Software Requirements Specification

For

Web Scraping Using Java

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Topic** | | **Page**  **No** |
| Table of Content | |  |
| Revision History | |  |
| 1 | Introduction |  |
|  | 1.1 Purpose of the Project |  |
|  | 1.2 Target Beneficiary |  |
|  | 1.3 Project Scope |  |
|  | 1.4 References |  |
| 2 | Project Description |  |
|  | 2.1 Reference Algorithm |  |
|  | 2.2 Data/ Data structure |  |
|  | 2.3 SWOT Analysis |  |
|  | 2.4 Project Features |  |
|  | 2.5 User Classes and Characteristics |  |
|  | 2.6 Design and Implementation Constraints |  |
|  | 2.7 Design diagrams |  |
|  | 2.8 Assumption and Dependencies |  |
| 3 | System Requirements |  |
|  | 3.1 User Interface |  |
|  | 3.2 Software Interface |  |
|  | 3.3 Database Interface |  |
|  | 3.4 Protocols |  |
| 4 | Non-functional Requirements |  |
|  | 4.1 Performance requirements |  |
|  | 4.2 Security requirements |  |
|  | 4.3 Software Quality Attributes |  |
| 5 | Other Requirements |  |
| Appendix A: Glossary | |  |
| Appendix B: Analysis Model | |  |
| Appendix C: Issues List | |  |

|  |  |  |
| --- | --- | --- |
| 1 | INTRODUCTION | |
|  | 1.1 Purpose of the Project | Describe the scope of this project by stating and justifying  the problem statement of the project. Present will clear motivation to execute the project. |
|  | 1.2 Target Beneficiary | Identify the prime beneficiaries of the project. |
|  | 1.3 Project Scope | Provide a short description of area of application of the  software, include relevant benefits, objectives, and goals. State clearly the requirement and deliverables of the project. |
|  | 1.4 References | List all documents or Web addresses to which this SRS refers. |
| 2 | PROJECT DESCRIPTION | |
|  | 2.1 Reference Algorithm | State the reference algorithm for the project and identify the required data structure Or/Add  design algorithm justifying the methodology of the project |
|  | 2.2 Characteristic of Data | Present with the characteristic of the dataset used for the project. Provide the primary and secondary source of the data, along with sampling techniques. Explain the statistical  method used for data processing (**if any**). |
|  | 2.3 SWOT Analysis | Present with a justification to support your project. |
|  | 2.4 Project Features | Summarize the major features the product contains or the  significant functions that it performs or lets the user perform. (Level 2 USE Case diagram) |
|  | 2.5 User Classes and Characteristics | Identify the various user classes that you anticipate will use this product. |
|  | 2.6 Design and Implementation Constraints | Present hardware boundary conditions (timing requirements, memory requirements); interfaces to other applications; specific technologies, and tools to be used; parallel operations; language requirements; communications  protocols; security considerations; design conventions or programming standards. |
|  | 2.7 Design diagrams | Present all the required Diagram (USE –Case, Class Diagram, Activity, Sequence, Data Flow diagram and State Diagram. (Major project should include Collaboration and  Deployment Diagram too) |

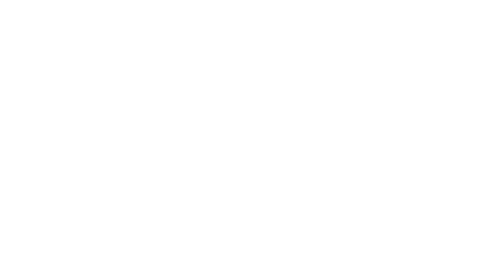
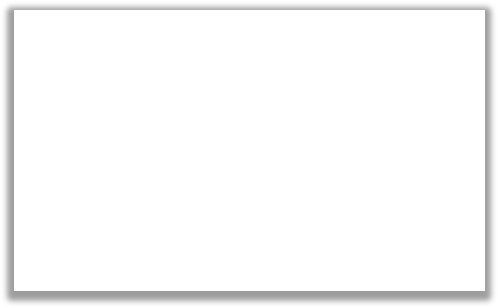
|  |  |  |
| --- | --- | --- |
|  | 2.8 Assumption and Dependencies | List any assumed factors (as opposed to known facts) that  could affect the requirements stated in the SRS. Also identify any dependencies the project has on external factors. |
| 3 | SYSTEM REQUIREMENTS | |
|  | 3.1 User Interface | Define the software components for which a user interface is needed. |
|  | 3.2 Software Interface | Describe the connections between modules. Describe the  services needed and the nature of communications. Describe detailed application programming interface protocols. |
|  | 3.3 Database Interface | Explain the Database management system used |
|  | 3.4 Protocols | Describe the requirements associated with any protocol deployed in the project. Specify any communication security or encryption issues, data transfer rates, andsynchronization mechanisms |
| 4 | NON-FUNCTIONAL REQUIREMENTS | |
|  | 4.1 Performance requirements | If there are performance requirements for the product under various circumstances, state them. Specify the timing relationships for real time systems. State performance  requirements for individual functional requirements or features |
|  | 4.2 Security requirements | Specify any requirements regarding security or privacy issues surrounding use of the product or protection of the data used or created by the product. Define authentication, verification and validation of the system. Refer to any external policies or regulations containing security issues that affect the product. |
|  | 4.3 Software Quality Attributes | Explain: adaptability, availability, correctness, flexibility,  interoperability, maintainability, portability, reliability, reusability, robustness, testability, and usability. |
| 5 | Other Requirements | Define any other requirements not covered elsewhere in the SRS. |
| Appendix A: Glossary | | Define all the terms necessary to properly interpret the SRS, including acronyms and abbreviations. |
| Appendix B: Analysis Model | | Pertinent analysis models used for this project |
| Appendix C: Issues List | | This is a dynamic list of the open requirements issues. |

|  |
| --- |
| INTRODUCTION |
| 1.1 Purpose of the Project  Over the past two years, we all have been facing the wrath of the COVID-19 pandemic which caused a lot of people to lose their livelihoods and face a psychological crisis. This motivated us to work on this project of extracting covid data of different countries which later can be used in analysing how did it affect different parts of the world. |
| 1.2 Target Beneficiary  Our project will be in the benefit of data analysts, who collect data and drive conclusions out of it, and growth hackers, who  use web scraping to gather the contact information of leads, as well as find out what customers are saying about their services, products, brands, and those of their competitors.  Businesses and firms also use web scraping to gather information about competitors’ price changes, campaigns or promotions, so they can react quickly and adjust their prices or operations if need be. Market research companies use scrapers to pull data from social media or online forums for things like customer sentiment analysis. Others scrape data from product sites like Amazon or eBay to support competitor analysis. |
| * 1. Project Scope   Here are some industries where web scraping is being used :   * + - News portals: to aggregate articles from different data sources like Reddit   / Forums / Twitter / specific news websites   * + - Search Engine     - The travel industry (flight/hotels prices comparators)     - E-commerce, to monitor competitor prices     - Journalism: also called “Data-journalism”     - SEO     - Data analysis     - “Data-driven” online marketing     - Market research   1. References |
| * [https://datascience.blog.wzb.eu/2020/12/04/robust-data-collection-](https://datascience.blog.wzb.eu/2020/12/04/robust-data-collection-via-web-scraping-and-web-apis/) [via-web-scraping-and-web-apis/](https://datascience.blog.wzb.eu/2020/12/04/robust-data-collection-via-web-scraping-and-web-apis/) * [https://able.bio/DavidLandup/introduction-to-web-scraping-with-](https://able.bio/DavidLandup/introduction-to-web-scraping-with-java-jsoup--641yfyl) [java-jsoup--641yfyl](https://able.bio/DavidLandup/introduction-to-web-scraping-with-java-jsoup--641yfyl) * [https://kevincurran.org/dissertations/2018%20Thesis%20Evan%20G](https://kevincurran.org/dissertations/2018%20Thesis%20Evan%20Gallagher%20-%20Scraping%20Websites%20for%20Law.pdf) [allagher%20-%20Scraping%20Websites%20for%20Law.pdf](https://kevincurran.org/dissertations/2018%20Thesis%20Evan%20Gallagher%20-%20Scraping%20Websites%20for%20Law.pdf) * <https://www.happiestminds.com/whitepapers/website-scraping.pdf> |

* [http://www.geog.leeds.ac.uk/courses/other/programming/practicals/g](http://www.geog.leeds.ac.uk/courses/other/programming/practicals/general/web/scraping-intro/2.html) [eneral/web/scraping-intro/2.html](http://www.geog.leeds.ac.uk/courses/other/programming/practicals/general/web/scraping-intro/2.html)
* <https://www.webscrapingapi.com/java-web-scraping/>
* <https://librarycarpentry.org/lc-webscraping/>

PROJECT DESCRIPTION

* 1. Reference Algorithm
  2. Characteristics of Data
  3. SWOT Analysis



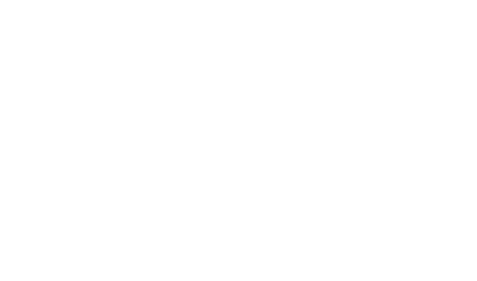
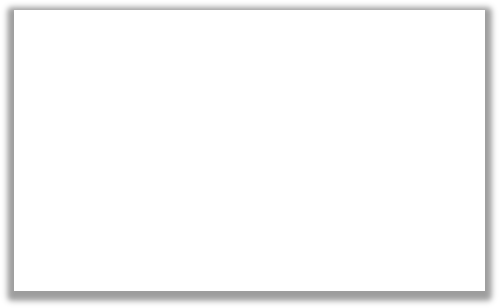
**Strength:**

Usable across all platforms

Saves a lot of time as day’s of manual work

can be done in hours

Easy to implement

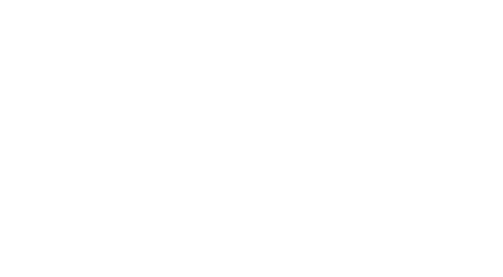
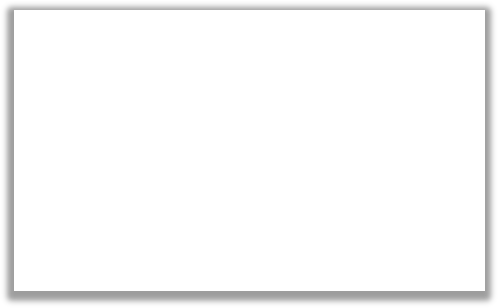


**Weakness:**

Some tools used for web scraping are not able to extract data on large scale

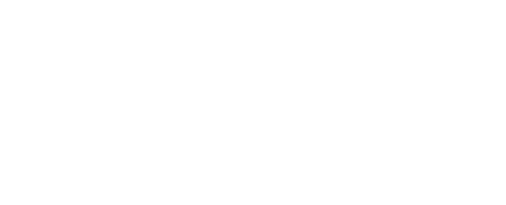
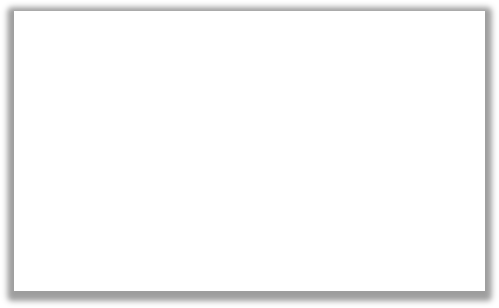
It is not easy to handle complex websites

The structure of the website can change anytime



**OPPORTUNITIES:**

Web scraping has variety of applications in various fields including social media sentimental analysis, e commerce compliance management, Industry statistics and insights, lead generation etc. Any application involving the use of data would require a web scraper to fetch data.



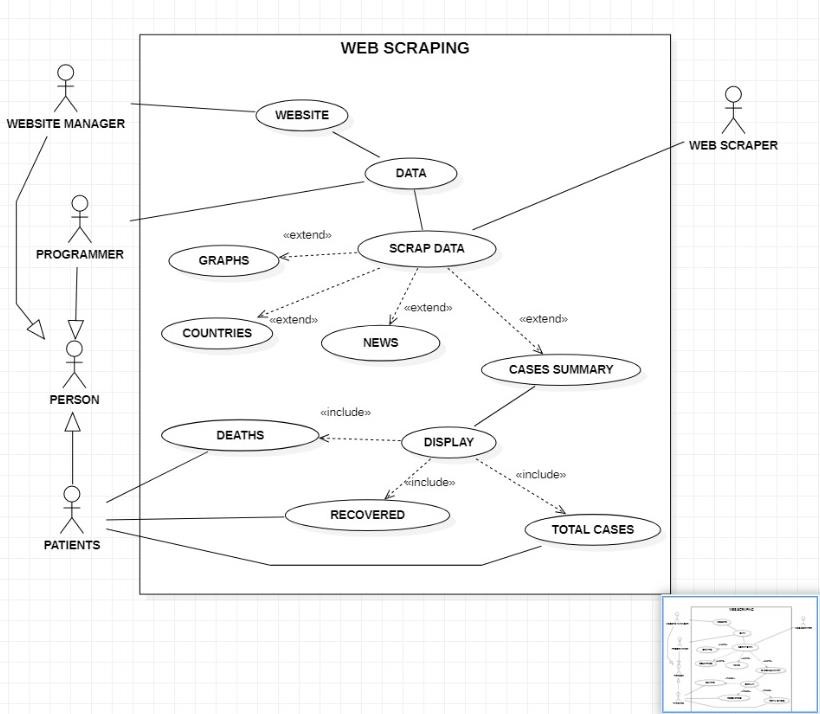
**THREATS:**

Web scraper might not work if the target url is not valid.

Web scraper might be used for suspicious

activities including hacking.

* 1. Project Features



Here programmer can access the data which is scrapped by the web scrapper. The scrapped data may include information regarding covid news, graphs, countries and case summary etc. The case summary data is displayed which includes deaths by covid, recovered patients and total cases.

# User Classes and Characteristics

The code contains one main class which is responsible for all the operations performed, from connecting to page to scraping the data.

This class contains a method which works along with exception handling in order to maintain the regular flow of our program.

If there is no problem, the program carries normally but in case of any exception it is handled by program as we throws an exception defined along with the definition of this method.

We take a variable of string type to store the URL of the website of which we need to scrap the data.

Using the library, we defined an object of document type by importing the required packages.

Then connect function is used to connect to the website whose URL is mentioned followed by get function.

This helps us to load the webpage in this document object.

Then another variable of type elements is used to fetch the specific part of the data from the website that we need.

We get the data in the HTML format for multiple elements.

A for loop is used to fetch the data elements one by one which we had requested for converting this HTML type data into the simple text format, we use the lambda expression.

This is a special function which can be created without belonging to any class. By passing the header value and the class name in which the data is stored, followed by

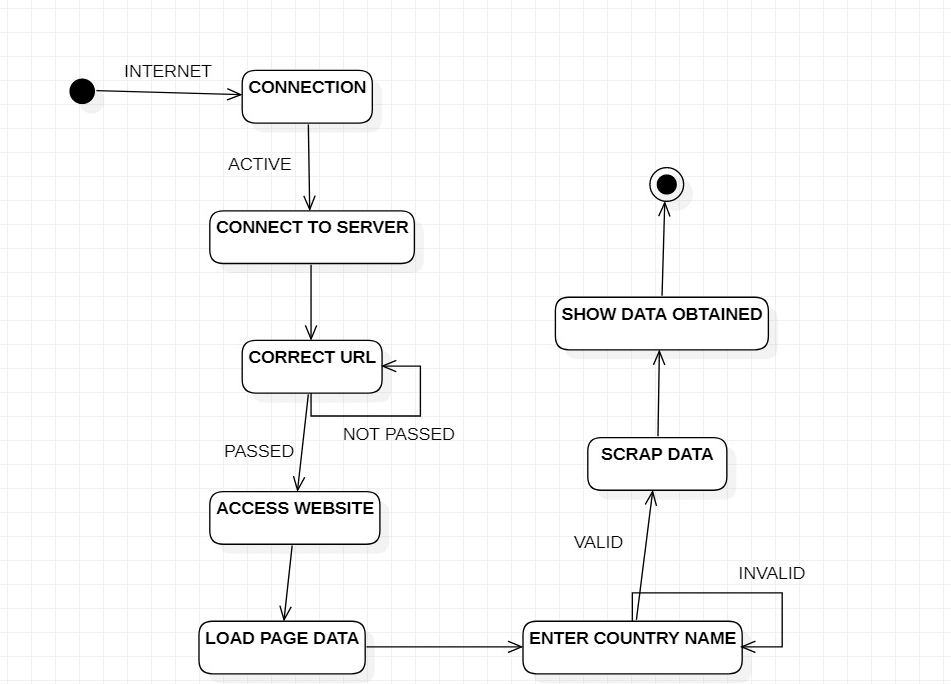
.text function, we can scrap data into text format.

# Design and Implementation Constraints

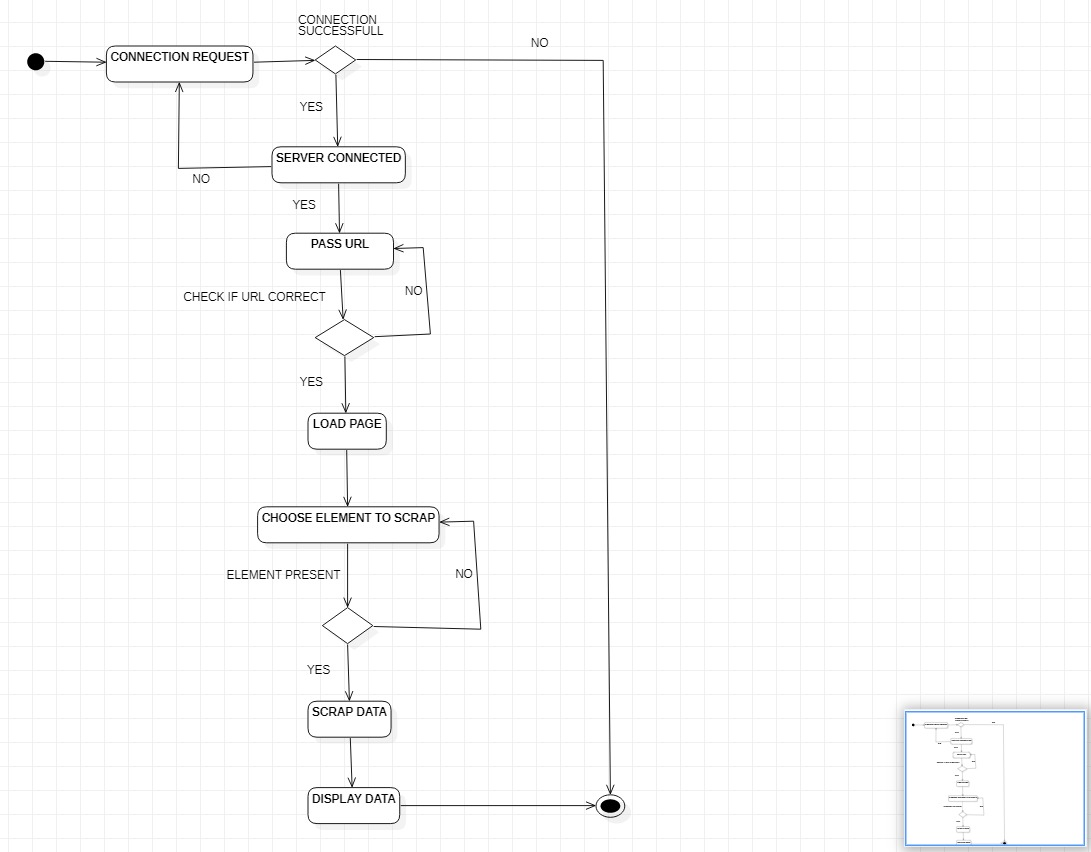
|  |  |
| --- | --- |
| **Name** | **Description** |
| Operating System | Windows 10 |
| RAM | 8GB |
| Processor | I5 processor |
| Screen Size | 14 inches |
| Interface | Console output |
| Tools & Technologies | IntelliJ, Maven project, Jsoup |
| Parallel Operation | Making connection with the web page and loading into document object |
| Language | Java, HTML |
| Communication Protocol | HTTP protocol |

* 1. Design diagrams

STATE CHART DIAGRAM



ACTIVITY DIAGRAM



|  |
| --- |
| USE CASE DIAGRAM |
| 2.8 Assumption and Dependencies  **Assumption: -**  Assuming the data published on the site is correct and to the best of the knowledge of the country providing it, we can get the proper data after scraping.  **Dependencies: -**  It depends on the internet connection speed, how quick you can get the scraped data Also, the connection request sent through the HTTP protocol is received properly and adequate reply is received. |
| SYSTEM REQUIREMENTS |
| 3.1 User Interface  The user interface is the system used by the user or the programmer to perform web scraping. The minimum system requirements recommended are: -   1. 8 GB RAM 2. 2 GHz processor 3. IDE for writing code 4. Windows 10 operating system or equivalent 5. I5 processor |

# Software Interface

* + 1. To establish the best communication, we require a web browser on which we will visit the required website from which we need to scrap the data. The web browser tab would help us in getting the URL of the website.
    2. Then we require an IDE i.e., the coding interface wherein we will write the code to pass the URL and fetch the data. As data scraping is done with the help of Java so we will be using the eclipse IDE or IntelliJ IDE to write our code.
    3. For the code, the required packages are imported and the functionalities contained in them are used. A library named JSOUP is used to fetch the data from the website. Document object is used to load the HTML page and we make the connection from the website using the connection function. A variable of type element is used to fetch the specific part of data from the website.
    4. Example, if we want to fetch the covid data of any region, we can load the HTML page using the Document type object by passing the URL of the website in the connection function. By inspecting the web page, we can check where the specific element is stored and then using the class name, we can fetch that part using the element type object in the code.

# Database Interface

* + 1. The data scrapped has to be stored for further analysis. Data extracted can be stored in various forms like a CSV file on the local machine which we are using or we can use any other database.
    2. For storing the extracted data in a proper structured format, we can use the MySQL database and for storing data in a random manner we can use a NoSQL database like MongoDB.
    3. MongoDB is very efficient to use and storing data and querying on the data is quite easier.
    4. In MySQL, we need to create a table and provide a structure to the table as to which columns need to be added and then accordingly store our data.
    5. For MongoDB, we create a database and then insert the data directly into the database and also process queries using commands.

# Protocols

Web scraper uses the Hyper Text Transfer Protocol (HTTP) to initiate the scraping process.

It uses a client/server model. The HTTP client is used to open the connection and send the request to the HTTP server to fetch a page. In the response to the client request, the server answers with a response and upon fulfilling the requirement, closes the connection. As it is a stateless protocol which means that each transaction performed is independent of each other. The protocol uses the get method which is used to request the data from a specific path.

Web scraping a web page involves fetching it and extracting from it. WSAPI is the platform that enables an organization to extend their existing web-based system, as well-designed set of services for creating new channels, developer integration or partner integration. It helps to offer clean and structured data from existing websites, so that the data can be effortlessly consumed by different systems. The data that is being

|  |
| --- |
| exposed through these APIs can be monitored, transformed and controlled easily. The inherent design helps developers to incorporate website changes without affecting the extraction logic by moving them to configurations. |
| NON-FUNCTIONAL REQUIREMENTS |
| 4.1 Performance requirements  We should implement optimized algorithms for faster and accurate comparisons. |
| 4.2 Security requirements  The general belief is that everything you see online is free to scrape and re-use. This is probably the biggest misconception out there regarding web scraping.  Scraping data from public sites is perfectly legal. This refers to the data and information on websites which is obtained without the need to log in or authenticate one’s identity. As a general rule, it is illegal to scrape any personal information without the person’s consent or without any legal motivation. It is illegal to scrape any openly accessible data like images, songs, articles, etc. that are intellectual properties of any business or individual. Because their owners have full control over their use and reproduction scrapers require explicit consent in order to extract them. |
| 4.3 Software Quality Attributes  USABILITY:  Web scraping saves us the trouble of manually downloading or copying any data and automates the whole process. With the help of web scraping, one can extract data from any website, no matter how large is the data, on your computer. Moreover, websites may have data that we cannot copy and paste. Web scraping can help extract any kind of data that we want.  When you extract web data with the help of a web scraping too, one would be able to save the data in a format such as CSV. You would then be able to retrieve, analyze and use the data the way you want.  RELIABILITY:  We are using our web scraper to scrape data regarding covid deaths and active cases. Since covid is such a disease where getting accurate data is very important that is where our web scraper proves out to be very reliable as it provides the data in real time with least possible lag and without any discrepancies. Also, our web scraper works efficiently as a failure free operation.  TESTABILITY:  Testability refers to that the program requirements must be designed to allow for testing. In this aspect our project is quite testable as it doesn’t require anything much except the URL of the webpage to be scrapped. |

## ROBUSTNESS:

By robust data collection via web scraping, we mean that the data collection process satisfies at least the following requirements:

* It runs largely without manual intervention and handles upcoming issues such as loss of internet connection, server failures, rate limit exceedance etc. by itself;
* If it breaks (or we manually interrupt it), there’s no or only few loss of already collected data.
* It reports and records failures or unexpected situations.

Since our web scraper fulfils all of the above requirements, we can say that our web

scraper is quite resilient.

## ADAPTABILITY:

Adaptability refers to processes being designed flexible or not, for example, require specific order of operations. In that aspect our web scraper is quite adaptable and flexible as we are using the jsoup library which is quite adaptable and easy to use.

## REUSABILITY:

Reusability refers to how easy and effective the reuse of the software is. Our web scraper is quite easy to use and understand and also can be used for multiple purposes just by making minor tweaking.

## PORTABILITY:

Portability refers that the program must not require a specific environment to run. Since our project is based on java programming language, it quite portable Java programs are **portable**, which means that the same bytecode program can run on any computer system that has a Java interpreter. The source program does not have to be changed to meet the particular needs of a particular computer systems. Also, we are using the jsoup java library to extract data which is Java library for working with real-world HTML. It provides a very convenient API for fetching URLs and extracting and manipulating data, using the best of HTML DOM methods and CSS selectors. jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup. Also, jsoup is an open-source project distributed under the liberal [MIT license](https://jsoup.org/license) which should not pose problem while working on any system or environment.

APPENDIX A (GLOSSARY)

|  |  |
| --- | --- |
| WORD | MEANING |
| OPERATING SYSTEM | An Operating System (OS) is a software that acts as an interface between computer hardware components and the user. Every  computer system must have at least one |

|  |  |
| --- | --- |
|  | operating system to run other programs. he OS helps you to communicate with the computer without knowing how to speak the computer’s language |
| IntellijIDEA | IntelliJ IDEA is an integrated development environment written in Java for developing computer software. It is developed by JetBrains, and is available as an Apache 2 Licensed community edition, and in a  proprietary commercial edition. |
| MAVEN | Maven is a project management and comprehension tool that provides developers a complete build lifecycle framework.  Development team can automate the project's build infrastructure in almost no time as Maven uses a standard directory layout and a default build lifecycle. |
| JAVA | Java is a [high-level,](https://en.wikipedia.org/wiki/High-level_programming_language) [class-based,](https://en.wikipedia.org/wiki/Class-based_programming) [object-](https://en.wikipedia.org/wiki/Object-oriented_programming) [oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) [programming language](https://en.wikipedia.org/wiki/Programming_language) that is designed to have as few implementation [dependencies](https://en.wikipedia.org/wiki/Dependency_(computer_science)) as possible. The [syntax](https://en.wikipedia.org/wiki/Syntax_(programming_languages)) of Java is similar to [C](https://en.wikipedia.org/wiki/C_(programming_language)) and [C++](https://en.wikipedia.org/wiki/C%2B%2B), but has fewer [low-level](https://en.wikipedia.org/wiki/Low-level_programming_language) facilities than either  of them. |
| HTML | The HyperText Markup Language, or HTML is the standard [markup](https://en.wikipedia.org/wiki/Markup_language)  [language](https://en.wikipedia.org/wiki/Markup_language) for documents designed to be displayed in a [web browser](https://en.wikipedia.org/wiki/Web_browser). It can be assisted by technologies such as [Cascading Style](https://en.wikipedia.org/wiki/Cascading_Style_Sheets) [Sheets](https://en.wikipedia.org/wiki/Cascading_Style_Sheets) (CSS) and [scripting languages](https://en.wikipedia.org/wiki/Scripting_language) such  as [JavaScript.](https://en.wikipedia.org/wiki/JavaScript) |
| HTTP Protocol | **HTTP** is a [protocol](https://developer.mozilla.org/en-US/docs/Glossary/Protocol) for fetching resources such as HTML documents. It is the foundation of any data exchange on the Web and it is a client-server protocol, which means requests are initiated by the recipient, usually the Web browser. A complete document is reconstructed from the different sub- documents fetched, for instance, text, layout description, images, videos, scripts, and more. |
| CSV File | A CSV (comma-separated values) file is a text file that has a specific format which allows  data to be saved in a table structured format. |
| MySQL | MySQL is a relational database management system based on the Structured Query Language, which is the popular language for  accessing and managing the records in the database. MySQL is open-source and free |

|  |  |
| --- | --- |
|  | software under the GNU license. It is supported by **Oracle Company.** |
| NoSQL | A NoSQL (originally referring to "non-[SQL](https://en.wikipedia.org/wiki/SQL)" or "non-relational") [database](https://en.wikipedia.org/wiki/Database) provides a mechanism for [storage](https://en.wikipedia.org/wiki/Computer_data_storage) and [retrieval](https://en.wikipedia.org/wiki/Data_retrieval) of data that is modelled in means other than the tabular relations used in [relational databases](https://en.wikipedia.org/wiki/Relational_database). |
| MONGODB | MongoDB is a [source-available](https://en.wikipedia.org/wiki/Source-available) [cross-](https://en.wikipedia.org/wiki/Cross-platform) [platform](https://en.wikipedia.org/wiki/Cross-platform) [document-oriented](https://en.wikipedia.org/wiki/Document-oriented_database)  [database](https://en.wikipedia.org/wiki/Document-oriented_database) program. Classified as  a [NoSQL](https://en.wikipedia.org/wiki/NoSQL) database program, MongoDB uses [JSON](https://en.wikipedia.org/wiki/JSON)-like documents with  optional [schemas.](https://en.wikipedia.org/wiki/Database_schema) MongoDB is developed by [MongoDB Inc.](https://en.wikipedia.org/wiki/MongoDB_Inc) and licensed under  the [Server Side Public License](https://en.wikipedia.org/wiki/Server_Side_Public_License) (SSPL). |
| JSOUP | Jsoup is an open-source Java library designed to parse, extract, and manipulate data stored in HTML  documents. |