International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

Intelligence Extraction Using Machine Learning Technics

Prof. Harish Patil[1], Varun Gaikwad[2], Dipali Pawar[3], Mayuri Nikam[4]

[1]Asst. Professor, Dept. of Computer, ISB&M School of Technology, Pune, Maharashtra, India
[234]Student, Bachelor of Engineering, Dept. of Computer Engineering, ISB&M School of Technology, Pune, Maharashtra, India

---

Abstract - *Intelligence Extraction or IE is technique of arranging unstructured information or data in a proper systematic manner by using machine learning algorithm. Structure information is a sorted information which can be easily understood and classify by human brain. Unstructured data as name suggest is an unstructured data format meaning dynamic format information which cannot be understand by machine or human. Hence, extracting meaningful information from them is not an easy task.*

Key Words: Intelligence Extraction, Structure Data, Unstructured Data.

1.INTRODUCTION

Every day we process large amount of data so processing and analyzing this data which is unstructured is complex task. Millions of Documents is uploaded every day on cloud and to handle those data we require system which is easy to handle, reliable, efficient, user friendly and through which we can get structured data.

World Wide Web is a central location in which data is stored and managed, so this organization which contain huge amount of information which is in the form of pdf, images, text, number, videos etc. from this huge data user wants only relevant data.

2. INTELLIGENCE EXTRACTION

Intelligence Extraction is nothing but extracting structured data. Intelligence Extraction contain Webpage Extraction, Csv Extraction, Video Extraction, Image Extraction, Pdf Extraction etc.

All these module extract data and store that data in some file format such as .csv, .txt etc. so it can be used by anyone. For example- Such kind of data is used in Crime Investigation Department

3. INTELLIGENCE EXTRACTION MODULES

Proposed system contains following modules-

1.CSV Extraction

2.Web-Page Extraction

a. Text Extraction

b. Image Extraction

c. E-mail Address Extraction

d. URL Extraction

e. Table Extraction

3.Video Extraction

4.Image Extraction

5.PDF's Extraction

3.1 CSV Extraction

This module is used for extracting the csv data from a particular column using a special character called as delimiter, Delimiter are those special character which separates data. This module uses 2 libraries:

a) Pandas
b) CSV

In this module the user has to provide the name of the file with extension to the program, module reads that file and asks the column that user wants to extract, after giving column name the user will provide the delimiter which is present between data of that column.

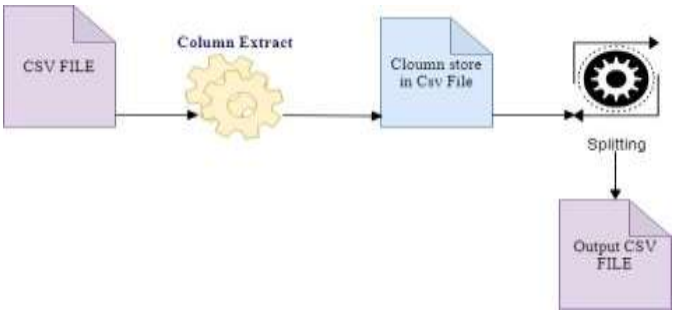

Figure 1: CSV Extraction Process

3.2 Web-Page Extraction

Web Extraction extract webpage contents and store it in file. This module is categorized in many forms such as-:

3.2.1 Text Extraction from web-page

This will extract text data from a static website. This uses 3 Libraries:

a) Request
b) Sys
c) Beautiful Soup

3.2.2 Image Extraction from web-page

This module is a part of Email Extraction, the difference in this is that it will extract image from given URL. There are 4 Libraries used:

a) Request
b) Urllib
c) Beautiful Soup
d) Re

3.2.3 E-mail Extraction from web-page

Email Extraction module is used for extracting the email address from a particular website provided by the user. We used 3 libraries here:

a) Re
b) Beautiful Soup
c) Request

3.2.4 URL Extraction from web-page

This module focuses on extracting linked URL of a particular Website. The user will provide a URL to this module and it will read that website and return linked URL from it. There are 4 Libraries used here:

a) Re
b) Sys

3.2.5 Table Extraction from web-page

This module extract table from web page and store it in csv file format.

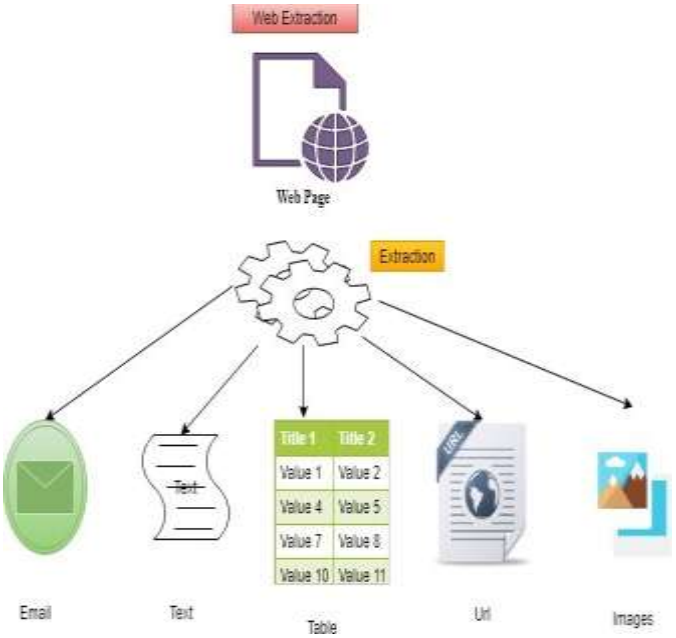

Figure 2: Web Extraction Process

3.3 Video Extraction

Video frame extraction will extract each and every frame of a particular video and store it in a file, the frame will be in an image format. This module used 2 Libraries:

a) CV2
b) Os

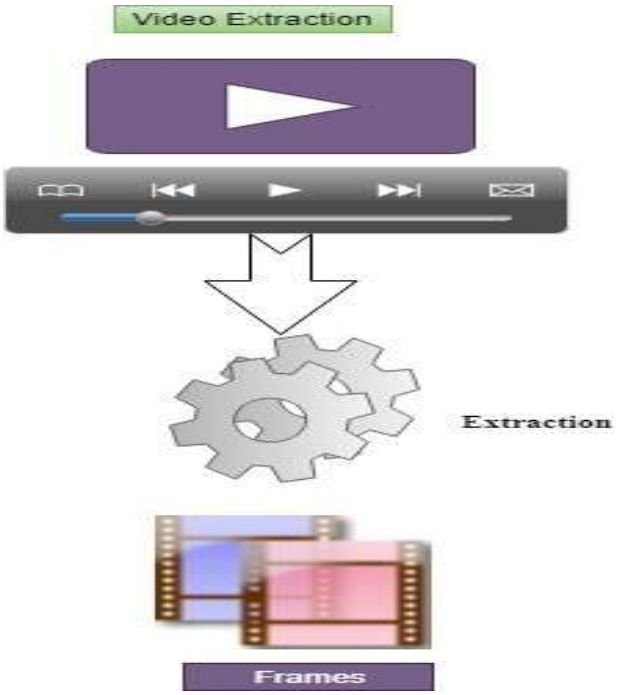

Figure 3: Video Extraction Process

3.4 Image Extraction

This module is used for extraction of text from image as a source. This uses 2 Libraries:

a) Pytesseract
b) PIL

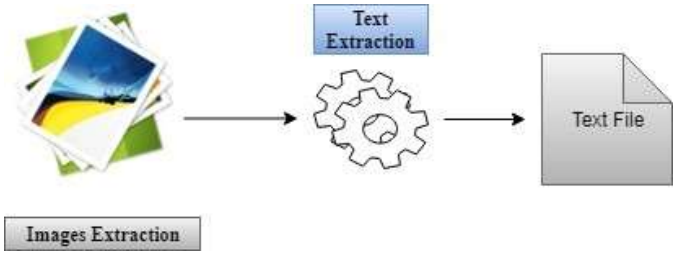

Figure 4: Image Extraction Process

3.5 PDF's Extraction

It will extract text from the pdf file format. We have used only one library here and that is PyPDF2.

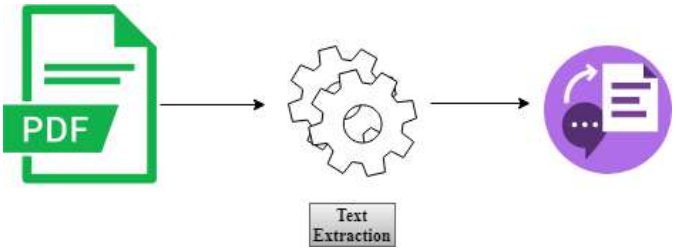

Figure 5: PDF's Extraction Process

4. Results and Discussion

This system work on data collected from various users such as company, organization data etc. As user wants only relevant data so our proposed system categories that data so it can be easily accessible by user. We use different types of machine learning algorithms for extraction and categorizing that data into graph format. After we store that data into database or cloud for future use.

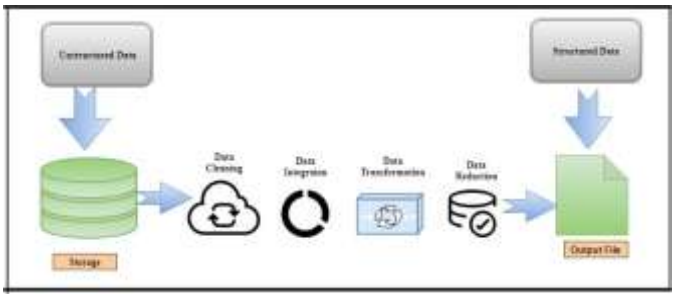

Figure 6: Working of Intelligence Extraction

5. CONCLUSIONS

We implement a system in which extraction is based on the different machine learning algorithms which sort an unstructured data into structured format so it may be user friendly for user.

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us the possibility to complete this paper. A special thanks we give to our project guide Prof. Harish Patil and our HOD of computer department Dr. Pallavi Jha whose contribution in suggestions and encouragement and helped us to coordinate in our project mainly in writing this paper.

REFERENCES

[1] Vidya V L, "A Survey of Web Data Extraction Techniques", International Journal of advance research in computer science and management studies, vol. 2, Issue 9, Sep. 2014.

[2] Information Extraction on Novel Text using Machine Learning and Rule-based System, Ria Chaniago School of Electrical Engineering and Informatics Bandung Institute of Technology Bandung, Indonesia.

[3] 2018 12th IEEE International Conference on Semantic Computing, Data Acquisition and Information Extraction for Scientific Knowledge Base Building Piotr Andruszkiewicz Institute of Computer Science Warsaw University of Technology Warsaw, Poland.S.S.Bhamare, Dr. B.V.Pawar" Survey on Web Page Noise Cleaning for Web Mining" International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013.

[4] Yanhong Zhai, Bing Liu," Web Data Extraction Based on Partial tree alignment", ACM 1-59593-046- 9/05/0005.

[5] H.L. You, W. Zhang, J.Y. Shen, and T. Liu, "A Weighted Voting Based Automatic Term Recognition Method," Journal of Chinese Information Processing, 2011, pp. 9-16

[6] L.L. Earl, ĀExperiments in automatic extracting and indexing,ā Information Storage and Retrieval, 1970, pp. 313-330.

[7] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic Recognition of Multi-Word Terms: The C-value/NC- value Method," International Journal on Digital Libraries, 2000, pp. 117~132.

[8] D.F. Zhai and B.S. Liu, "Automatic Domain – specific Term Extraction in Administrative – domain ontology," Data Analysis and Knowledge Discovery, 2010, pp. 59- 65.

[9] Z.Y. Fu, Information Theory: Fundamental Theory and Applications. Beijing: Electronic Industry Press Pub, 2007.