BOHR

BOHR International Journal of Computer Science
2022, Vol. 1, No. 1, pp. 6–10
DOI: 10.54646/bijcs.2022.02
www.bohrpub.com

METHODS

# Implementation of web application for disease prediction using AI

**Manasvi Srivastava, Vikas Yadav and Swati Singh***

IILM Academy of Higher Learning, College of Engineering and Technology, Greater Noida, India

***Correspondence:**
Swati Singh,
swati.singh@iilm.edu

**Received:** 04 August 2021; **Accepted:** 03 January 2022; **Published:** 19 January 2022

The Internet is the largest source of information created by humanity. It contains a variety of materials available in various formats, such as text, audio, video, and much more. In all, web scraping is one way. There is a set of strategies here in which we get information from the website instead of copying the data manually. Many web-based data extraction methods are designed to solve specific problems and work on *ad hoc* domains. Various tools and technologies have been developed to facilitate web scraping. Unfortunately, the appropriateness and ethics of using these web scraping tools are often overlooked. There are hundreds of web scraping software available today, most of them designed for Java, Python, and Ruby. There is also open-source software and commercial software. Web-based software such as Yahoo! Pipes, Google Web Scrapers, and Firefox extensions for Outwit are the best tools for beginners in web cutting. Web extraction is basically used to cut this manual extraction and editing process and provide an easy and better way to collect data from a web page and convert it into the desired format and save it to a local or archive directory. In this study, among other kinds of scrub, we focus on those techniques that extract the content of a web page. In particular, we use scrubbing techniques for a variety of diseases with their own symptoms and precautions.

**Keywords:** web scraping, disease, legality, software, symptoms

## Introduction

Web scraper is a processfor downloading and extracting important data by scanning a web page. Web scrapers work best when page content is either transferred, searched, or modified. The collected information is then copied to a spreadsheet or stored in a database for further analysis. For the ultimate purpose of analysis, data need to be categorized by progressively different developments, for example, by starting with its specification collection, editing process, cleaning process, remodeling, and using different models and various algorithms and end result. There are two ways to extract data from websites: the first is the manual extraction process and the second is the automatic extraction process. Web scrapers compile site information in the same way that a person can do that by removing access to a web page of the site, finding relevant information, and moving on to the next web page. Each website has a different structure that is why web scrapers are usually designed to search through a website. Web deletion can help in finding any kind of targeted information. We will then have the opportunity to find, analyze, and use information in the way we need it. Web logging therefore paves the way for data acquisition, speeds up automation, and makes it easier to access extracted data by rendering it in comma-separated values (CSV) pattern. Web publishing often removes a lot of data from websites, for example, monitoring consumer interests, price monitoring (e.g., price checking), advancing AI models, data collection, tracking issues, and so on. So, there is no doubt that web removal is a systematic way to get more data from websites. It requires two stages, mainly crawling and removal. A search engine is an algorithm designed by a person who goes through the web to look for specific information needed by following online links. Deleter is a specific tool designed to extract data from sites.

Web scraper will work that way; if the patient is suffering from any kind of illness or illness, he will add his symptoms and problems and when the crawl work starts, he will start scrolling and look for a disease from the database provided on the website and it will show the best disease like patient symptoms. When those specific diseases show up, they will also show the precautionary measures that the patient needs to take care in order to overcome them and treat the infection.

## Overview of web scraping

Web scraping is a great way to extract random data from websites and convert that data into organized data that can be stored and analyzed in a database. Web scraping is also known as web data extraction, web data removal, web harvesting, or screen scanning. Web scraping is a form of data mining. The whole purpose of the web crawling process is to extract information from websites and convert it into an understandable format such as spreadsheets, a database, or a CSV, as shown in **Figure 1**. Data such as item prices, stock prices, various reports, market prices, and product details can be collected with web termination. Extracting website-based information helps to make effective decisions for your business.

## Practices of web scraping

- Data scraping
- Research
- Web mash up—integrate data from multiple sources
- Extract business details from business directory websites, such as Yelp and Yellow pages
- Collect government data
- Market analysis

The web data scraper process, a software agent, also known as a Web robot, mimics browsing communication between web servers and a person using a normal web browser. Step by step, the robot enters as many websites as it needs, transfers its content to find and extract interesting data, and builds that content as desired. AP scraping APIs and frameworks address the most common web data scrapers involved in achieving specific recovery goals, as described in the following text.

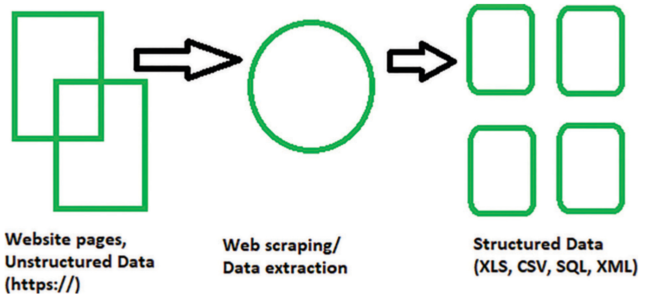

**FIGURE 1** | Web scraping structure.

## Hypertext transfer protocol

This method is used to extract data from static and dynamic web pages. Data can be retrieved by sending Hypertext Transfer Protocol (HTTP) requests to a remote web server using a socket system.

## Hyper text markup language

Exploring languages for query data, such as XQuery and Hyper text query language, can be used to scan hyper text markup language (HTML) pages and retrieve and modify content on the page.

## Release structure

The main purpose is to convert the published content into a formal representation for further analysis and retention. Although this last step is on the side of web scraping, some tools are aware of post-results, providing memory data formats and text-based solutions, such as cables or files (XML or CSV files).

## Literature survey

Python has a rich set of libraries available for downloading digital content online. Among the libraries available, the following three are the most popular ones: BeautifulSoup, LXml, and RegEx. Statistical research performed on the available data sets indicated that RegEx was able to deliver the requested information at an average rate of 153.6 ms. However, RegEx has limitations of data extraction of web pages with internal HTML tags. Because of this demerit, RegEx is used to perform complex data extraction only. Some libraries, such as BeautifulSoup and LXml, are able to extract content from web pages in a complex environment that has yielded a response rate of 457.66 and 203 ms, respectively.

The main purpose of data analysis is to get useful information from data and make decisions based on that analysis. Web deletion refers to the collection of data on the web. Web scraping is also known as data scraping. Data analysis can be divided into several steps such as cleaning and editing. Scrapy is the most widely used source of information needed by the user. The main purpose of using scrapy is to extract data from its sources. Scrapy, which crawls on the web and is based on python programing language, is very helpful in finding the data we need by using the URLs needed to clear the data from its sources. Web scraper is a useful API to retrieve data from a website. Scrapy provides all the necessary tools to extract data from a website, process data according to user needs, and store data in a specific format as defined by users.

The Internet is very much looking at web pages that include a large number of descriptive elements including text, audio, graphics, video, etc. This process, called web scraping, is mainly responsible for the collection of raw data from the website. It is a process in which you extract data automation very quickly. The process enables us to extract specific data requested by the user. The most popular method used is to create individual web data structures using any known language.

# Experimental work

## Technology used

### *Firebase*

For database, we have used Cloud Firestore from firebase. It is a real-time NoSQL database that stores two key-value data in the form of collections and documents.

### *Tensorflow*

TensorFlow is used to train the database model and to make predictions. There are various algorithms for modeling training or using line format in our project.

## JavaScript frameworks

### *Node.js*

Node.js is an open-source, cross-platform, JavaScript runtime environment running back to V8 engine and extracting JavaScript code without a web browser.

Our rewriting code is written for Node.js as it is a fast and platform language.

### *ElectronJS*

Electron is a framework for building native applications with web technologies such as JavaScript, HTML, and CSS. As electron is used to create a short web application, it helps us write our code and thus reduces the development time.

### *ReactJS*

React makes it less painful to create interactive UIs. Design a simple view of each state in your app and React will carefully review and provide relevant sections as your data changes.

React can also render to the server using Node and has the power of mobile applications using React Native.

# Python

Python is a high-level programing language translated into high-level translations.

In this project, various libraries, such as pandas, NumPy, good soup, etc., are used to create our database. Pandas and NumPy are used to filter and process data needed to train our model by extracting and removing them from a separate data source.

# Compatibility

## OS X

Only 64-bit binaries are provided for OS X, and the lower version of OS X is supported by OS X 10.9.

## Windows

Electron supports Windows 7 and later, but older versions of the OS are not supported.

Both x86 and amd64 (x64) binary are provided for Windows and are not supported in the ARM version of Windows.

## Software used

### *VSCode*

Visual Studio Code is a freeware source code editor developed by Microsoft for Windows, Linux, and macOS. Features include debugging support, syntax highlighting, intelligent coding, captions, reuse code, and embedded Git.

### *Google colab notebook*

Colaboratory, or Colab for short, is a product of Google Research that allows developers to write and use Python code through their browsers. Google Colab is an excellent tool for in-depth learning activities. It is a compact Jupyter notebook that needs no setup and has an excellent free version, providing free access to Google computer resources such as GPUs and TPUs.

### *PyCharm*

PyCharm is an integrated development platform used for computer programs, especially in the Python language, developed by the Czech company JetBrains.

## Data source

As we did not get more than 40 diseases, to get a dataset we have created our own dataset. The dataset that we have used for our training and testing process has taken from various sources. One of them is added below[1],[2].

### Use of scrapy

Scrapy is a framework for crawling and retrieving non-fiction data that can be used for the size of a supportive application, such as data mining, managed, or actual reported data. Apart from the way it was originally expected for Scrapy to be removed from the web, it could be used in the same way to extract data using APIs, for example, Amazon AWS, or as a very important web browser. Scrappy is written in Python. Let us take a Wiki example related to one of these problems. A simple online photo gallery can provide three options to users as defined by HTTP GET parameters at URL. If there are four ways to filter images with three thumbnail-sized options, two file formats, and a user-provided disabling option, then the same content set can be accessed with different URLs, all of which can be linked to the site. This carefully crafted combination creates a problem for the pages as they have to plan with an endless combination of subtitle changes to get different content.

## Methodology

The method used by the project to collect all the required data is extracted from various sources such as the CDC's database and Kaggle resources. Then, analyze the extracted data using texts written in the Python language according to project requirements. Pandas and NumPy are widely used to perform various functions on the database.

After sorting the data according to each need, it is then uploaded to the database. In the database, we have used Cloud Firestore as it is a real-time NoSQL database with extensive API support.

Furthermore, the TensorFlow project is used to train our model according to needs.

In this project, we predict the disease because of the given symptoms.

Training data set: 70%

Setting test data: 30%

TensorFlow supports Linear Regression, which is used to predict diseases based on the given indicators.

### Coding

Project Frontend is written using ReactJS & TypeScript. However, we have used the MaterialUI kit from Google ReactJS to speed up our development process.

To provide our app, Electron is used. Our web system supports macOS and Windows. Most of today's web app are written with the help of ElectronJS.

### Testing

The project is tested using an Electron built-in test framework called Spectron.

The project is being implemented in the browser. The output generated turns out to be completely consistent and the generated analysis is approximate.

Electron's standard workflow with Spectron can involve engineers who write unit tests in the standard TDD format and then write integration tests to ensure that acceptance criteria are met before approving a feature to be used. Continuous integration servers can ensure that all these tests are passed before they are incorporated into the production.

### Algorithm used

Linear regression is a standard mathematical method that allows us to study a function or relationship in a given set of continuous data. For example, we are given some of the corresponding x and y data points, and we need to study the relationship between them called a hypothesis.

In the event of a line reversal, the hypothesis is a straight line, i.e., Where the vector is called weights and b is a scale called bias. Weights and bias are called model parameters.

All we need to do is estimate the values of w and b from the set of data, given that the result of the assumption has produced the minimum cost J defined by the next cost function where m is the number of data points in the data provided. This cost function is also called the mean squared error.

To find the optimized value of the J's minimum parameters, we will be using a widely used optimizer algorithm called gradient descent. The following is a fake gradient descent code:

## Result discussion

The overall results of the project are useful in predicting diseases with the given symptoms. The script that was written to extract data can be used later to compile and format it according to needs.

Users can pick up symbols by typing them themselves or by selecting them from the given options. The training model will predict the disease according to it. Users are able to create their own medical profile, where they can submit their medical records and prescribed medication; this greatly helps

[1] https://www.sciencedirect.com/science/article/pii/S2352914817302253
[2] https://github.com/DiseaseOntology/HumanDiseaseOntology

us to feed our database and better predict disease over time, as some of these diseases occur directly during the season.

Moreover, the analysis performed showed a very similar disease, but the training model lacks the size of the database.

## Conclusion and future scope

The use of the Python program also emphasizes understanding the use of pattern matching and general expressions for web releases. Database data are compiled from factual reports and sent directly to government media outlets for local media where it is considered reliable. A team of experts and analysts who validate the information from a continuous list of more than 5,000 items is likely to be the site that collects data effectively. User-provided inputs are analyzed and deleted from the website, and the output is extracted as the user enters the user interface encounters. Output is generated in the form of text. This method is simple and straightforward to eradicate the disease from companies and provides vigilance against that disease.

For future work, we plan tests that aim to show the medication that a patient can take for treatment. In addition, we are looking to link this website to various hospitals and pharmacies for easy use.

## References

1. Sakthivadivel T, Gopalakrishna S, Sarathambekai S. A survey on python librariesused for social media content scraping. *Proceedings of the international conference on smart electronics and communication (ICOSEC 2020)*. Piscataway, NJ: IEEE (2020). p. 361–6.
2. Upadhyay S, Pant V, Bhasin S, Pattanshetti MK. *Articulating the construction of a web scraper for massive data extraction*. Piscataway, NJ: IEEE (2017).
3. Kulkarni A, Kalburgi D, Ghuli P. Design of predictive model for healthcare assistance using voice recognition. *2nd IEEE international conference on computational systems and information technology for sustainable solutions*. Piscataway, NJ: IEEE (2017). p. 61–4.
4. Dojchinovski D, Ilievski A, Gusev M. Interactive home healthcare system with integrated voice assistant. *42nd International convention on information and communication technology, electronics and microelectronics (MIPRO)*. Piscataway, NJ: IEEE (2019). p. 284–8.
5. Shahnawaz M, Singh P, Kumar P, Konidena A. Grievance redressal system. *Int J Data Min Big Data*. (2020) 1:1–4.