer without coding.

For example, you might want to extract posts and comments from Twitter. All you have to do is to paste the URL to the scraper, select desired posts and comments

and execute. Therefore, it saves time and effort from the mundane work of copy-and-paste.



REQUEST PARSING USING URL

What is a URL?

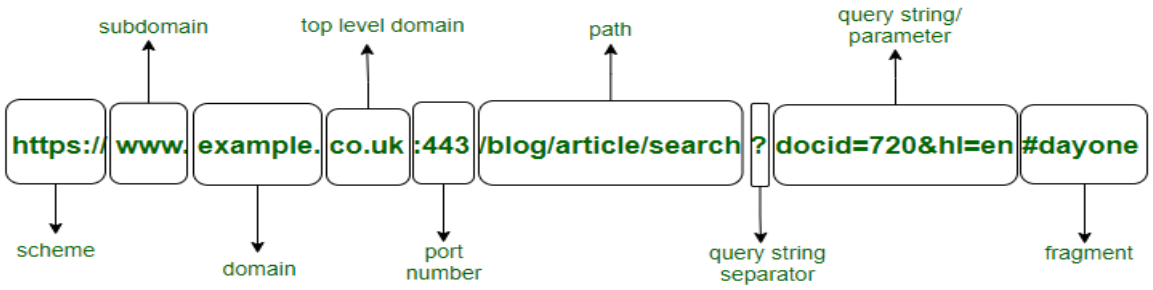
A Uniform Resource Locator or URL,is a short string containing an address which refers to an object in the web.

Parts of a URL

- **The protocol or scheme.** Used to access a resource on the internet. Protocols include http, https, ftps, mailto and file. The resource is reached through the domain name system (DNS) name. In this example, the protocol is https.
- **Host name or domain name.** The unique reference represents a webpage. For this example, whatis.techtarget.com.
- **Port name.** Usually not visible in URLs, but necessary. Always following a colon, port 80 is the default port for web servers, but there are other options. For example :port80.
- **Path.** A path refers to a file or location on the web server. For this example, search/query.
- **Query.** Found in the URL of dynamic pages. The query consists of a question mark, followed by parameters. For this example, ?.
- **Parameters.** Pieces of information in a query string of a URL. Multiple parameters can be separated by ampersands (&). For this example, q=URL.
- **Fragment.** This is an internal page reference, which refers to a section within the webpage. It appears at the end of a URL and begins with a hashtag (#). Although not in the example above, an example could be #history in the URL.

Parts of a URL

URL : https://www.example.co.uk:443/blog/article/search?docid=720&hl=en#dayone




```php
$searchtext="";


if ($_SERVER["REQUEST_METHOD"] == "POST"){
$searchtext = $_POST["searchtext"];

}
$searchtext = str_replace(' ', '%20', $searchtext);
$flp_str1="https://www.flipkart.com/search?q=";
$flp_str2="&otracker=search&otracker1=search&marketplace=FLIPKART&as-show=on&as=off";
$flp_query=$flp_str1.$searchtext.$flp_str2;

$html = file_get_html($flp_query);


#1
if(isset($html->find('div[class="_4rR01T"]',0)->plaintext)){

echo    $html->find('div[class="_4rR01T"]',0)->plaintext;
}
echo "<span class='float-right'>";

if(isset($html->find('div[class="_3tbKJL"]',0)->plaintext)){

echo    $html->find('div[class="_3tbKJL"]',0)->plaintext;
}

echo "</span><br><br>";
```

SECTION-6

FEATURES

- Search Engine to search products for over websites.

  (we included amazon and flipkart in our project)
- Listings where comparisons are shown.
- Latency from the point of clicking from search button to getting results is less than 5 seconds.
- Geolocation to know user's location and time to know when user is viewing
- Page views to track number of viewers
- Visit to site button ,provides direct redirection towards the clicked website.
- Supports multi-word queries in Search bar.
- Built on PHP Simple DOM web scraping platform.
- Speed.
- Data extraction at Scale.
- Cost-effective.
- Flexibility and Systematic approach.
- Performance Reliability and Robustness.
- Low maintenance cost.
- Automatic delivery of structured data.
- Web scraping has a learning curve.
- Affiliate Marketing.
- Revenue Optimization.

SECTION-7

PROS/CONS

The advantages of using a web scraping

Easy Integration

The ease with which an API can be integrated into a developer's application is one of its most appealing features. Only a set of credentials and a basic understanding of the API documentation are required. After you've completed the first request, you can concentrate solely on the parts that interest you, which brings us to another major benefit of APIs.

Built-in Solutions

The tool's built-in solutions are the most noticeable advantage of web scraping

APIs. They help you overcome some difficult problems by offering you Javascript rendering, datacenter, and residential proxies, custom headers, CAPTCHA bypass, IP rotations, and geolocation features.

Customization

A web scraping API allows you to personalize it and use its capabilities to its full potential to achieve all of your scraping goals, from API calls and geotargeting to dedicated accounts and custom scrapers.

Costs

Choosing an API for web scraping is an advantage over outsourcing a web scraping project, which can be costly. APIs aren't the cheapest option, but they're still not the most expensive in terms of the benefits they provide to developers. Prices vary based on the number of API calls you'll make per month and the amount of bandwidth you'll need. However, the return on investment is what makes a web scraping API worthwhile.

Time Saved

When time is your most valuable resource, a web scraping API is exactly what you need. Because you won't have to worry about building it, downloading it, or installing it, the process will be very short. So, you just have to start scraping after you've completed the integration and setup steps.

The disadvantage of using a web scraping

Learning Curve

An API, like any other tool, has its drawbacks. Learning how to use it would be one of them. You can't just start using an API and expect it to function properly. An API's documentation might be a little too light, depending on its complexity. Learning how to use the API will take a long time if the documentation is lacking.

SECURITY

Another minor discomfort will be a security issue. Once a hacker has gained access to an API, all applications that use it are at risk.

1. INJECTION
2. BROKEN AUTHENTICATION
3. SENSITIVE DATA EXPOSURE
4. XML EXTERNAL ENTITIES
5. BROKEN ACCESS CONTROL
6. SECURITY MISCONFIGURATION
7. CROSS SITE SCRIPTING
8. INSECURE DESERILATION
9. USING COMPONENTS WITH VULNERABILITIES
10. INSUFFICIENT LOGGING AND MONITORING

SECTION-8

Bibliography

1. (IDC), I. D. C., 2017. Data Age 2025, Framingham, USA: IDC.

2. Adamuz, P. L., 2015. Development of a generic test-bed for web scraping, Barcelona: European Education and Training Accreditation Center.

3. Apress, 2009. Using Web Scraping to Create Semantic Relations. Scripting Intelligence, pp. 205-228.

4. B.C., B., 2016. Scraping Data. In: Data Wrangling with R. Use R!.. Cham: Springer.

5. Berlind, D., 2015. APIs Are Like User Interfaces--Just With Different Users in Mind. [Online] Available at: https://www.programmableweb.com/news/api-economy-deliverslimitless-possibilities/analysis/2015/12/03 [Zugriff am 17 November 2017]. 6. Daly, M., 2016. Dublin Globe - Legal briefs:

6 Reasons Why Not to Scrape Data. [Online] Available at: http://www.dublinglobe.com/community/toolbox/legal-briefs-6-reasonsnot-scrape-data [Zugriff am 30 November 2017].

7. Daniel Glez-Pen, M. R.-J. a. F. F.-R., kein Datum Webscrapingtechnologies in an API world. s.l., s.n.

8. Emilio Ferraraa, P. D. M. G. F. R. B., 2014. Web data extraction, applications and techniques: A survey. Knowledge-Based Systems, Band 70, pp. 301-323.

SECTION-9

TECHNOLOGIES USED

- **Front end**:

HTML5 is the latest and enhanced version of HTML. HTML5 is a standard for structuring and presenting content on the internet.

CSS is used to control the style of a web document in a simple,easy way. CSS3 is a collaboration of CSS2 with its new specifications.

- **Framework**:

Bootstrap 4 is the most popular front end framework in the recent time. It is a sleek, intuitive, and powerful front-end framework for faster and easier web development.

PHP is an acronym for Hypertext Preprocessor (PHP) is a programming language that allows web developers to create dynamic content that interacts with databases.

JavaScript is a lightweight, interpreted programming language with object-oriented capabilities that allows to build interactivity into static HTML pages.

- **Tools**:

Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It is used to develop computer programs, as well as websites, web apps, web services and mobile apps.

XAMPP is a free and open-source cross-platform web server solution stack package developed by Apache Friends, consisting mainly of the Apache HTTP Server, MariaDB database, and interpreters for scripts written in the PHP and Perl programming languages.

SECTION-10

CONCLUSION

In summary, there is no specific piece of legislation which forbids Web Scraping to gather information. The website owners, however, may have legal rights against the company under intellectual property law and contract law. Each case will turn on its own facts though and this is very much dependent upon what information is scraped from the websites. Companies should beware of contractual provisions which they have agreed to in respect of a website's terms of use these may prohibit the user from taking and using the data off the site. If the data being scraped includes personal data, then compliance with data protection law must also be borne in mind. The only way to be truly certain that the rights of a website owner have not been infringed is to obtain their express consent to the screen scraping and subsequent use of the information. (Rezai, 2017) The overall outcome of forbidden Web Scraping could be for the Website owner. The possibility of less visitors, less links from content aggregator websites and less income from advertising. Thus, Data hosts should only use legal actions against scrapers when the scraper presents a threat to the data host's core business and the data host has a strong enough claim to prevail legally against the scraper. From the law perspective it is necessary to adjust the terms of use on the websites. Restriction of Web Scraping techniques can be directly included. Such a step does not require much resources and allows a direct argumentation at the court.