WEB SCRAPING USING MACHINE LEARNING

Sakshi Nakate

Student, Department of Information Technology,
Pune University, Baramati, Pune, Maharashtra., India,
sakshi.nakate.it.2019@vpkbiet.org

Shruti Narkhede

Student, Department of Information Technology,
Pune University, Baramati, Pune, Maharashtra., India,
shruti.narkhede.it.2019@vpkbiet.org

Sonali Gawade

Student, Department of Information Technology,
Pune University, Baramati, Pune, Maharashtra., India,
sonali.gawade.it.2019@vpkbiet.org

Abstract - Web scraping is the process of collecting or extracting information from a particular website. It is a technique to convert any unstructured data into structured data and then analyze the obtained data based and is the stored in required format file type. Web scraping is becoming well known due to large amount of data available on internet and want to collect the data without wasting time. Web scarping can be applied to obtain a huge amount of data for better decision making. We can achieve this using BeautifulSoup tool and other algorithms. The obtained data after web scraping will be processed for Text Recognition and Text Classification using NLP and Classification.

Keywords: Web scraping, unstructured data, data-format, text classification, text recognition, NLP, classification

1. Introduction

WEB SCRAPING: Web scraping is the process of extracting data from websites automatically using a software program or script. It involves retrieving data from the HTML source code of a webpage and transforming it into a structured format that can be analyzed or stored for later use Web scraping is the process of extracting data from websites automatically using a software program or script. It involves retrieving data from the HTML source code of a web-page and transforming it into a structured format that can be analyzed or stored for later use. It's important to note that while web scraping can be a powerful tool for data collection, you should always respect the website's terms of service and follow ethical guide lines. Some websites may have restrictions or prohibit scraping their content, so it's crucial to ensure you're acting within legal and ethical boundaries. TEXT RECOGNITION: Text recognition, also known as Optical Character Recognition (OCR), is the technology that enables computers to recognize and extract text from images or scanned documents. It involves converting the text present in an image or document into machine readable and editable text. Text recognition finds applications in a wide range of fields, such as digitizing printed documents, extracting information from invoices or receipts, converting scanned books into editable text, automatic license plate recognition, and more. Many programming languages provide OCR libraries or APIs that simplify the implementation of text recognition in your applications, such as Tesseract OCR for Python or Google Cloud Vision API. TEXT CLASSIFICATION: Text classification is a natural language processing (NLP) task that involves

categorizing or assigning predefined labels or categories to textual data. It aims to automatically classify text documents or snippets into different classes or categories based on their content. Text classification has various applications, including sentiment analysis, spam filtering, topic classification, news categorization, intent detection, and many more. The choice of model and techniques depends on the specific task and the characteristics of the text data you are working with.

2. Literature Survey

1. D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019

The work in this paper is focusing on – Data analysis, Web Scraping, Implementing of Web Scrape which is a web scraping software to scrape e-commerce sites such as Flipkart, Amazon and analyze product details which aren't available, analyze variation, comments, ratings, etc. The point of the paper is to remove the information from different sources with the assistance of programming known as the web crawler Scrapy utilizing the programming language Python adaptation 3.6. The Database is created which collects all the unstructured data from various sources and then analyses them by the analytic process of its specifications, assembling, organizing, cleaning, reanalyzing, applying models and algorithms and finally providing the desired results. In this paper Reddit by XPath method was used to find details of each element of the frequent searches. Main outcome was user friendly search interface, indexing, query processing and effective data extraction based on web structure.

2. M. S. Parvez, K. S. A. Tasneem, S. S. Rajendra, and K. R. Bodke, "Analysis Of Different Web Data Extraction Techniques," 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, 2018.

The work in this paper is focusing on - Web scraping, Data Extraction, Web crawler, Machin Learning Approach systems -WIEN, WHISK, Rapier. Web data Scraping process includes Web Crawler and Data Extractor. Data Extraction Techniques involved are – a) Human Copy and Paste, b) HTML Parser-JAVA library – Jsoup and Python library – BeautifulSoup. c) HTTP Programming, d) Tree based technique i.e DOM (Document Object Model).The techniques that use DOM are i) Addressing element in the document tree(XPath), ii) Tree edit distance matching algorithms., e)Web Scrap per includes 3 approaches - Regular expression based approach, Logic based approach, ML approaches. Different ML approach specifies are Statistical ML approach, Adaptive Search, WIEN, RAPIER, WHISK, SRV. WHISK system is the most advantageous as compared to others according to the survey.

3. M.S. Akopyan, O.V. Belyaeva, T.P. Plechov, D.Y. Turdakov (2019). Text Recognition on Images from Social Media.

The work in this paper is text recognition, social networks, image processing, deep neural networks. Text recognition pipeline is provided to address text extraction from various quality images collected from social media. Input images are categorized into different classes and then class specific preprocessing is applied to them for illumination improvement, text localization. Then OCR engine is used to recognize text. The results are experiments of dataset collected from social media. For Image preprocessing. Image Resolution Enhancement (IRE) is used before applying OCR engine. They are based on Deep Neural Networks and use general Adversial Networks with Sub-Pixel

Convolutional Images. The dataset used contains 64554 symbols in 67 images. Images are divided into 4 categories – Demotivators, Certificates, Scanned, Smartphones. To solve classification problems in this paper Neural Network (ResNet50, MobileNet) and gradient boosting (GB) approaches are used. The rounded stamp on documents are determined using Hough Circle Search Algorithm.

4. Y. Su, H. Peng, K. Huang and C. Yang, "Image processing technology for text recognition," 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), Kaohsiung, Taiwan, 2019.

The work in this paper is - Image processing, optical character recognition, object detection component. This paper demonstrates how image processing technology can be used in combination with OCR to improve recognition accuracy and to improve the efficiency of extracting text from images. Two software systems are developed and tested – i) A character recognition system applied to cosmetic-related advertising images which include the process as a) Image processing, b) Establishing contours and lassosing the region of interest(ROI), c) Character Recognition. The techniques in preprocessing were edge detection, binarization and erosion and dilation. Edge detection algorithm used was Sobel Edge detection with Tesseract - OCR tool used in combination with python. ii) A text detection and recognition system for natural scenes which include the process as a) Operating the Raspberry Pi camera and detecting the target object containing text, b) Image processing. In this the Cascade classifier was used which was given the image trained set downloaded from ImageNET. Advertising images and images from ICDAR Robust Reading Competition were used as test images for this result study.

5. W. Zhuo and C. Lili, "The algorithm of text classification based on rough set and support vector machine," 2010 2nd International Conference on Fu ture Computer and Communication, Wuhan, 2010.

The work demonstrates - rough set, support vector machine, classification. It rep resents a new algorithm of text classification based on Rough Set and Support Vector Machine. As SVM is a tool for solving the problem of ML based on optimization method. It has a simple structure and good classification ability but its processing speed is slow when we deal with large amount of data. To overcome this bottleneck problem of SVM, Theory of Rough Set was introduced. Theory of Rough Set is a math tool of quantitative analysis which could analyze correlations between the information, which needn't any prior knowledge and it has a powerful foundation to process information of high capacity and dimensions. The experiment used Rossta software to process the initial training set data. The SVM aims to construct objective function which could make a distinction between two patterns of modes as far as possible and give consideration to maximize of interval of classification and minimize the error. The key of RS-SVM algorithm is how to delete attributes which uncorrelated and unimportant by the algorithm of attribute reduction and decrease the dimensions of SVM training.

3. Proposed Work

3.1 System Architecture

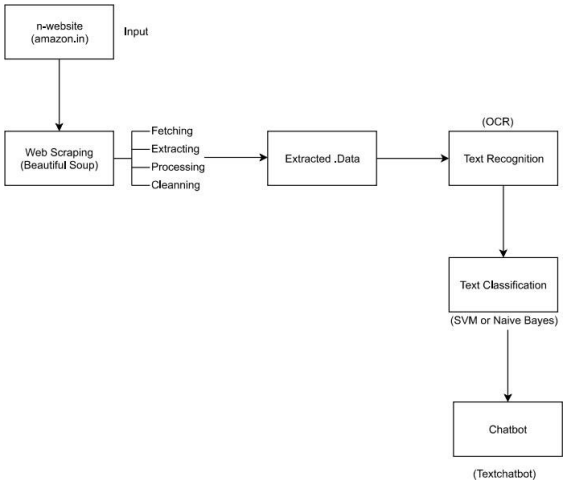

Fig – 1: Proposed System Architecture

4. Conclusion

The proposed extraction model is capable of extracting the data and storing data in required format. The stored or saved data can be used for obtaining confidential data It contains modules as web scraping, storing, text recognition and text classification. In this first, data is extracted and stored in required format. Further it is given for text recognition and text classification and finally the overall saved data format is provided to the server using chatbot as platform.

References

[1] D. M. Thomas and S. Mathur, ”Data Analysis by Web Scraping using Python,” 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019.

[2] M. S. Parvez, K. S. A. Tasneem, S. S. Rajendra, and K. R. Bodke, ”Analysis Of Dif ferent Web Data Extraction Techniques,” 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, 2018.

[3] M.S. Akopyan, O.V. Belyaeva, T.P. Plechov, D.Y. Turdakov (2019). Text Recogni tion on Images from Social Media.

[4] Y. Su, H. Peng, K. Huang and C. Yang, ”Image processing technology for text recognition,” 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), Kaohsiung, Taiwan, 2019.

[5] W. Zhuo and C. Lili, ”The algorithm of text classification based on rough set and support vector machine,” 2010 2nd International Conference on Future Computer and Communication, Wuhan, 2010.