www.revistageintec.net
ISSN: 2237-0722

revista GEINTEC
Gestão, Inovação e Tecnologias

Conversion of Unstructured to Structured: A Solution Using Data Science and NOSQL

Shagufta Praveen[1]; Mazhar Afzal[2]; Aamir Khan[3]

[1]Computer Science Department, Glocal University, India. E-mail: shaguftasaeed125@gmail.com

[2]Computer Science Department, Glocal University, India.

[3]Computer Science Department, Glocal University, India.

Abstract

Exponentially rise of unstructured data is a question to all data scientists today. The graph of unstructured data is at such height that it consumes most of the storage of all clouds present today. Analysis through unstructured data is not easy due its high complexity and abstraction computations. There are different forms of unstructured data found inside web and different operations are done for their conversion. In this paper we tried to touch most of the unstructured data types and their related solution for specific conversion using data science analytics. Mongodb, famous for horizontal scaling is also used here for collection and storage of scalable and high volume data.

Key-words: Unstructured Data, Python, R and Mongodb.

1. Introduction

The fourth step of data mining process is “Data transformations [1]” where data is transformed from one form to another for accurate outcomes which are transformed data based. Data conversion and transformation are one of the vital processes for data analysis. For better data pattern evaluations and visualization these transformed and converted data is widely used during operations. There are different types of data is present which is collectively called as Big Data [2] now. Before the concept of big data, data was categorized into transactional data, multidimensional data, spatial data, and audio and sensor data.

With revolving earth there are innumerable revolution occurred in our data market too. Now data is categorized into three basic types: Structured, Unstructured and Semi-Structured [3]. Arrival of unstructured data (audio, videos and sensor data) is much faster than other data presently. The

Received: 24.05.2021 – Accepted: 21.06.2021

cause behind this arrival is social media, web based applications and web based upcoming technologies. As per data scientists unstructured data is like a raw material to them that needed to be processed to convert them into particular shape and result.

2. Related Work Done

In 2010, Tao peng et al. published his work for the research of unstructured data transformation using XML[4]. In 2010, authors like Justin Langseth, Nithi Vivatrat, Gene Sohn, worked to perform methods that will available unstructured data for structured data analysis [5]. In 2013, Octavian Rusu et al. published a paper that convert unstructured and structured data into knowledge[6]. In 2018, kuldeep smabrekar et al. tried to convert unstructured agro-data to semi-structured or structured using Cassandra couchbase tools [7].

3. Data Science Programming and Data Conversion

A data scientist extracts knowledge from set of data through fundamental principles provided by data science [8]. Importantly procedures and methods that work in a system to treat a problem in respect of data are performed by data science. Multiple predictive and descriptive analyses are going on throughout the world by data science concept. The way out for its implementation is done by some data science programming languages that help in data extraction through scraping [9]. Not only scraping but combination of these languages with ML gives amazing outcomes for multiple fields. Among multiple data science languages R and python are widely used in all data based organizations and MNCs.

Python and R not only provide better data extraction. They also contribute in data conversion a lot. Most of the data scientists know that data conversion is one of the biggest and complex challenge present in data science world.

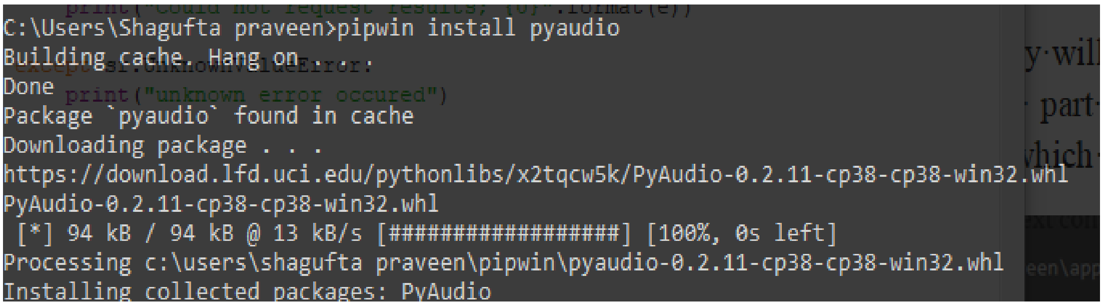
3.1. Speech to Text Conversion

Python has a library name Speech recognition and pyaudio[10][11]. These libraries will help the speech to take text as input and convert it into a text. This is also a famous part of NLP[12](Natural language processing).It can deal with different number of languages which is the best

Received: 24.05.2021 – Accepted: 21.06.2021

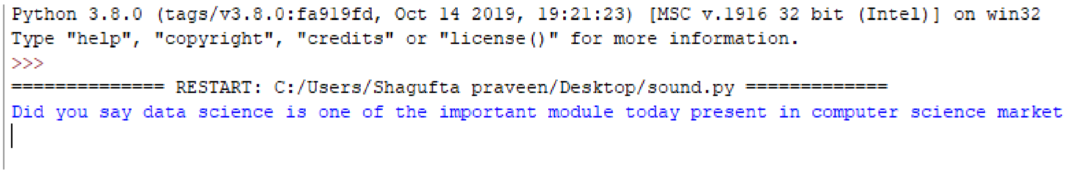
part. Most of the lectures in various fields can be converted to a text with the help of speech to text conversion.

Fig. 1 - Installing library



```
C:\Users\Shagufta praveen>pipwin install pyaudio
Building cache. Hang on . . .
Done
Package `pyaudio` found in cache
Downloading package . . .
https://download.lfd.uci.edu/pythonlibs/x2tqcw5k/PyAudio-0.2.11-cp38-cp38-win32.whl
PyAudio-0.2.11-cp38-cp38-win32.whl
 [*] 94 kB / 94 kB @ 13 kB/s [##################] [100%, 0s left]
Processing c:\users\shagufta praveen\pipwin\pyaudio-0.2.11-cp38-cp38-win32.whl
Installing collected packages: PyAudio
```

Fig. 2 - Converting speech into text



```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:21:23) [MSC v.1916 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
============== RESTART: C:/Users/Shagufta praveen/Desktop/sound.py =============
Did you say data science is one of the important module today present in computer science market
```

3.2.Image to Text

3.2.1. Extraction of Text from Image

Text written over an image can be extracted through library of python using Python Image library and using image to string function. Library used here for data transformation is pytesseract [13].Python move image into this module and finally do image to text conversion with the help of this module.

3.2.2. Recognizing Images

Tensor Flow[14] and keras[15] like library in python made a drastic change in python world. Many of the images or images of collection of different object can be recognized with the help of already prepared data consist of millions of images and their related information. Not only can this many of the videos (moving frames) be converted into structured data by telling about their activities through text

ISSN: 2237-0722
Vol. 11 No. 4 (2021)
Received: 24.05.2021 – Accepted: 21.06.2021

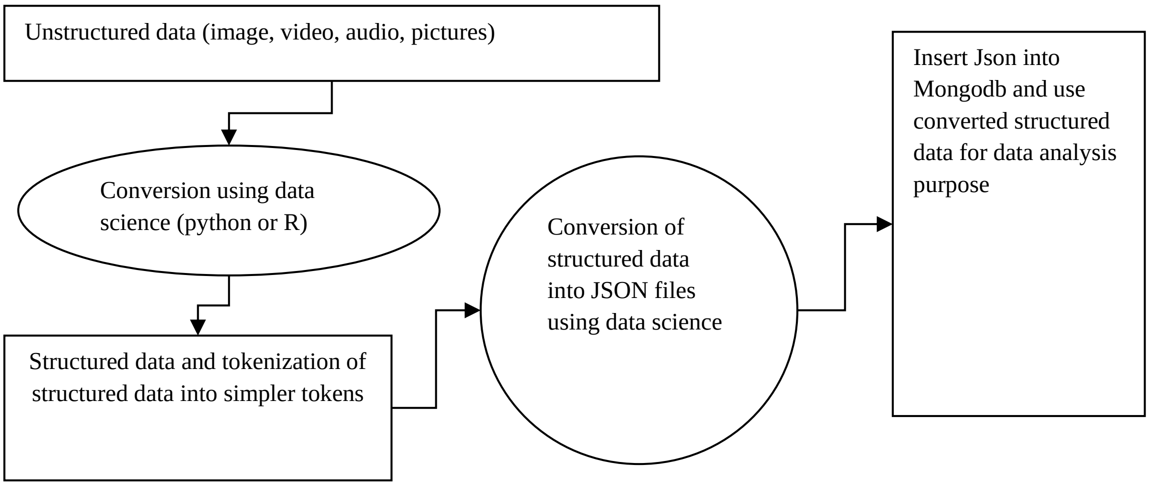
3.3.Extraction of Data and Tables from Web Pages

Web pages are always the best source for data collection and there are multiple web sites that update data related to weather and crypto-currency for data analysis and data prediction. Most of the data i.e. images, text and pictures can be extracted from web pages through web scrapping which is accomplished by beautiful Soup[15] library of python.

4. Using Unstructured Data for Analytical Purpose through Mongodb after Conversion

Unstructured data conversion is a task but another task is how make it relatively useful than before for analysis purpose and there is a process to do so by using NoSQL database and storing this converted data in Mongodb. Mongodb is one of the famous and most horizontal scalable data used world widely. As it is already known that unstructured data is rising like anything then mongodb and that will be the best option to store such large data after conversion. Even by using python a text file (converted structured) can be converted to json then eventually inserted into Mongodb for different analysis.

Fig. 3 - Conversion of unstructured data into structured using data science and NOSQL for data analysis purpose


Unstructured data (image, video, audio, pictures)
Conversion using data science (python or R)
Structured data and tokenization of structured data into simpler tokens
Conversion of structured data into JSON files using data science
Insert Json into Mongodb and use converted structured data for data analysis purpose

5. Algorithm

Steps Using Python

1. Import library tkinter# for UI
2. Import library speech_recog//for sound

Received: 24.05.2021 – Accepted: 21.06.2021

3. Import pyaudio,sys,os,pyttsx3
4. Create window dialog using
    a. Label command =(“attribute name”,”attribute-color”,”background-color”)
    b. Label command(dimensions of window)
5. Define function on click
    a. Call(sound()) //call sound function

sound() //sound function

1. Import libraries pyaudio,os,urllib
2. Get audio source using microphone()
3. Use audio () to listen data
4. Write data on my_text
5. Run application notepad
6. Print data on the notepad

Steps using R

1. Import library tokenizer
2. Assign data of the notepad to variable
3. Use tokenize function to tokenize the variable
4. Convert tokens to JSON file

Mongodb Application

- Open cmd
- Give path
- Make connection, run mongodb shell
- Connect using mongodb compass
- Insert json file as table into mondodb
- Start doing analysis

6. Conclusion

Unstructured data was always a challenge for data scientist. It is really hard to use unstructured data for analysis purpose. In this research article, unstructured data is converted to structured data with the help of data science language and then after conversion, text is converted to json file so that it could be use for analysis in Mongodb. Mongodb is used here as it belongs to family of NoSQL and can store a large excessive data easily. This way data science and NoSQL technologies could give various numbers of solutions for different challenges.

References

Han, J., Kamber M., and Pei, J., (2011). *Data mining concept and techniques,* The Morgan Kaufmann series in data management systems, Morgan Kaufmann publishers.

Jain, V.K. (2017) *Big Data and Hadoop,* Khanna Publishing, 600 pages.

Praveen, S., Chandra U., Influence of structured (2017), Semi-structured and unstructured data on various data models. *International journal of scientific & engineering research* volume 8, issue 12.

T. Peng, L. Sun and H. Bao, "Research of Unstructured Data Transformation Based on XML," 2010 *International Conference on Internet Technology and Applications,* Wuhan, China, 2010, pp. 1-4, doi: 10.1109/ITAPP.2010.5566508.

Justin Langseth, Nithi Vivatrat, and Gene Sohn. “Analysis and transformation tools for structured and unstructured data”. January 11, 2007, US20070011183 A1.

Rusu, Octavian, et al. et al. "Converting unstructured and semistructured data into knowledge." *Roedunet International Conference (RoEduNet),* 2013 11th. IEEE, 2013.

K. Sambrekar, V. S. Rajpurohit and J. Joshi, "A Proposed Technique for Conversion of Unstructured Agro-Data to Semi-Structured or Structured Data," 2018 *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA),* Pune, India, 2018, pp. 1-5.

Provost F and Fawcett T. (March 2013)” *Data science and its relationship to big data and data-driven decision making”,* Mary Ann Liebert, Inc, New York, California.

Mitchell R. (2018)” *Web scraping with python: collecting data from the modern web”,* O'Reilly Media, Inc.

https://pypi.org/project/SpeechRecognition/ [17 April,2021]

https://pypi.org/project/PyAudio/[17 April,2021]

https://pypi.org/project/indic-nlp-library/[17 April,2021]

https://pypi.org/project/pytesseract/[17 April,2021]

Install TensorFlow with pip[17 April,2021]

https://pypi.org/project/keras/[17 April,2021]

https://pypi.org/project/beautifulsoup4/[17 April,2021]

Received: 24.05.2021 – Accepted: 21.06.2021