Indian Institute of Technology, Indore

Software Engineering Project SRS

CS258 Minor Project

Presented by:

Ram Preetham Tammireddi
Sujal Mishra
Shivashish Sharma
Vivek Bhojwani
Shreyash Raj

Professor:

Dr. Puneet Gupta

March 2023

Índice

1. Introduction

The purpose of this SRS document is to define the requirements for a web crawler and scraper application that uses Optical Character Recognition (OCR) technology to extract data from websites. The application will be designed to crawl and scrape data from multiple websites and save it in a structured format. The OCR technology will be used to extract data from images, PDFs, and other non-textual data

A web crawler is a program that automatically navigates the web by browsing links and downloading content from websites. The purpose of a web crawler can range from indexing the web for search engines to analyzing websites for specific information.

A web scraper is a program that extracts specific information from websites. It is used to extract data from websites for a variety of purposes such as data analysis, data migration, or to gather specific information for a database. Web scraping can be done manually or with the use of a web scraper program.

Both web crawlers and web scrapers are important tools for data extraction and analysis, and they are widely used in various industries. However, it's important to be mindful of ethical and legal considerations when using these tools, as some websites may restrict or prohibit automated scraping.

2. Scope

Web scraping, also known as web data extraction, is the process of collecting data from websites. The scope of web scraping is vast and can be applied in various fields, including e-commerce, market research, social media analysis, and more. Web scraping can be used to extract different types of data such as text, images, videos, pricing data, product descriptions, customer reviews, and more. Web scraping can be done manually, but it is time-consuming and tedious. Automated web scraping using software tools can significantly speed up the process and make it more efficient.

The web crawler and scraper application will be developed for users who require data from websites and images for research or analysis purposes. The application will be capable of crawling and scraping data from websites of different types and formats, and from images in different languages. The application will be designed to run on desktop computers running Windows, macOS, and Linux operating systems.

3. Functional Requirements

The web crawler and scraper application should provide the following functionality:

- User should be able to input the list of URLs of the websites to be crawled
- User should be able to select the data to be crawled and scraped, such as text, images, and videos
- The application should be able to crawl the websites and extract the data
- The application should be able to process the extracted data using OCR technology
- The application should be able to save the processed data in a structured format, such as CSV, JSON, or XML The application should be able to handle different types of data such as text, images, and videos.
- The application should be able to handle different types of websites such as static, dynamic and AJAX based sites. The application should be able to handle images in different languages.
- The application should be able to handle images in different formats such as JPG, PNG, and TIFF.
- The application should provide a user-friendly interface for input and output of data

4. Non-Functional Requirements

The web crawler and scraper application should also meet the following non-functional requirements:

- The application should be able to handle large volumes of data and crawl websites quickly
- The application should be scalable and easy to maintain The application should be secure and protect user data from unauthorized access
- The application should be compatible with the latest versions of web browsers and operating systems The OCR technology used in the application should be accurate and reliable.
- The application should have a low memory footprint and be optimized for performance

5. Constraints

The following constraints should be considered while designing the web crawler and scraper application:

- The application should comply with website terms of service and usage policies.
- The application should not harm or damage websites or servers.
- The OCR technology used in the application may not be accurate for certain languages and fonts.

6. Assumptions

The following assumptions are made while developing the web crawler and scraper application:

- The user has basic knowledge of websites and their structure.
- The user has internet access and a compatible web browser installed on their computer.
- The images provided for processing by the user are of high quality and readable.

7. Dependencies

The following dependencies should be considered while developing the web crawler and scraper application:

- The application should use compatible web technologies and libraries to interact with websites and extract data.
- The OCR technology used in the application should be compatible with the latest versions of operating systems and web browsers.

8. User Interface

The user interface of the web scraper application should be intuitive and easy to use. The application should provide a graphical user interface (GUI) with the following features:

- Input box for the website URL
- Options for selecting data to scrape
- Options for saving data in different formats
- Start and Stop buttons for scraping data
- Progress bar to indicate scraping progress

9. Purpose of SRS

The purpose of a Software Requirement Specification (SRS) is to document the requirements and specifications for a software project. It serves as a foundation for the software development team to understand what needs to be built, and how it should function. The SRS outlines the goals, features, functionalities, and constraints of the software project. It also defines the scope of the project, along with the system and user requirements. The SRS document helps ensure that the software product meets the needs and expectations of the stakeholders, and provides a clear understanding of the software system to be developed. It also serves as a reference for the development team throughout the project lifecycle, and helps ensure that the final product meets the quality standards and is delivered on time and within budget.

10. Conclusion

This SRS document has defined the requirements for a web scraper application. The application should be designed to scrape data from different types of websites and save it in a structured format. The application should also be scalable, secure, and easy to maintain. The user interface should be user-friendly and intuitive. The application should comply with website terms of service and usage policies and should not be used for any illegal purposes.