Web Scraping using Python

Index

1. PRELIMINARY INVESTIGATION

1.1 **Introduction**

The World Wide Web is a network of interconnected websites that provide users with information. The way we exchange, acquire, and distribute data has radically altered because to the Internet. The amount of data available is always increasing.

As the volume, diversity, and value of data expands, business executives must concentrate their efforts on the data that matters most. For organisations and customers, not all data is equal. Enterprises who can identify and take use of the crucial fraction of data that will produce substantial positive impact for user experience, solving complex challenges, and establishing new economies of scale will thrive throughout this data transformation. To exploit the immense potential of data, business executives should concentrate on identifying and servicing that one-of-a-kind, crucial slice.

The quantity of material has also increased as a result of the use of the Internet as a new marketing and sales channel. Large amounts of data are provided by online retailers to characterise their items. Access to databases is provided by knowledgebase providers. According to IDC, the global data sphere will reach 163 zeta bytes by 2025. (that is a trillion gigabytes). That's 10 times more data than the 16.1ZB created in 2016. All of this data will open up new commercial options and unique user experiences.

(According to the International Data Corporation (IDC), 2017) It is no longer practicable to manually track and record all accessible sources due to the disorderly expansion. Web scraping took off at that point. In comparison to human data extraction, automated procedures allow for the capture of a large volume of data from the Web.

Another word that has become quite essential in conjunction with Web Scraping is Meta Data. Web scraping collects a large amount of data that may be analysed using Meta Data. This thesis will discuss the most prevalent causes for Web scraping as well as legal implications.

Web scraping techniques are covered in distinct chapters with relevant explanations. A list of currently available software tools is provided, along with a brief description of their capabilities. A practical example towards the conclusion of the text explains the complete process of web scraping.

1.2 Purpose of Web Scraping

The structure of a Hypertext Markup Language (HTML) page still holds a large quantity of source material on the World Wide Web. Because the intended reader was a person, automated extraction is challenging. This chapter explains why information extraction via web scraping is important and why it's done.

The rapid expansion of the World Wide Web has fundamentally altered how we exchange, acquire, and distribute data. Online, a vast quantity of data is kept in both organised and unstructured formats. This has created a new difficulty in terms of particular queries or study topics: conquering the tangled masses of internet data, rather than the worry of data scarcity and inaccessibility.

Market analysis and research

One of the market research strategies has been data collecting from internet sources. In comparison to traditional surveys, it provides a considerably faster answer.
While conventional surveys are preferred by Know it, web scraping is considered as a cost-effective supplement for such instruments. Multiple sources should be used to have a complete view of the market and to learn about the instruments available.

Consumers are engaged in the internet realm, sharing their frustrations, motivations, and experiences. Companies that want to learn more about their customers might use web resources. Web scraping is one of the methods for gathering this information.

Indexes are updated using targeted data collected from e-commerce and advertising servers. Which are based on pricing that vary regularly. Automated Web scraping indexes can provide more regular updating periods. With the rising relevance and availability of on-line pricing, it's reasonable to wonder if the consumer price index (CPI) projection, or related information, may be computed more often than monthly schedules now allow.

Social Network mining

On a variety of issues, social networks show trends and user reactions. Over the last decade, social media (such as blogs, online social networks, and micro blogs) has become one of the most important data sources for quantitative communication research. Researchers may extract important messages from social media networks for numerous study goals using basic programming techniques.

Scraping, leveraging site-provided applications, and crawling are all methods for obtaining social media data from publically available sources. The ability to get social media data makes social media data study appealing. Social media data is similar to traditional data in that it has the potential to be a treasure mine, but it takes data mining to unearth hidden gems.

1.3 Future Scope

In the future new features to the system can be added as the development will progress, features like storing data in a specific file format, downloading data from dynamic websites, better user interface, etc.

User may use this to scrape huge quantities of data and this software will be able to handle it and may be able to save the data in various file formats if the user desires so.

2. System Requirement and Specifications

2.1 Aim

The aim was to create a Website Scraping project that was quick, easy, and accurate, which I feel I accomplished.

The goal of this project is to provide "Website Scrapping," which requires various user inputs such as the website's url, the name of the Tags/Classes/ids to scrape, and the time in seconds to schedule scraping of the website, among other things, before accessing any type of information from the website, including text and images.

2.2 Objectives and Scopes

The primary goal of the Website Scraping system is to make scraping websites and storing data on a local computer as simple as possible.

This technology is primarily designed to give customers with a more dependable mechanism for collecting large amounts of data from the internet.

2.3 Functional Requirement

I examined the following functional needs of the project after a thorough conception and elaboration:
This project will require human participation as well as some processing requirements in order to work properly:

GUI Window

The windows displays all the entry boxes required for user to input the necessary url and tags and optional time values for the scraper to work properly.

Scrapping module

Required for processing user input, scraping websites, and downloading and storing data to a local computer.

2.4 Proposed Approach

The aim for this project is to make it easy for the user to collect large amounts of data from the desired website.

We purpose that the software allows and encourages the user to gather more information from the internet for cyber security purposes.

This is a simple software which allows user to select specific tags, class or id in order to scrape data from the website.

2.5 System requirements

Software Requirements:

Windows Xp, Windows 7(ultimate, enterprise)
Chrome or equivalent browser

Hardware Components:

Processor – i3
Hard Disk – 5 GB
Memory – 1GB RAM
1 Mbps internet Connection

3. Problem Statement & Problem Solutions

3.1 Problem Statement

Not being able to gather information fast enough has been a huge problem in cyber investigation. This delays the investigation and we risk the criminal running free.

This can lead to loss of trust in the judicial and police institutions hence, efficient way of data gathering and analyzing is the need of the hour.

3.2 Problem Solution

When it comes to collecting public information the python based web scraping can be used by the user to ensure fast and timely and easy collection and management of huge data.

Ideally this is done in a secure way from the comfort of our home or office with just a computer and an internet connection.

Also, user can specify the time in seconds in order to schedule the periodic scrapping of the desired website. This can help us gather data from dynamic websites which keep updating their information on a regular basis.

4. Feasibility Study

4.1. Feasibility Study

The viability of the project and the possibility that the system would be valuable to the organisation are examined during the preliminary study. The feasibility study's main goal is to evaluate the technical and financial feasibility of adding new modules and troubleshooting old operational systems. If given limitless resources and infinite time, all systems are possible. A fresh examination of the viability of using keystroke dynamics on touch screen devices was offered in this study.

4.2 Technical

The following are some of the technical issues that are frequently highlighted during the feasibility stage of an inquiry.
Is the technology required to carry out the suggestions?
Is the planned equipment technically capable of storing the data needed to operate the new system?
Will the proposed system, regardless of the number or location of users, offer appropriate response to inquiries?
Is it possible to upgrade the system if it is developed?
Is there any technological assurance of accuracy, dependability, accessibility, and data security?

4.3 Operational

User-Friendly

The customer can use the graphical window with forms provided with the project in order to insert data into the system. It has been taken care that it is quite intuitive and easy to understand for the user.

Reliability

The software is quite reliable when it comes to collecting large amounts of data from the desired website.

Portability

The application is quite portable and can used on any computer without the need for installation from the user. It is simple and has plug and play capabilities.

Availability

The application is always available for the user to use.

4.4 Economic

The computerised system should entirely replace the current system's data flow and operations, as well as create all of the manual system's reports plus a slew of extra management reports. It should be developed as a web application with its own web server and database server. Because the operations are dispersed throughout the business, the client need a centralised database.

5. Requirement Analysis

5.1 Project Specification

Project Title

Website Scrapping

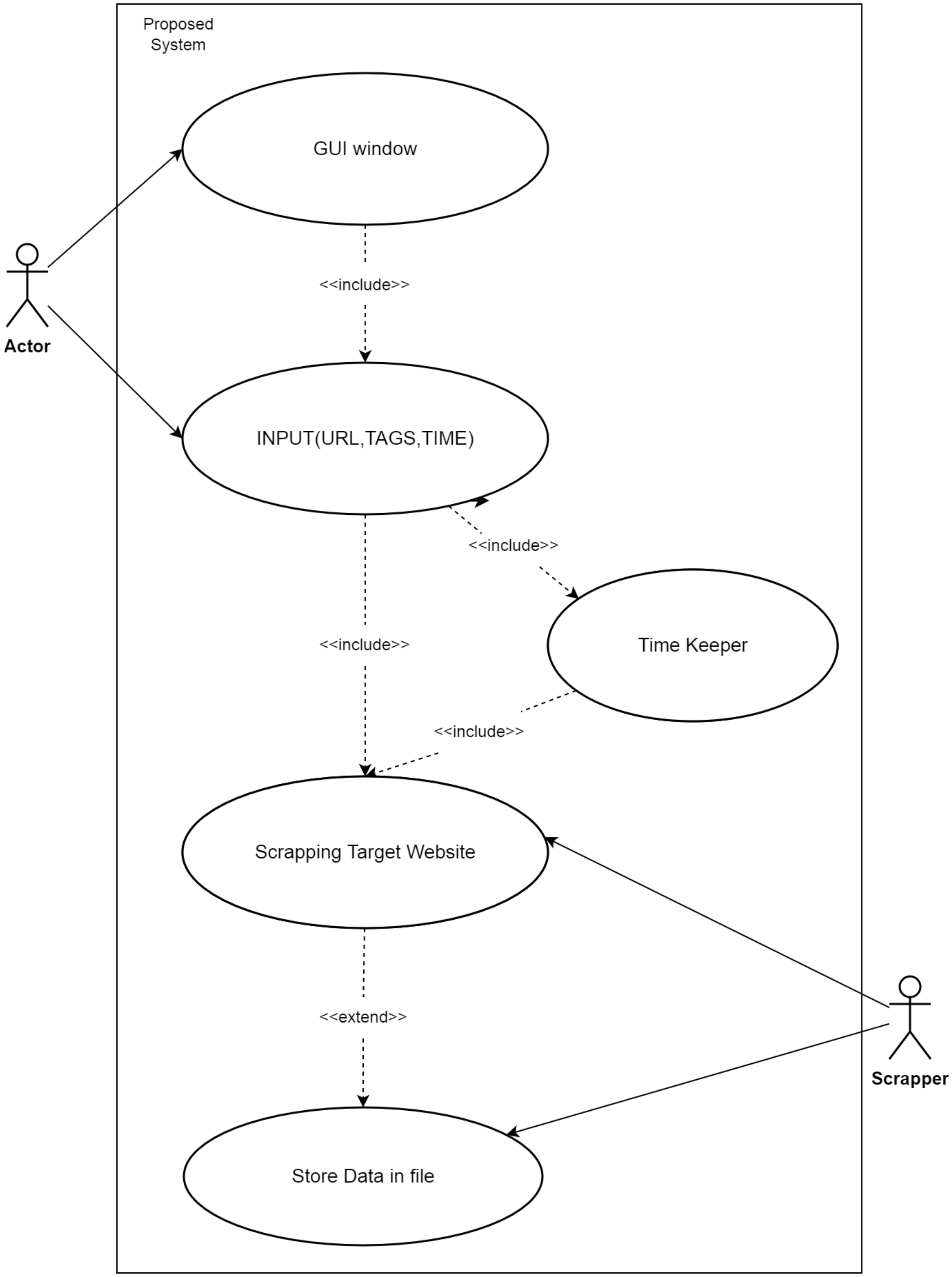
Aim of the project

The project involves data collection using Website Scraping software in order to collect and save data from a desired website.

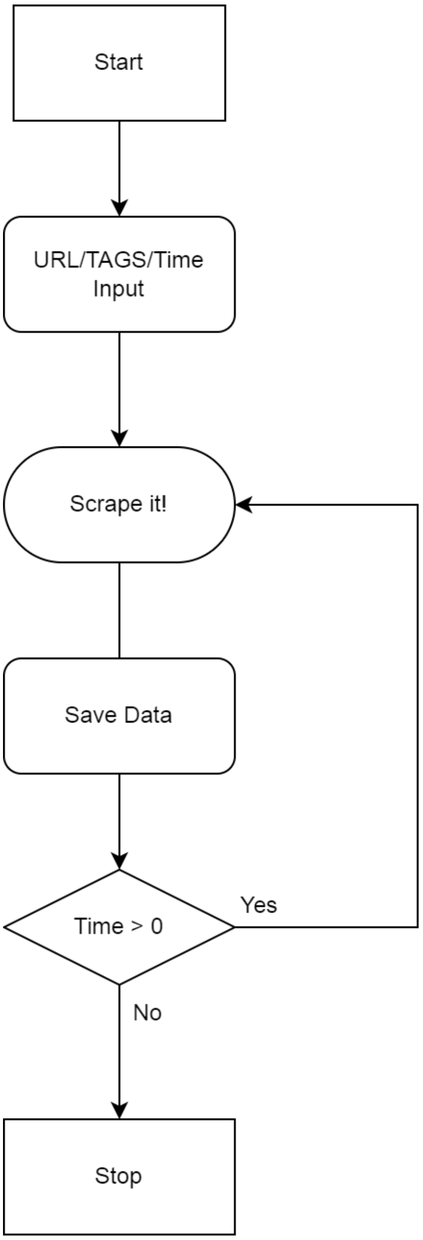
User Requirements

User needs a windows computer with an internet connection in order to run the software and scrape data from the desired website. User requires to provide website url and the tags or classes or ids in order to download and save data on the user computer.

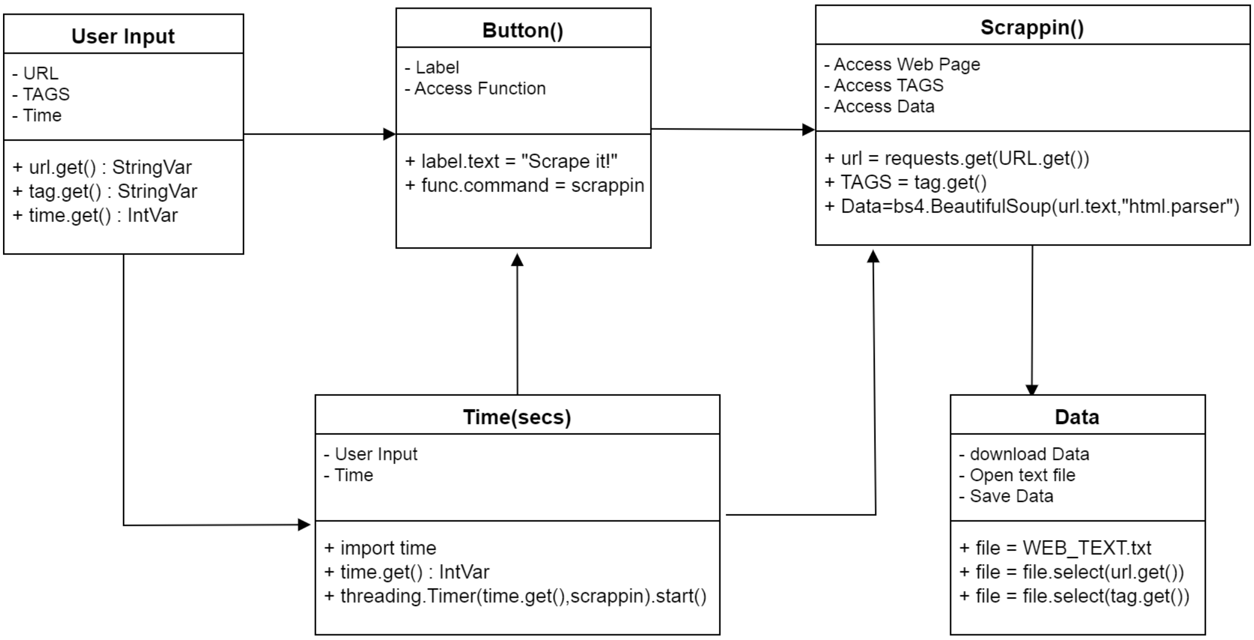
5.2 Use Case Diagram



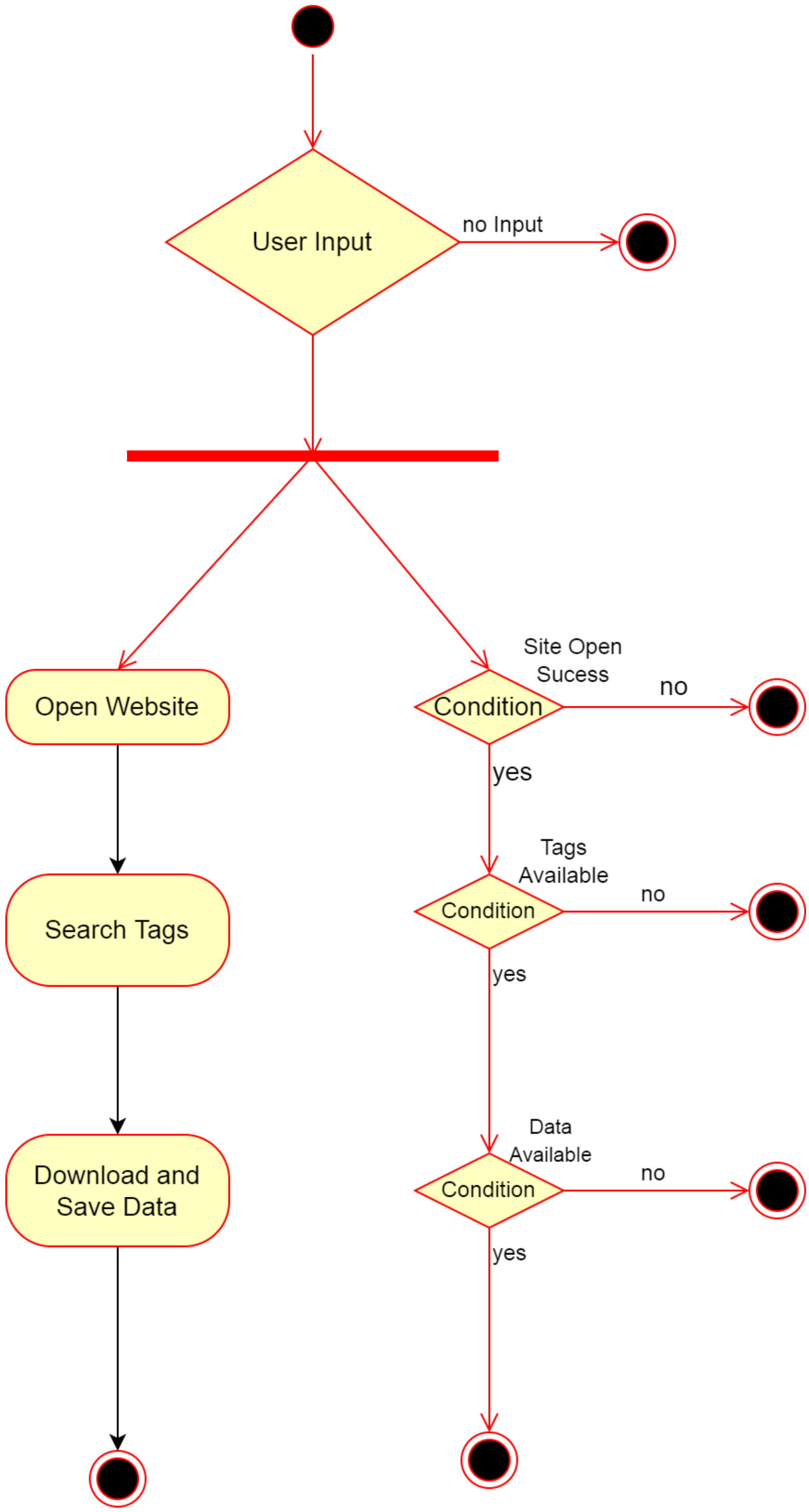
5.3 Flowchart Diagram



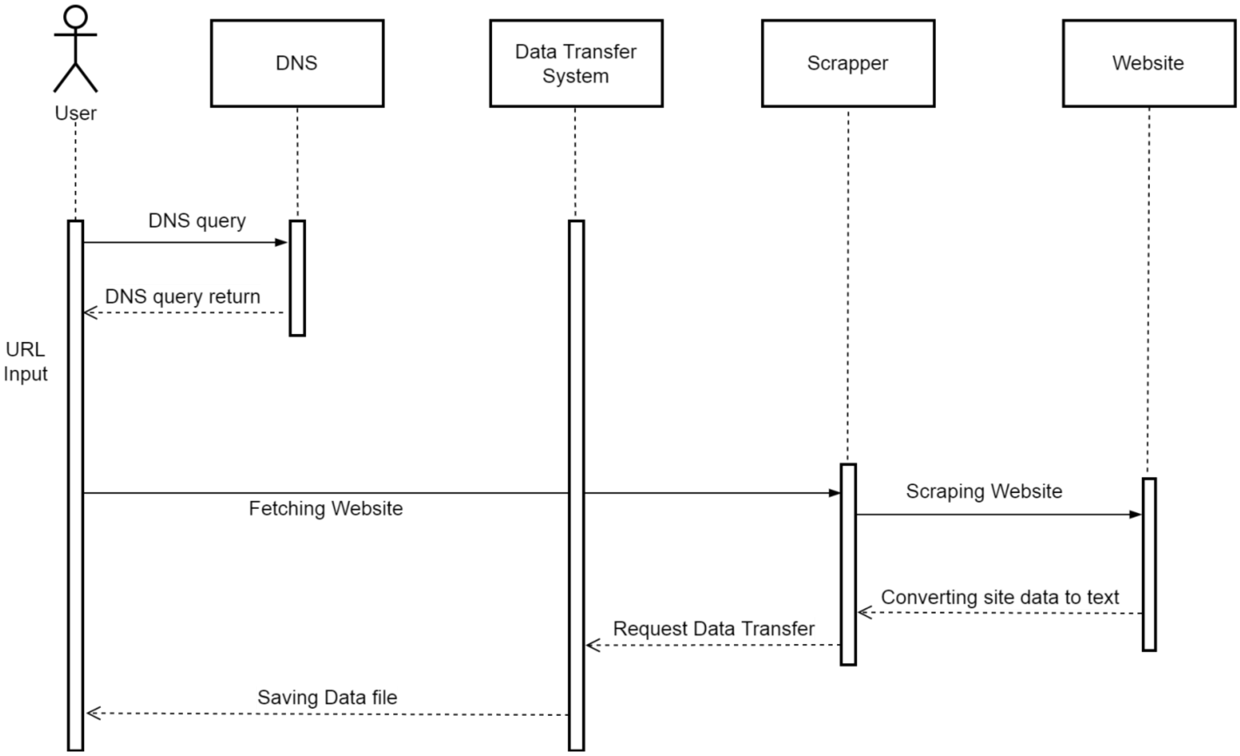
5.4 Class Diagram



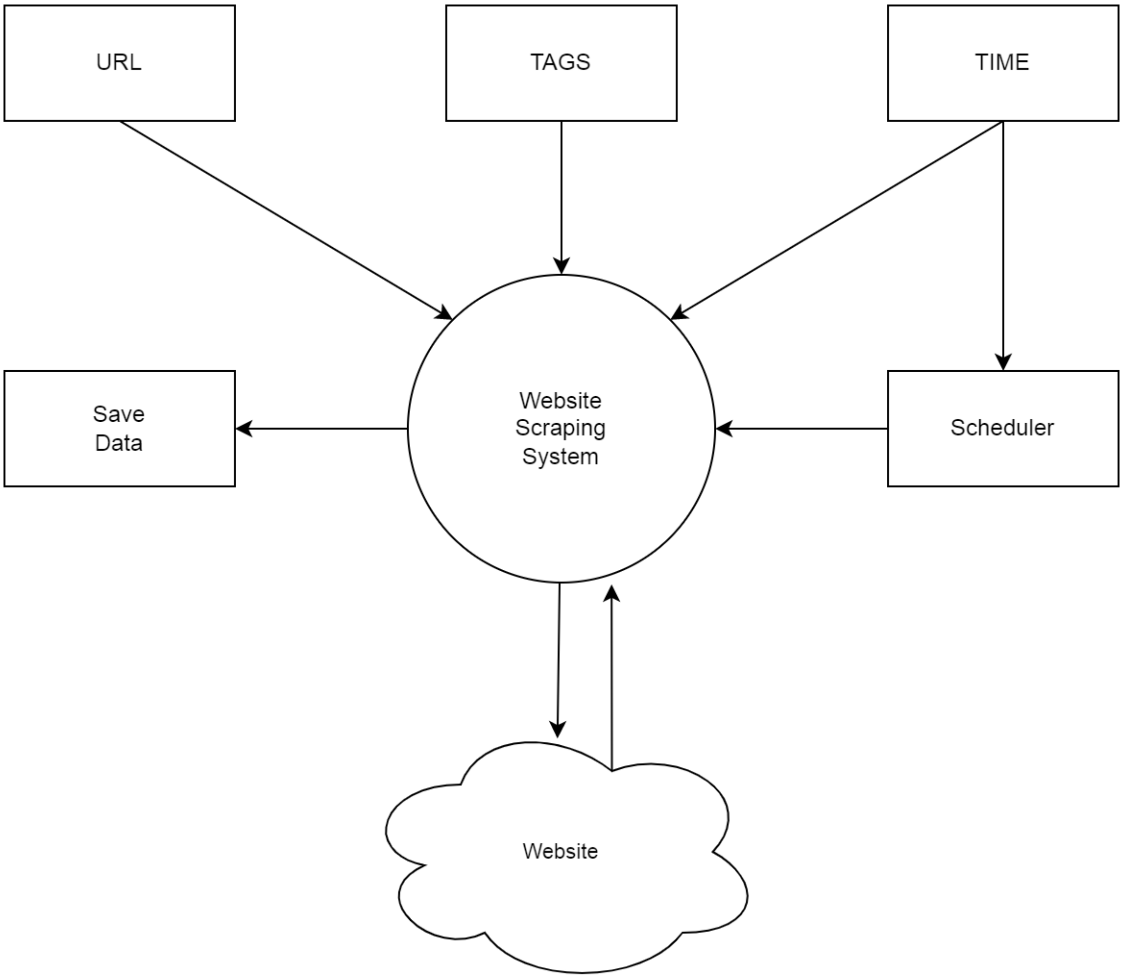
5.5 Activity Diagram



5.6 Sequence Diagram

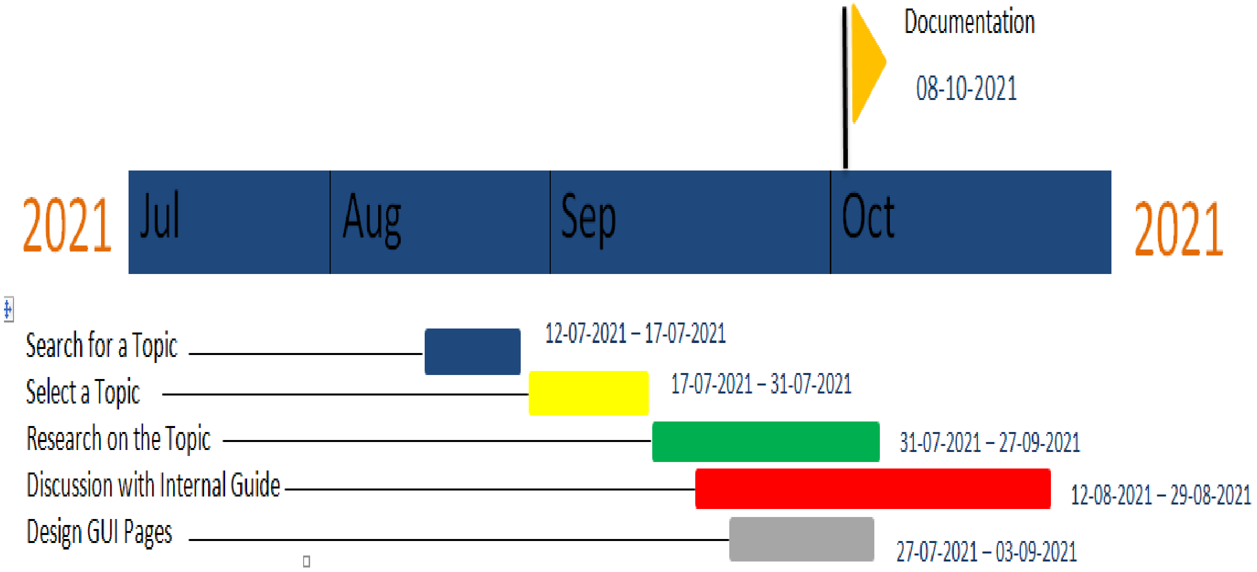


5.7 Data Flow Diagram

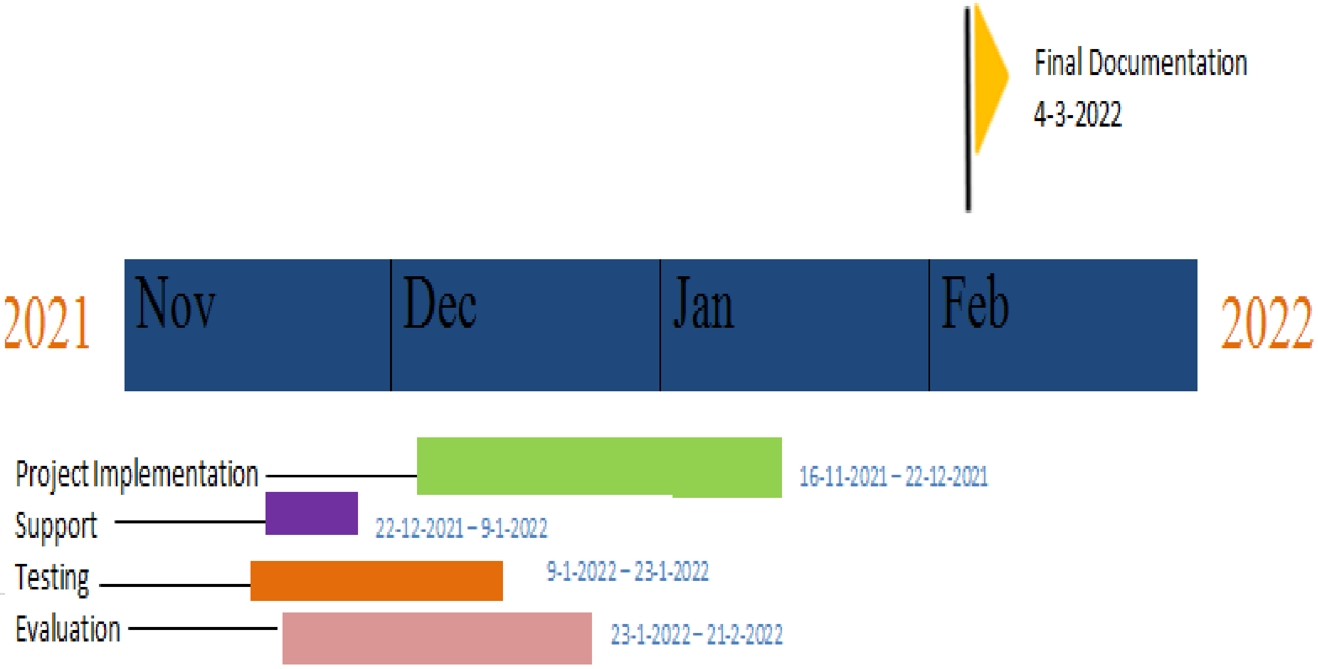


5.8 Project Timeline

Project Timeline for Semester 3



Project Timeline for Semester 4



6. Testing

6.1 Testing Principle

Testing is the process of running a software to see whether it has any bugs. It is the last step in the verification and validation process. During the testing step, we attempted to confirm the product's quality. We've also attempted to correct any problems from the previous phases.

Why is testing carried out?

Testing is the process of putting a system through its paces in order to uncover flaws.
Testing improves the integrity of a system by finding design flaws and system problems.
The goal of testing is to find places that are prone to errors. This aids in the prevention of system faults.
Testing adds value to the product by validating that it meets the user's needs.

6.2 Testing Objectives

Testing is the process of running a software in order to detect faults. A excellent test case is one that has a high chance of uncovering a flaw.

6.3 Strategic approach to Software Testing

A spiral can be drawn to represent the software engineering process. System engineering begins with the definition of software's purpose and leads to software requirement analysis, which establishes the information domain, functionalities, behaviour, performance, restrictions, and validation criteria for software. As we progress down the spiral, we reach design and then code. We spiral in along streamlines to construct computer software, lowering the degree of abstraction with each turn.

In the framework of the spiral, a software testing approach may likewise be examined. Unit testing starts at the top of the spiral and focuses on each software unit as it is written in source code. Testing progresses outward along the spiral to integration testing, where the focus is on software architecture design and construction. Validation testing is a step farther outward on the spiral, when requirements created as part of software requirements analysis are validated against the software that has been built. Finally, we get to system testing, which involves evaluating the software and other system components as a whole.

7. Coding and Output

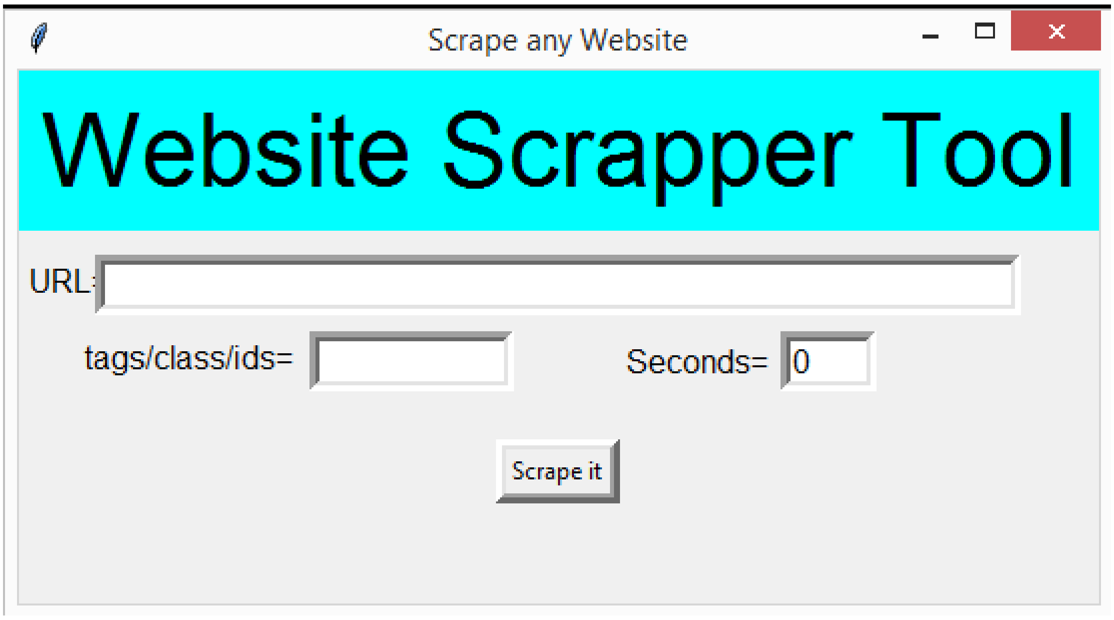
7.1 Coding Logic

```
import requests
import html5lib
import bs4
import sys
import time
import threading
from tkinter import *

####################################

win = Tk()
win.title("Scrape any Website")

def scrappin():

        url = requests.get(URL.get())
        res = bs4.BeautifulSoup(url.text , "html.parser")
        timestr = time.strftime("_%Y_%m_%d_%H_%M_%S")


        #######webcode only file
        try:
                saveFile2 = open("WEB_CODE"+timestr+".txt","a")

                for i in res.select(pvar.get()):
                        try:
                                saveFile2.write(i.prettify())

                        except:
                                error1 = open("LOG"+timestr+".txt","a")
                                error1.write("error1\n")
                                #print('error1')

        except:
                e1 = open("LOG"+timestr+".txt","a")
        e1.write("e1\n")
        #print('e1')

    finally:
        saveFile2.close()

    ######file to save web text
    try:
        saveFile1 = open("WEB_TEXT"+timestr+".txt","a")

        for i in res.select(pvar.get()):
            try:
                saveFile1.write(i.getText())
                saveFile1.write("\n\t")
            except:
                error2 = open("LOG"+timestr+".txt","a")
                error2.write("error2\n")
                #print("error2")

    except:
        e2 = open("LOG"+timestr+".txt","a")
        e2.write("e2\n")
        #print("e2")

    finally:
        saveFile1.close()

    #will execute def scrappin periodically
    if perio.get() > 0:
        threading.Timer(perio.get(),scrappin).start()


################################

#var variable is a String variable
var = StringVar()

#contains the text of label
var.set("Website Scrapper Tool")

#Label within the window....
LABEL_OF_WEB=Label(win,textvariable=var,bd=10,bg="cyan",font=("Calibre",38)).grid(row=0,column=0)

#StringVar to pass the url to the function
```

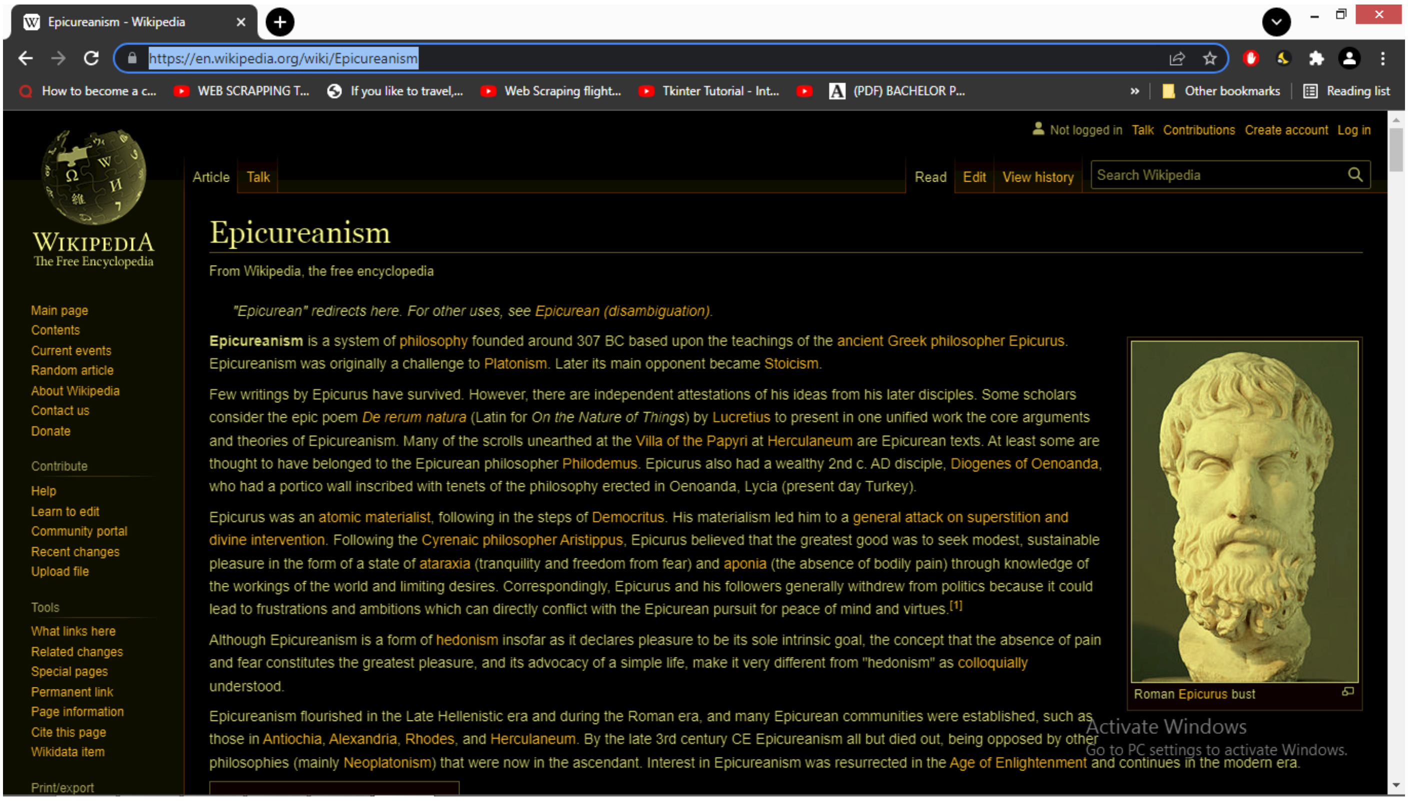
```
URL = StringVar()
pvar = StringVar()
perio = IntVar()

#entry box
url_label = Label(win,text="URL= ",font=7).place(x=2,y=93)
E1=Entry(win,bd=5,font=7,textvariable=URL,width=50).grid(row=1,column=0,padx=0,pady=12)

tags_label = Label(win,text="tags/class/ids= ",font=7).place(x=30,y=131)
E2=Entry(win,bd=5,font=7,textvariable=pvar,width=10).place(x=145,y=130)

time_label = Label(win,text="Seconds=",font=7).place(x=300,y=133)
E3=Entry(win,bd=5,font=7,textvariable=perio,width=4).place(x=380,y=130)

#Button
button=Button(win,text="Scrape it",bd=5,command=scrappin).grid(row=2,column=0,pady=50)

###########################################
win.mainloop()
```

7.2 Code Execution and Output

Output Image:



Website to be Scrapped:



Epicureanism

From Wikipedia, the free encyclopedia

"Epicurean" redirects here. For other uses, see Epicurean (disambiguation).

Epicureanism is a system of philosophy founded around 307 BC based upon the teachings of the ancient Greek philosopher Epicurus. Epicureanism was originally a challenge to Platonism. Later its main opponent became Stoicism.

Few writings by Epicurus have survived. However, there are independent attestations of his ideas from his later disciples. Some scholars consider the epic poem *De rerum natura* (Latin for *On the Nature of Things*) by Lucretius to present in one unified work the core arguments and theories of Epicureanism. Many of the scrolls unearthed at the Villa of the Papyri at Herculaneum are Epicurean texts. At least some are thought to have belonged to the Epicurean philosopher Philodemus. Epicurus also had a wealthy 2nd c. AD disciple, Diogenes of Oenoanda, who had a portico wall inscribed with tenets of the philosophy erected in Oenoanda, Lycia (present day Turkey).

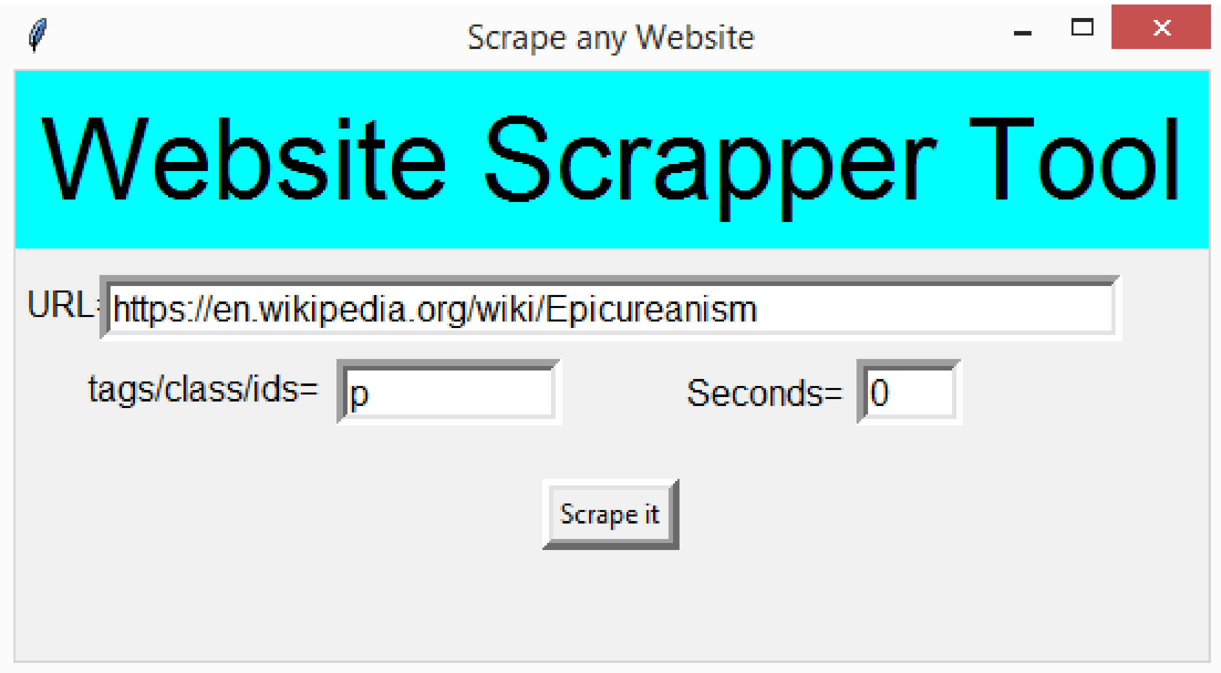
Epicurus was an atomic materialist, following in the steps of Democritus. His materialism led him to a general attack on superstition and divine intervention. Following the Cyrenaic philosopher Aristippus, Epicurus believed that the greatest good was to seek modest, sustainable pleasure in the form of a state of ataraxia (tranquility and freedom from fear) and aponia (the absence of bodily pain) through knowledge of the workings of the world and limiting desires. Correspondingly, Epicurus and his followers generally withdrew from politics because it could lead to frustrations and ambitions which can directly conflict with the Epicurean pursuit for peace of mind and virtues.[1]

Although Epicureanism is a form of hedonism insofar as it declares pleasure to be its sole intrinsic goal, the concept that the absence of pain and fear constitutes the greatest pleasure, and its advocacy of a simple life, make it very different from "hedonism" as colloquially understood.

Roman Epicurus bust

Epicureanism flourished in the Late Hellenistic era and during the Roman era, and many Epicurean communities were established, such as those in Antiochia, Alexandria, Rhodes, and Herculaneum. By the late 3rd century CE Epicureanism all but died out, being opposed by other philosophies (mainly Neoplatonism) that were now in the ascendant. Interest in Epicureanism was resurrected in the Age of Enlightenment and continues in the modern era.

Window after filling the details by user:



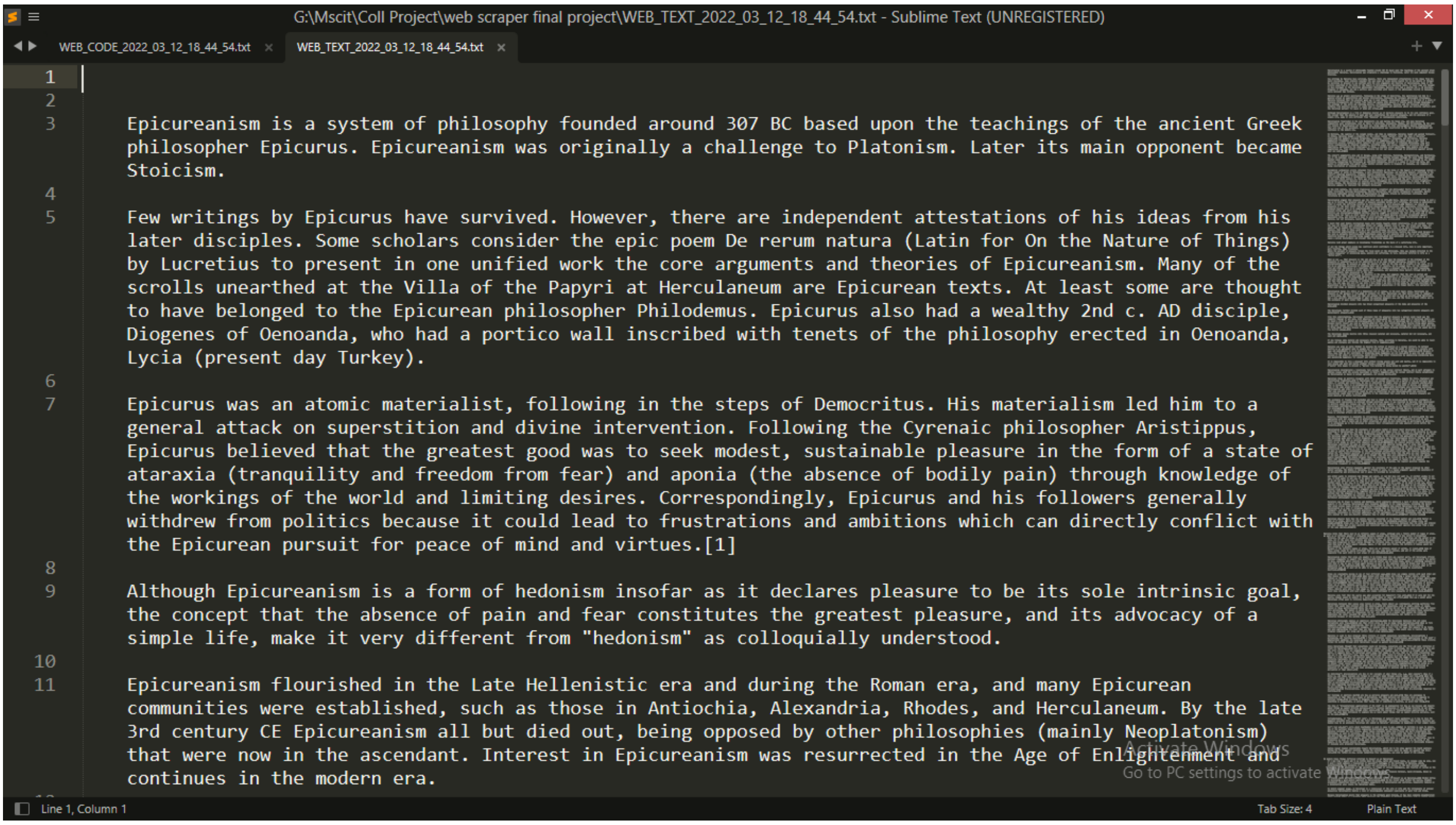
Website Scrapper Tool

URL: https://en.wikipedia.org/wiki/Epicureanism

tags/class/ids= p

Seconds= 0

Output Data after Scraping:


G:\Mscit\Coll Project\web scraper final project\WEB_TEXT_2022_03_12_18_44_54.txt - Sublime Text (UNREGISTERED)
WEB_CODE_2022_03_12_18_44_54.txt
WEB_TEXT_2022_03_12_18_44_54.txt

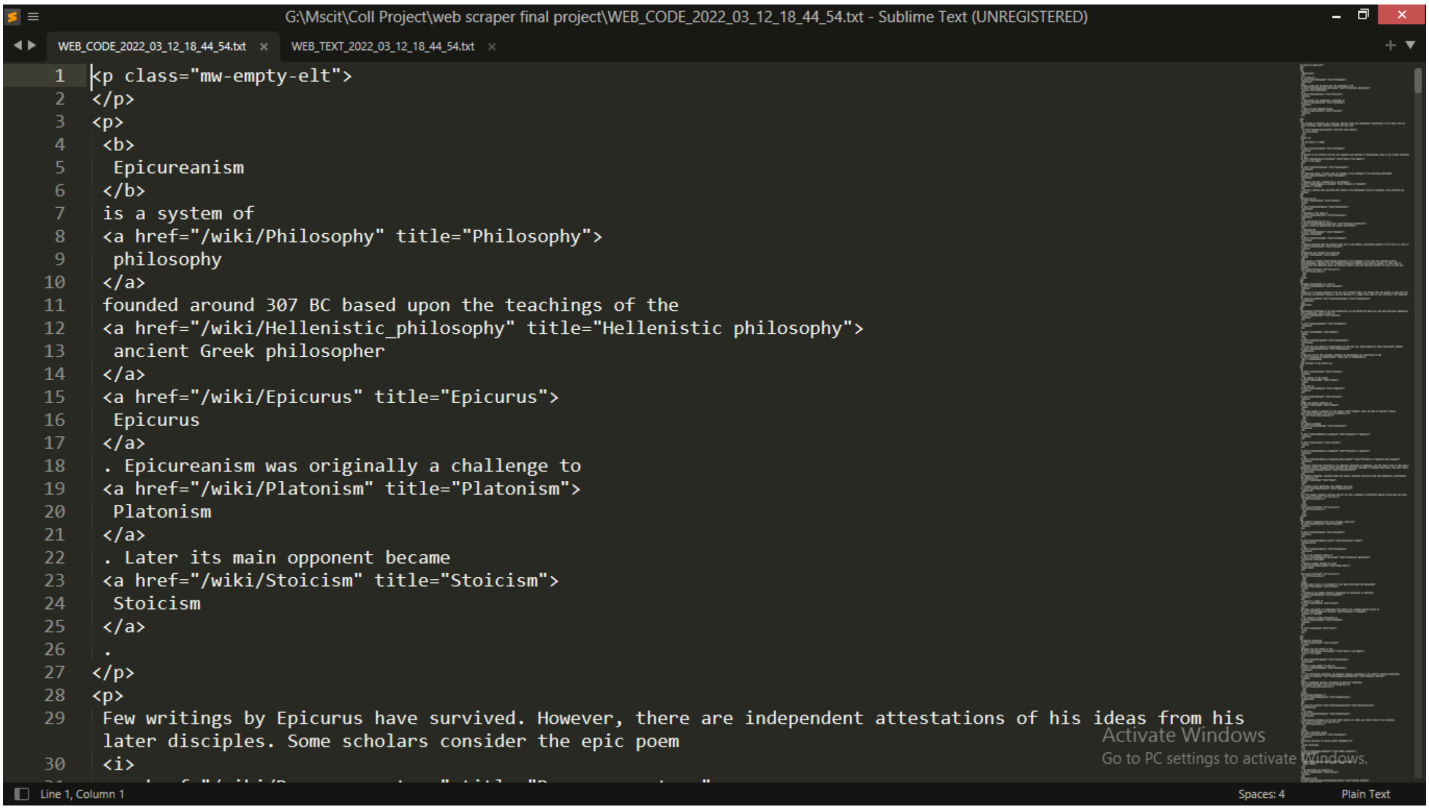
Epicureanism is a system of philosophy founded around 307 BC based upon the teachings of the ancient Greek philosopher Epicurus. Epicureanism was originally a challenge to Platonism. Later its main opponent became Stoicism.

Few writings by Epicurus have survived. However, there are independent attestations of his ideas from his later disciples. Some scholars consider the epic poem De rerum natura (Latin for On the Nature of Things) by Lucretius to present in one unified work the core arguments and theories of Epicureanism. Many of the scrolls unearthed at the Villa of the Papyri at Herculaneum are Epicurean texts. At least some are thought to have belonged to the Epicurean philosopher Philodemus. Epicurus also had a wealthy 2nd c. AD disciple, Diogenes of Oenoanda, who had a portico wall inscribed with tenets of the philosophy erected in Oenoanda, Lycia (present day Turkey).

Epicurus was an atomic materialist, following in the steps of Democritus. His materialism led him to a general attack on superstition and divine intervention. Following the Cyrenaic philosopher Aristippus, Epicurus believed that the greatest good was to seek modest, sustainable pleasure in the form of a state of ataraxia (tranquility and freedom from fear) and aponia (the absence of bodily pain) through knowledge of the workings of the world and limiting desires. Correspondingly, Epicurus and his followers generally withdrew from politics because it could lead to frustrations and ambitions which can directly conflict with the Epicurean pursuit for peace of mind and virtues.[1]

Although Epicureanism is a form of hedonism insofar as it declares pleasure to be its sole intrinsic goal, the concept that the absence of pain and fear constitutes the greatest pleasure, and its advocacy of a simple life, make it very different from "hedonism" as colloquially understood.

Epicureanism flourished in the Late Hellenistic era and during the Roman era, and many Epicurean communities were established, such as those in Antiochia, Alexandria, Rhodes, and Herculaneum. By the late 3rd century CE Epicureanism all but died out, being opposed by other philosophies (mainly Neoplatonism) that were now in the ascendant. Interest in Epicureanism was resurrected in the Age of Enlightenment and continues in the modern era.


G:\Mscit\Coll Project\web scraper final project\WEB_CODE_2022_03_12_18_44_54.txt - Sublime Text (UNREGISTERED)
WEB_CODE_2022_03_12_18_44_54.txt
WEB_TEXT_2022_03_12_18_44_54.txt
Activate Windows
Go to PC settings to activate Windows.
Line 1, Column 1
Spaces: 4
Plain Text

```
<p class="mw-empty-elt">
</p>
<p>
 <b>
  Epicureanism
 </b>
 is a system of
 <a href="/wiki/Philosophy" title="Philosophy">
  philosophy
 </a>
 founded around 307 BC based upon the teachings of the
 <a href="/wiki/Hellenistic_philosophy" title="Hellenistic philosophy">
  ancient Greek philosopher
 </a>
 <a href="/wiki/Epicurus" title="Epicurus">
  Epicurus
 </a>
 . Epicureanism was originally a challenge to
 <a href="/wiki/Platonism" title="Platonism">
  Platonism
 </a>
 . Later its main opponent became
 <a href="/wiki/Stoicism" title="Stoicism">
  Stoicism
 </a>
 .
</p>
<p>
 Few writings by Epicurus have survived. However, there are independent attestations of his ideas from his
 later disciples. Some scholars consider the epic poem
 <i>
```

8. Conclusion

The project has a lot of data collection capabilities. The merits of this project are its simplicity and friendliness. The programme has been designed to be as user-friendly as possible, allowing anyone with internet connection to operate it. This project takes care of all the intricacies without putting the project at danger. All of the goals were met satisfactorily. The system's performance has been judged to be satisfactory.