VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELGAUM, KARNATAKA

MINOR-PROJECT-I REPORT

ON

"Web Scraping Tool Sets"

Submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING

IN

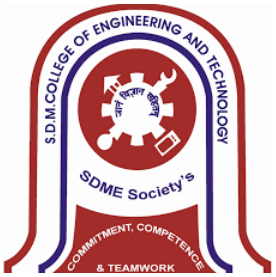
INFORMATION SCIENCE AND ENGINEERING

Submitted by

USN	NAME
2SD19IS003	Aditya Bammangoudar
2SD19IS021	Guruprasad Bhagwat
2SD19IS023	Kanishkvardhan A N
2SD19IS024	Karthik Kavathekar

Under the Guidance of Prof Leena Sakri

Project Coordinators: Dr. S. R. Biradar and Dr. Rajashekarappa



S.D.M COLLEGE OF ENGINEERING & TECHNOLOGY,

DHARWAD –580002

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the Minor-Project-1 work and presentation entitled **"Web Scraping Tool Sets"** *is a bonafide work carried out by* **Aditya Bammangoudar (2SD19IS003)**, **Guru Prasad Bhagwat (2SD19IS021), Kanishkvardhan A N (2SD19IS023), and Karthik Kavathekar (2SD19IS024),** *students of S.D.M.College of Engineering & Technology,Dharwad,in partial fulfilment for the award of* **Bachelor of Engineering in Information Science and Engineering of *Visvesvaraya Technological University, Belgaum***, *during the year 2021-2022. It is certified that all corrections/suggestions indicated for internal Assessment have been incorporated in the report deposited in the department library.The Minor project has been approved, as it satisfies the academic requirements in respect of project report prescribed for the said degree.*

----------	----------
Prof Leena Sakri	**Dr. J. D. Pujari**
Project Guide	HOD-ISE

ABSTRACT

Main objective of Web Scraping is to extract information from one or many websites and process it into simple structures such as spreadsheets, databases, or CSV files. However, in addition to being a very complicated task, Web Scraping is resource and time-consuming, mainly when it is carried out manually.

This workshop will introduce basic techniques for web scraping using popular open-source tools. The first part of the workshop will provide an overview of basic HTML elements and Python tools for developing a custom web scraper. The second part will enable participants to practice accessing websites, parsing information, and storing data in a CSV file.

The Internet grants a wide scope of facts and data sources established by humans. Though, it shall consist of an enormous assortment of dissimilar and ailing organised data, challenging in the collection in a physical means and problematic for its usage in mechanical processes. Since the recent past, procedures along-with various outfits have been developed to permit data gathering and alteration into organised information to be accomplished by B2C and B2B systems. This paper will focus on various aspects of web scraping, beginning with the basic introduction and a brief discussion on various software's and tools for web scraping.

Table of Contents

PROBLEM STATEMENT

When the user requests a website the website throws a huge amount of data and for the user, it is very hard to read to avoid that we can use the Web scraping tools to retrieve the data in a structured manner.

1.1 Why your project is important

Students and Teachers - For a faster understanding of concepts

Authors and Writers - For quick comprehension of their topics

Journalists - Attainable ideas for market demands/sentiments.

Stock Investors/Beginners - To get valuable ideas and information about the top 15 companies to understand the market

A web app which displays summarised data to-the-point information helps the user to comprehend on the spot without wasting any time.

1.2 Where is it used?

Some of the main use cases of web scraping include price monitoring, price intelligence, news monitoring, lead generation, and market research among many others.

1.3 What we will do

We will be developing a web application which will provide a user interface for the user to easily navigate through and provide the details like what he/she cooks, their food preferences, physical activity details, etc. based on which the web-application will tell the user where they should shift.

The web-application is developed using Python. Here, Streamlit is used for developing the front-end of the web-application and Python will be used to develop the Logic side, or the backend is written with the help of Python.

1.4 About the further report

The chapter mentioned below is the complete description of the project and activities performed to achieve the project goals. In chapter 2, we have mentioned the complete background about the project, what was the idea, what is the thought of the project, resources, and what was our approach. In

chapter 3, we have mentioned the technical aspects of the project by providing an overview of the aspects like what technology we have used.

CHAPTER 2: BACKGROUND

2.1 Technology Used

Different Technology used here to develop different parts of the application:

1. Front-end:

Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. Streamlit allows you to write an app the same way you write a python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app.

2. Back-end:

Python3 is object-oriented programming rich in different libraries used to write complex codes easily. Related to our project python is used to develop the backend and the Machine learning part of the application.

2.2 Perspective of the crowd

Different people had different perspective but overall, the view we got was:

1. This was a good idea as it will help many people to learn new things.
2. Some had doubts about the trust of the application in terms of correctness.
3. Appreciation from most of the people studying in schools and colleges.
4. People had doubts on the sensitivity of the application.

2.3 What we have referred

Survey is the way to reach people's expectations on accommodation of People in different cities when they travel. Customer segmentation is another way or process of dividing the people based on their income, health parameters, and basic requirements which helps people to make their life easy and healthy, even if they are not in their own home, town, country.

- Osmar Castrillo-Fernández, “Web Scraping: Applications and Tools”, European Public Sector Information Platform Topic Report No. 2015 10, December 2015.

- Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40:D109–14.
- Glez-Prenatal. , “Web scraping technologies in an API world”, Briefings in Bioinformatics Advance Access, doi:10.1093/bib/bbt026, published April 30, 2013.

2.4 What other people did on this topic

No particular solution which is this much user focused is yet developed. This is one of the initial moves made in developing such an application which takes care of so much of the condition provided by the user and provides them with the best possible option present in the market.

2.5 What and why out work is different

As there is no such application present in the market right now, it is totally a new thing and there is no work done earlier so it is a first move. Although there is no such software or application existing, there are some applications which apply some of the features like one of the projects has the features of finding the location of a restaurant according to the user preferences. The other application we looked on the internet was just a property allocating website which tells all the properties that are out there for renting.

Our project is totally different as:

1. It takes user preferences that a person needs as an amenity and surrounding needs.
2. The properties that are recommended are of the preferences that are provided by the user.

CHAPTER 3: PROJECT

3.1 Our approach for the problem

We will be using Python packages to retrieve information on articles based on users’ search URL. These articles will be limited to the number of search results in terms of data. The extracted and formatted article (derived from the main URL) is given back to the user in the form of the text output. The working web application will be hosted on the internet, using the streamlit library.

We came up with an idea for building something that helps the user by taking the user preferences and processing them accordingly for getting better results. Our overall approach for achieving our project goal is:

1. We took the data of the people which include their daily interests.
2. We visualised and studied the data and found the attributes necessary for developing the better model.
3. We have used web scraping in our backend for finding the best possible results for the user.
4. Python3 is used for developing the logical part of the web-application and it also serves the purpose for the core part of it. Python3 has a rich source of libraries and functions that made it easy to implement our project.
5. For the development of the frontend, Streamlite is used.

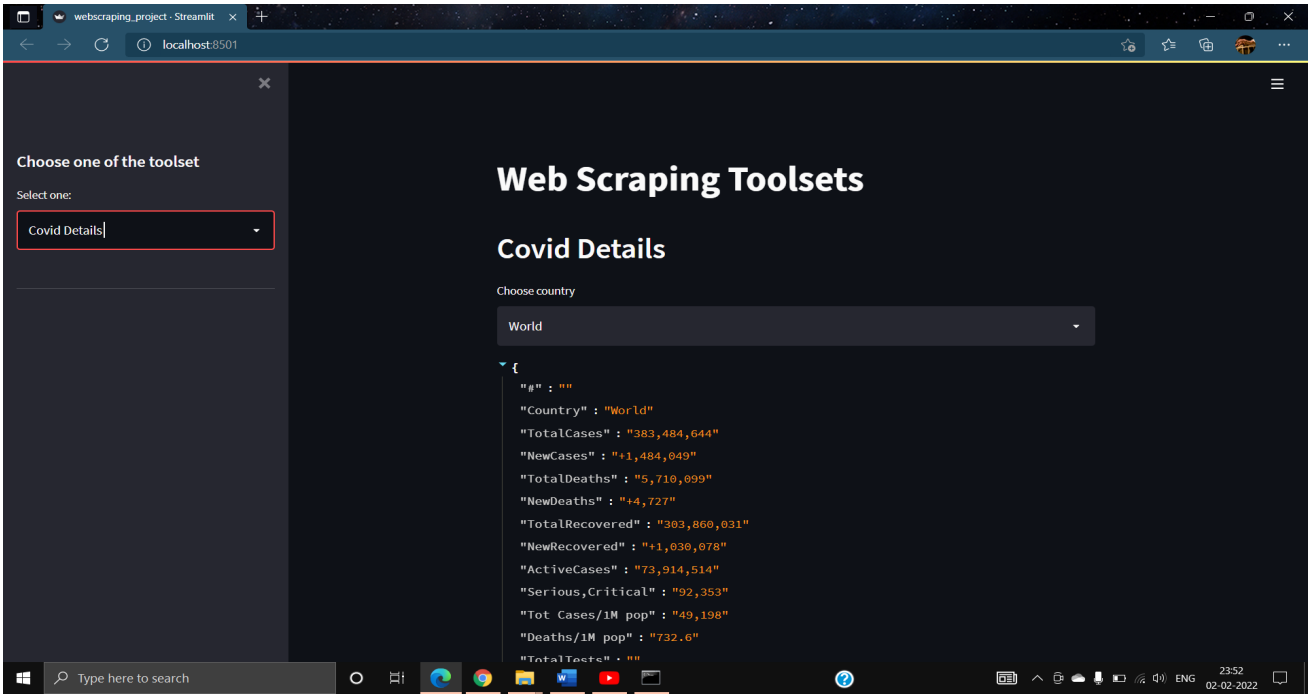
3.2 What we did

When the user requests a website the website throws a huge amount of data and for the user, it is very hard to read to avoid that we can use the Web scraping tools to retrieve the data in a structured manner.

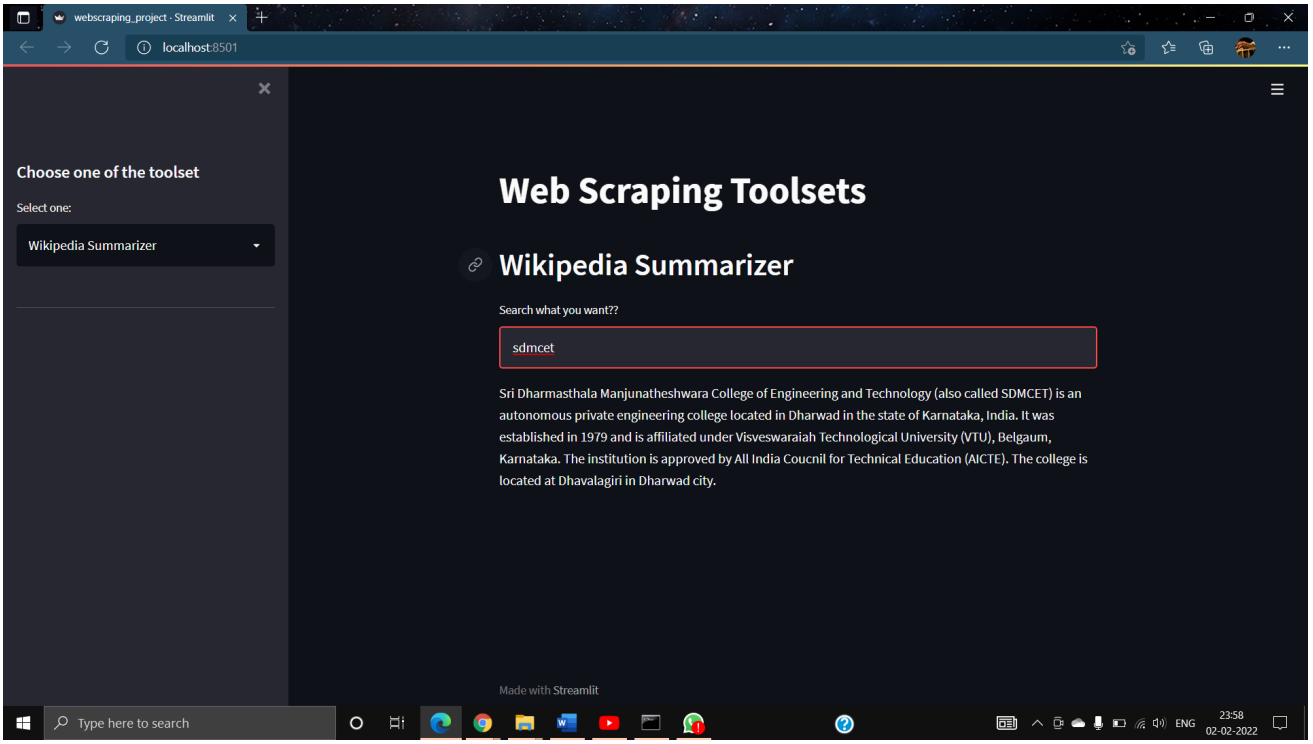
By using various modules and programming them using Python, we plan to display the abridged version of the data extracted from the internet so that the user can digest them instantaneously. The summarised to-the-point information helps the user to comprehend on the spot without wasting any time

3.3 Designs

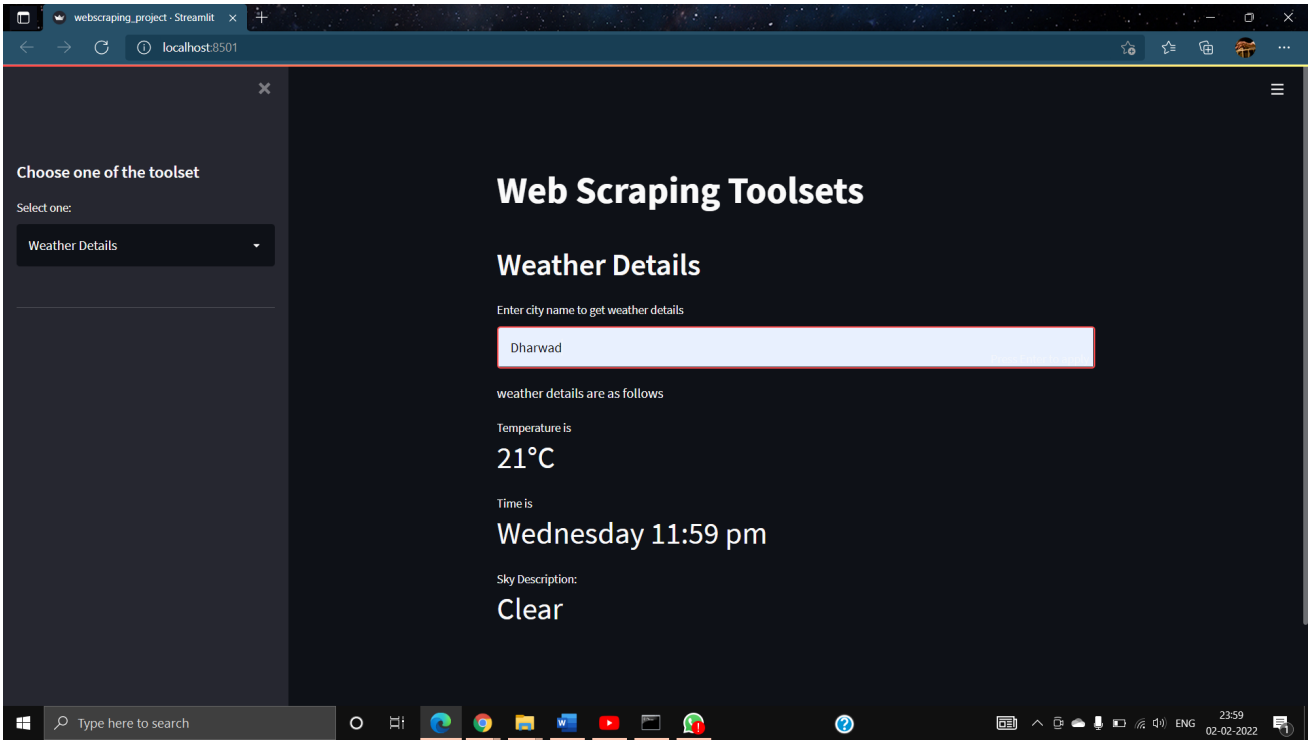
3.3.1 Covid Details



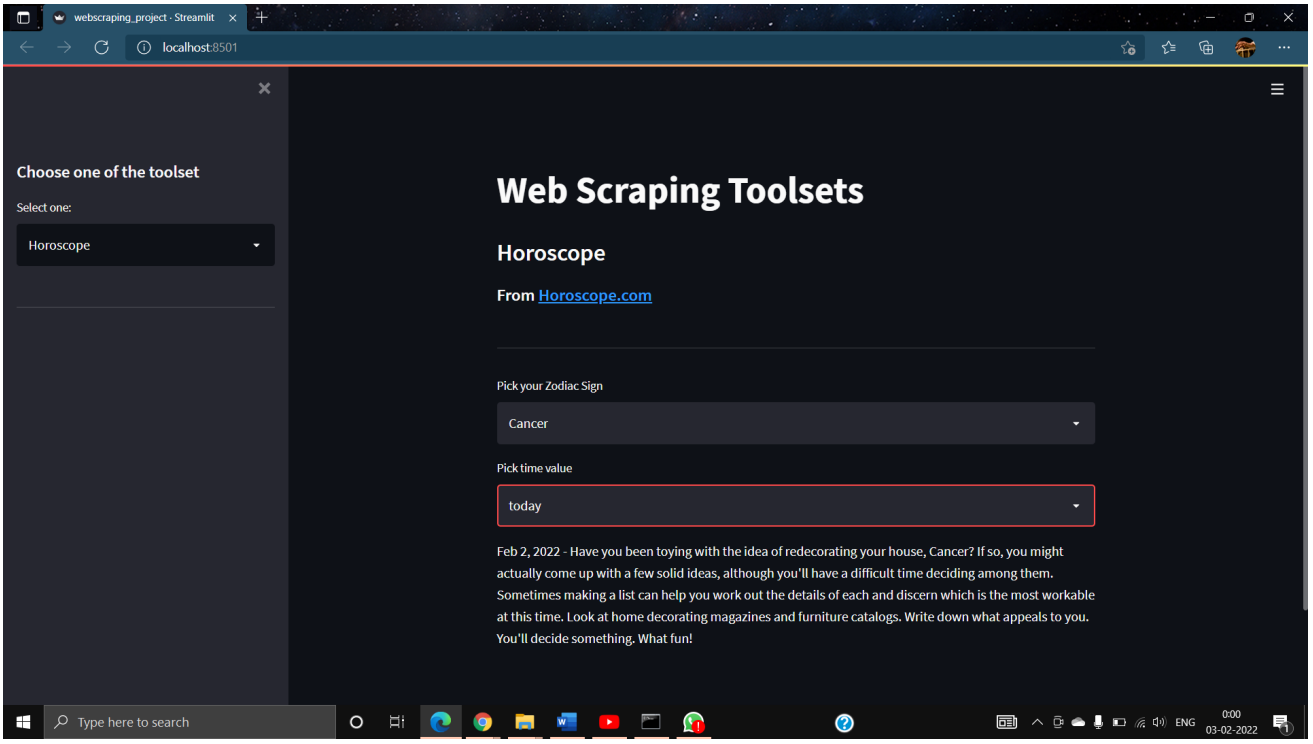
3.3.2 Wikipedia Summarizer



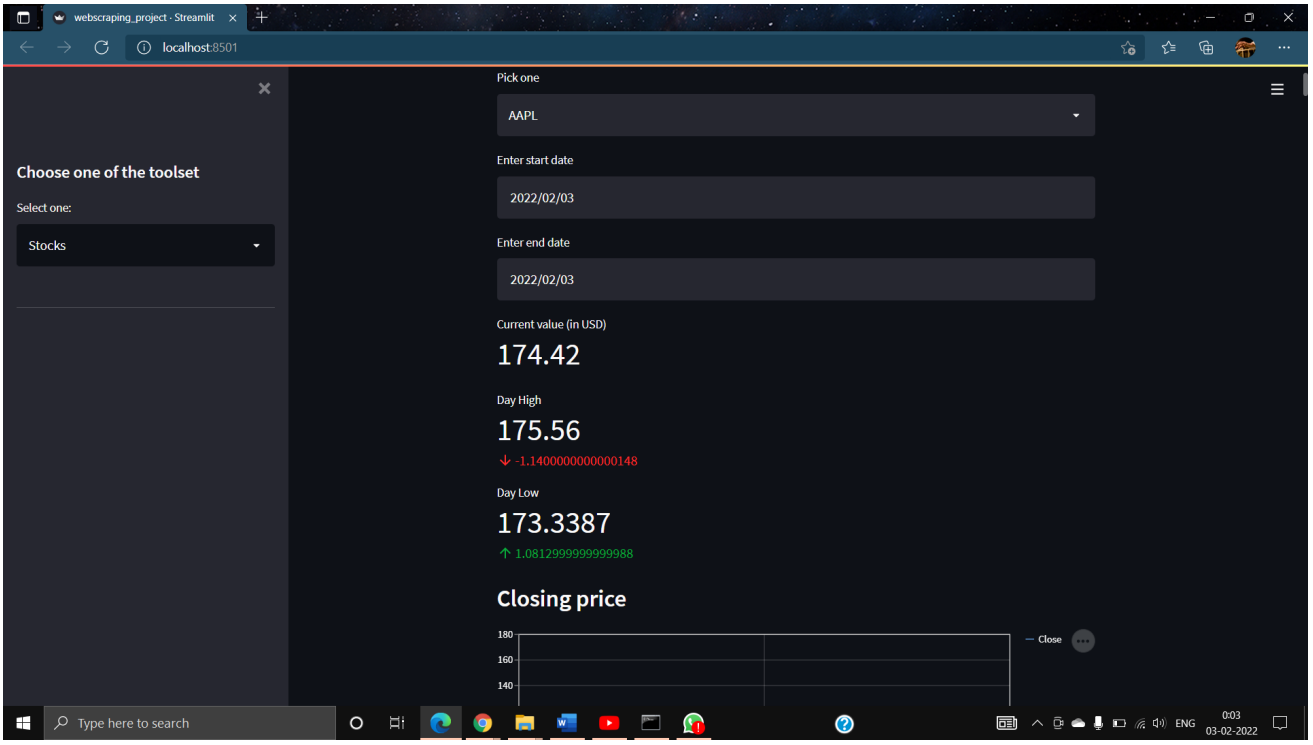
3.3.23 Weather Details



3.3.4 Horoscope

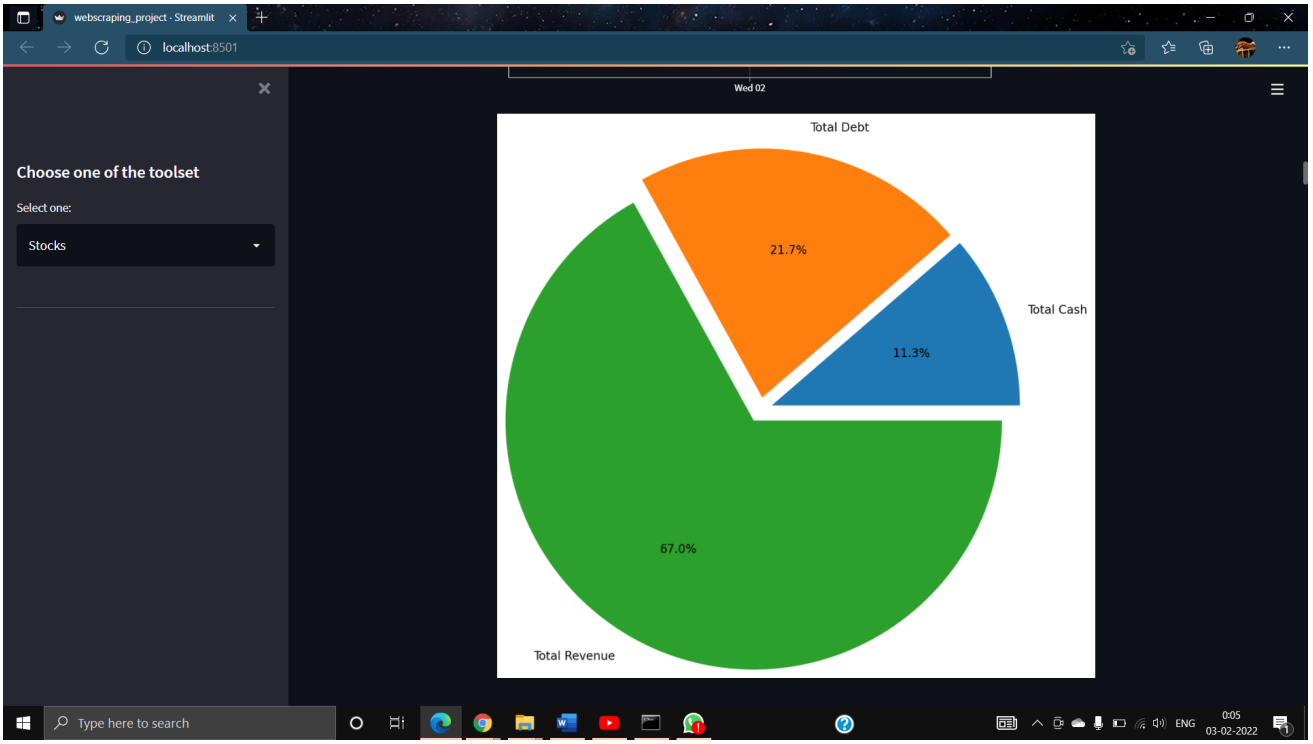


3.3.5 Stock Market



3.4 Graphs

3.3.6 Pie chart for Covid Details



3.5 Result

The above graphs represent the final set of data/information scraped by our scraper. The final result is based on the inputs provided by the user and their preferences.

3.6 Limitation

- The structure of websites changes frequently. Scraped data is arranged according to the structure of the website.
- It is not easy to handle complex websites.
- To extract data on a large scale is way harder.
- A web scraping tool is not omnipotent.

CHAPTER 4: RESULT/SUMMARY

The final result is the project according to the expectation. The project takes the user data about what they want to know about. A web app which displays summarised data to-the-point information helps the user to comprehend on the spot without wasting any time. Programming Language - Python. Python Libraries - Beautiful Soup, Requests, yFinance, Streamlit. This web app is open source and free for all the users.

CHAPTER 5: CONCLUSION

Web scraping is a recognizable phrase that has expanded significance owing to the requirement of “free” data accumulated in PDF documents or web pages. Numerous professionals and researchers require the data for processing, analysis, and extraction of significant consequences. Alternatively, people dealing with B2B use cases require the admittance of data from several sources for its integration into innovative applications which will offer supplementary values and novelty. Throughout this paper we have reviewed the various aspects of Web Scraper. Starting with the tools and software for web scraping, we have seen the operating principle, strengths, and drawbacks and finally viewed the applications of web scraping systems.

References

- Grimmer, Justin. 2013. Representational Style in Congress: What Legislators Say and Why It Matters. Cambridge University Press.
- William Marble, “Web Scraping With R”, stanford.edu, August 11, 2016 Carlos A. Iglesias Mercedes Garijo Jose Ignacio Fernandez-Villamor, Jacobo Blasco-Garcia. A Semantic Scraping Model for Web Resources, Applying Linked Data to Web Page Screen Scraping.
- Muntasir Mashuq MichelZiyan Zhou. Web Content Extraction Through Machine Learning.
- Diffbot: Extract content from standard page types: articles/blog posts, front pages, image, and product pages.
- Alex Gimson. This Just In: A Data Journalism Webinar with BeaSchofield.
- Daan Krijnen, “Automated Web Scraping APIs”.