BOHR International Journal of Computer Science
2022, Vol. 1, No. 1, pp. 6–10
https://doi.org/10.54646/bijcs.002
www.bohrpub.com

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, Uttar Pradesh, India
**Corresponding Author: swati.singh@iilm.edu*

Abstract. 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. It 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. To enable Web Scraping, a variety of tools and technologies have been created. Regrettably, the propriety and ethics of employing these Web Scraping programmes are frequently neglected. There are hundreds of online scraping applications available today, the most of which are written in Java, Python, or Ruby. There is both commercial software and open source software. For novices in web cutting, web-based applications such as YahooPipes, Google Web Scrapers, and Outwit Firefox plugins are the finest options. 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 paper, among others the kind 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, Symptons.

INTRODUCTION

Web Scraper is a process for 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 needs 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 webpage of the site, finding relevant information, and moving on to the next web page. Each website has a different structure which 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. Web logging therefore paves the way for data acquisition, speeds up automation and makes it easier to access extracted data by rendering it in CSV pattern. Web publishing often removes a lot of data from websites for example, monitoring consumer interests, price monitoring eg 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 and he will start scrolling and look like a disease from the database provided on the website and will show the best disease like patient symptoms. And when those specific diseases show up, it will also show the precautionary

measures that the patient needs to take care of in order to overcome them and to treat the infection.

OVERVIEW OF WEB SCRAPING

Web scraping is an excellent method for extracting random data from websites and organising that data so that it can be saved and examined in a database. Web scraping is also known as data extraction from the online, data removal from the web, web harvesting, or screen scanning. Web scraping is a type of data mining. The goal of web crawling is to collect information from websites and transform it into a usable format, such as spreadsheets, databases, or comma-separated files (CSV), as illustrated in Figure 1. With web termination, data such as item pricing, stock price, different reports, market prices, and product details may be gathered. Extracting information from websites allows you to make more informed business decisions.

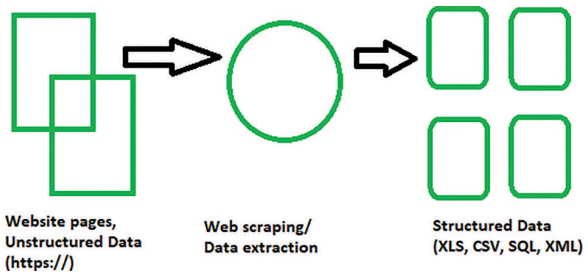

Figure 1. Web scraping structure.

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 on a normal Web browser. Step by step, the robot enters as many Websites as it needs, transfers its content to find and extracts interesting data and builds that content as desired. The following text describes how AP scraping APIs and frameworks meet the most frequent online data scrapers engaged in attaining various recovery goals:

Hypertext Transfer Protocol (HTTP)

This approach extracts data from both static and dynamic web pages. Data may be obtained by utilising a socket system to send HTTP requests to a remote web server.

Hyper Text Markup Language (HTML)

Exploration languages for query data, such as XQuery and Hyper Text Query Language (HTQL), can be used to scan HTML pages and obtain and alter material 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 final stage is on the Web scraping side, some technologies are aware of post-results, including 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: BeautifulSoup, LXml and RegEx. Statistical research was performed on the available data sets; indicates that RegEx was able to deliver the requested information at an average rate of 153.6 ms. However, RegEx has those 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 under a complex environment that has yielded a response rate of 457.66 ms and 203 ms respectively.

The main purpose of data analysis is to get useful information from data and make decisions based on data analysis. Web deletion refers to the collection of data on the web. Web scraping is also known as data scraping. For the purpose of data analysis can be divided into several steps such as cleaning, editing etc. 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 programming 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 and 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, 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 structure 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.

Tensor Flow

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

JavaScript Frameworks

- **NodeJS**

  Node.js is an open source, cross-platform JavaScript runtime environment that runs back to the V8 engine and extracts JavaScript code without the use of a web browser.

  Our rewriting code is written for Nodejs as its fast and platform language.

- **ElectronJS**

  Electron is a framework for developing native apps using web technologies like as JavaScript, HTML, and CSS. As Electron is used to create a short web application it helps us to write our code and thus reduce 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 may also be used to render to the server using Node and to power mobile applications using React Native.

PYTHON

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

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

COMPATIBILITY

OS X

Only 64bit 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, 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**

  Microsoft Visual Studio Code is a freeware source code editor available for Windows, Linux, and MacOS. Debugging assistance, syntax highlighting, intelligent coding, captions, code reuse, and integrated Git are among the features.

- **Google Collab Notebook**

  Collaboratory, or Colab for short, is a Google Research tool that allows developers to write and run Python code in their browser. Google Colab is an amazing tool for hands-on learning activity. It is a little Jupyter notebook that requires no installation and includes a fantastic free edition that gives you free access to Google computer resources like GPUs and TPUs.

- **PyCharm**

  PyCharm is an integrated development environment for computer applications, mostly in the Python programming language. JetBrains, a Czech firm, created it.

Data Source

As we did not get more than 40 diseases so to get dataset we have created our own dataset. And the dataset which we have used for our training and testing process have taken from various sources. One of them is added below.

- https://github.com/DiseaseOntology/HumanDiseaseOntology

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's 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 and extracted from various sources such as the CDC's database database and Kaggle resources. Then analyze the extracted data using texts written in 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.

Further in 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. Although 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 is written with the help of Electron JS.

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 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 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 value 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 the number of data points in the data provided. This cost function is also called 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 in 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 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.

CONCLUSIONS 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 is compiled from factual reports, directly to Government media outlets for local media where it is considered reliable. A team of experts and analysts who validate 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. Also, we are looking to link this website to various hospitals and pharmacies for easy use.

REFERENCES

[1] Thivaharan. S, Srivatsun. G and Sarathambekai. S, "A Survey on Python Libraries Used for Social Media Content Scraping", Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020) IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9, PP: 361–366.

[2] Shreya Upadhyay, Vishal Pant, Shivansh Bhasin and Mahantesh K Pattanshetti "Articulating the Construction of a Web Scraper for Massive Data Extraction", 2017 IEEE.

[3] Amruta Kulkarni, Deepa Kalburgi and Poonam Ghuli, "Design of Predictive Model for Healthcare Assistance Using Voice Recognition", 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2017, PP: 61–64.

[4] Dimitri Dojchinovski, Andrej Ilievski, Marjan Gusev Interactive home healthcare system with integrated voice assistant MIPRO 2019, PP: 284–288 Posted: 2019.

[5] Mohammad Shahnawaz, Prashant Singh, Prabhat Kumar and Dr. Anuradha Konidena, "Grievance Redressal System", International Journal of Data Mining and Big Data, 2020, Vol. 1, No. 1, PP. 1–4.

Hyperlink of Research Paper

1. https://www.sciencedirect.com/science/article/pii/S2352914817302253