Application of Python Programming Language in PDF-Text Based Information Extraction

CHINEMEREM BENITA A[1], NJOKU DONATUS O[2], TAIWO AHMED O.[3]
[1, 2] *Department of Computer Science, Federal University of Technology, Owerri-Imo State, Nigeria*
[3] *Pladis Global, The Polytechnic, Ibadan, Oyo State, Nigeria*

Abstract- ***Information Extraction has become a vital aspect of research over the decades which allow millions of researchers have access to only what seems important to them, amidst the enormous pieces of information around. The motivation of carrying this work out was to mitigate on the time factor that faces every researcher with regards to meeting stipulated deadlines. After a careful literature review on the existing studies it was gathered that a proposed system framework was developed to match keywords during the word extraction formation. The research followed a System Structured Analysis Design Methodology (SSADM), which utilizes the powerful libraries of python programming to mine text data known as keywords from a structured file and parse through these binary data. The extracted input data are encrypted via parallel encryption. The system was tested and the sample results achieved.***

Indexed Terms- ***Information, Extraction, Text data, Binary data, Encryption, Python***

I. INTRODUCTION

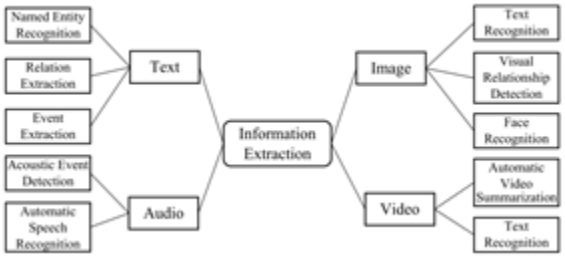
Information Extraction (IE) is the field that primarily deals with text structuring. IE work automates the recognition of interesting information related to pre-specified types of events, entities or relationships in text from sources such as newswire articles and the Web. IE systems extract information that are not structured from unstructured automatically. In other to do this, most IE systems rely on a set of patterns extractions. Extraction patterns are defined based on the syntactic and semantic constraints on the positions of desired entities within natural language sentences [14]. Information Extraction (IE) is the name now being applied to a process analogous to reading a document and ling out a form which captures the crucial information. Specifically, IE requires the structuring of information from texts into a record format suitable for creating databases, or for use in other computer-based applications. [1], in a survey of Unsupervised Approaches for Textual Semantic Annotation, projected that successful automatic extraction depends on limiting the subject domain being considered and limiting the scope of the information being extracted. The grammars used for this task are normally limited in their coverage and hence parsing techniques must be robust. Another complementary approach is that of natural language processing (NLP) which has solved the problem of modelling human language processing with considerable success when taking into account the magnitude of the task. In terms of both difficulty and emphasis, IE deals with tasks in between both IR and NLP. IN [9] research shows that numerous parsers have been proposed by the Computational Linguistics community. These parsers can generally be divided into two broad categories based on their underlying grammatical formalism: Constituency parsers and dependency parsers. Parsers (also known as tree bank parsers) produce syntactic analysis in the form of a tree that shows the phrases comprising the sentence and the hierarchy in which these phrases are associated. Constituency parsers have been used for pronoun resolution, labeling phrases with semantic roles and assignment of functional category tags.

Relation Extraction is the task of detecting and classifying pre-defined relationships between entities identified in the text. In other words, it is way of transforming unstructured (free) text into structural form which can be used in web-search, question answering and lot more [10]. Part-of-Speech tagging is fundamental step in various NLP tasks such as speech recognition, speech synthesis, machine translation, information retrieval, information extraction, and so on. Two factors that determine the tag of word are its lexical probability and its contextual probability. Part-of-Speech tagging approaches can

generally fall into two categories: Rule based approaches and statistical approaches. Rule based approaches apply language rules to improve the accuracy of tagging.

II. RELATED LITERATURE

Holding strong to the concept of [14] relates that a more specific goal is to allow logical reasoning to draw inferences based on the logical content of the input data structured data is semantically well-defined data from a chosen target domain, interpreted with respect to category and context. It is often, according to [13] an early stage in pipeline for various high-level tasks such as Question Answering Systems, Machine Translation, event extraction, user profile extraction, and so on. Constituency parsers overlook functional tags when training. Therefore, they cannot use them when labeling unseen text. Dependency parsers analyze the sentence as a set of pair wise word-to-word dependencies. As per [10], various low-level tasks in NLP such as Parts-of-Speech Tagging (POST). the effectiveness of these low-level tasks highly determine the performance of high end tasks Error in low level tasks get propagated to high level tasks and degrading their overall performance.



Classification of IE Subtask [13]

There are lots of many processes involved in the pipeline of NLP. At the syntactic level, statements are segmented into words, punctuation (i.e. tokens) and each token is assigned with its label in the form of noun, verb, adjective, adverb and so on (Part of Speech Tagging). At the semantic level, each word is analyzed to get the meaningful representation of the sentence [9]. Noun Phrase Recognizer finds the noun phrases from the text. For e.g. "the president of USA", the president is noun phrase and it refers to a person, whereas USA represents noun phrase and refers to name of the country. Named Entity Recognizer, finally assigns particular named entity class from various classes such as: person, organization, location, date, time, money, percent, e-mail address and web-address. As Part-of-Speech tagging and Syntactic Parsing forms the building block and initial phase in the pipeline of various Information Extraction tasks, it is important to look at their role, state-of-the-art systems and how they affect downstream IE tasks.

A. Named Entity Recognition (NER)

In the Named Entity Recognition (NER) task, systems are required to recognize the Named Entities occurring in the text. More specifically, the task is to find Person (PER), Organization (ORG), Location (LOC) and Geo Political Entities (GPE). For instance, according to[9], in the statement "Michael Jordan lives in United States", NER system extracts Michael Jordan which refers to name of the person and United States which refers to name of the country. NER serves as the basis for various crucial areas in Information Management, such as Semantic Annotation, Question Answering, Ontology Population and Opinion Mining

B. Temporal Information Extraction (Event Extraction)

Temporal information extraction or event extraction refers to the task of identifying events (le information which can be ordered in a temporal order) in free text and deriving detailed and structured information about them, ideally identifying who did what to whom, where, when and why. Hence, [11] temporal expression (also called timex) refers to the task of detecting phrases in natural language text that denote a unit of temporal entity in the form of an interval, a particular instance of time or certain frequency related to particular event.

According to [4] surveyed literature and some limitation in the issue of information extraction from research articles using data mining techniques the research observed the synergy between information extraction and data mining techniques that helps to discover different interesting text patterns in the retrieved articles. Text mining is different from data mining as stated in [3] data mining is focused on discovering interesting patterns from large databases rather than textual information. In [2] Information recovery methodologies like text indexing techniques have been developed for handling unstructured

documents. In conventional researches, it is assumed that a user mostly searches for known terms, which have been previously used or written by someone else. The main problem is that the search results are not relevant to the user's requirements.

C. Text Mining Processing Framework

The research carried by [5] developed a customized framework as shown in Fig 2. The work proposed three steps which include in text mining: text pre-processing, text mining operations, and post processing. Text pre-processing involves the following tasks: data selection, classification, feature extraction and text.

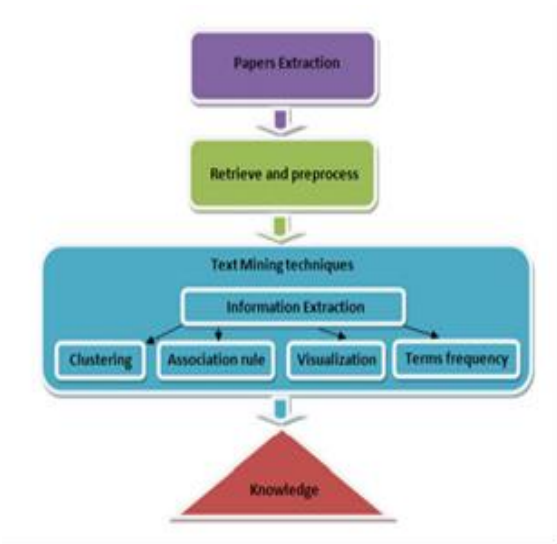


Fig. 2: Text mining processing framework [5]

Normalization involves the process of transforming the documents into an intermediate form for ensuring compatibility for various mining tools. The second step deals with different text mining techniques like clustering, association rule detection, visualization, and terms frequency. During the third step, alterations and changes are made on the data (i.e., research articles) through text mining functions like evaluation and choice of knowledge, analysis and visualization of knowledge. The main aim of this study is to extract interesting information from the collected articles using the text mining techniques [5]

In the research done by [6] provides detailed view on Text recognition to extract the data from bills which are either handwritten or printed and updated to the database automatically. Their research gives fast and accurate way of yielding text from the bill. The research proposes deep learning techniques for text detection and extraction where we use EAST algorithm to analyse the letter and word from image or scanned document into machine readable form.

Optical Character Recognition (OCR) OCR can read text from scanned document a image. It can convert all formats of images containing text in handwritten or printed. OCR operates in 2 steps such as text detection and text recognition. OCR is the technology which is used to convert the scanned image, paper document images captured by digital camera and extract the text from images which automatically updated in the database. The image is processed using OpenCV libraries. The work done in [6] applying filters and image classification is pretty straightforward but contour matching is a very difficult problem and requires a lot of manual effort and is not generalized. Next is the Deep Learning method. Deep Learning is one of the best methods. One of the most popular approaches for text detection is EAST (Efficient and Accurate Scene Text Detector). It is a powerful algorithm for text detection. The EAST is made up of networks like U-Net, which is superior for detecting features from images of different formats and size.

A. Web Page Analysis and Information Extraction

As previously argued, obtaining the necessary information for discovering or invoking the vast majority of Web APIs nowadays requires interpreting highly heterogeneous HTML pages providing documentation for developers. Although, to the best of our knowledge, no other approaches to automating the extraction of information from Web APIS documentation have been devised so far, considerable effort has been devoted to extracting information from Web pages which is relevant to this work.

Tag-based Segmentation approaches [12] are used to analyze the DOM tree of Web page and automatically divide it into subtrees in order to eventually extract information. The essence of this approach relies on the observation that useful information is usually wrapped into so-called important blocks. These techniques usually rely on specific HTML tags, e.g. <table>, as block separators. This approach is, however, not performant when dealing with heterogeneous cases

where various tags are used as separators. To address this, proposals like contemplate a wider range of tags e.g., <tr>, <hr>

Template-based Segmentation techniques rely on the fact that, for repetitive information Web pages often use a recurring structure to capture information, these approaches exploit templates provided either by humans or derived automatically by machine learning techniques in order to extract information from the Web pages. Although, these techniques are performant when dealing with Web pages that share (at least partly) a common structure, in cases where the heterogeneity of the pages is very high the performance is considerably affected. As we shall see, although it is possible to exploit local patterns within a Web API documentation, currently diverse Web APIS use highly diverging structured which prevents us from applying successfully this approach.

First and foremost, more often than not, these registries contain out of date information or even provide incorrect links to APIS documentation pages. Indeed, the manual nature of the data acquisition in APIs registries aggravates these problems as new APIs appear, disappear or change. Secondly, the fact that the data listed is often not that accurate and [14] rather coarse-grained hampers significantly the development of advanced search functionality since automated algorithms produces non relevant information. Therefore, despite the increasing relevance of Web APIs, there is hardly any system available nowadays that is able to adequately support their discovery.

In the research done by [6] provides detailed view on Text recognition to extract the data from bills which are either handwritten or printed and updated to the database automatically. Their research gives fast and accurate way of yielding text from the bill. The research proposes deep learning techniques for text detection and extraction where we use EAST algorithm to analyse the letter and word from image or scanned document into machine readable form.

Optical Character Recognition (OCR) OCR can read text from scanned document a image. It can convert all formats of images containing text in handwritten or printed. OCR operates in 2 steps such as text detection and text recognition. OCR is the technology which is used to convert the scanned image, paper document images captured by digital camera and extract the text from images which automatically updated in the database. The image is processed using OpenCV libraries. The work done in [6] applying filters and image classification is pretty straightforward but contour matching is a very difficult problem and requires a lot of manual effort and is not generalized. Next is the Deep Learning method. Deep Learning is one of the best methods. One of the most popular approaches for text detection is EAST (Efficient and Accurate Scene Text Detector). It is a powerful algorithm for text detection. The EAST is made up of networks like U-Net, which is superior for detecting features from images of different formats and size.

In the work done by [7] stated different varieties of approaches to text information extraction (TIE)from images. The proposed specific applications including page segmentation, address block location, license plate location, and content based image indexing. Fig. 3

Fig. 3: Multi-color document images [7]

Text in images can exhibit many variations with respect to the properties like geometry, color, motion, edge and compression [7]. The problem of Text Information Extraction TIE system receives an input in the form of a still image or a sequence of images. The images can be in gray scale or color, compressed or un-compressed, and the text in the images may or may not move. The TIE problem can be divided into the following sub-problems: (i) detection, (ii) localization, (iii) tracking, (iv)extraction and enhancement (v) Optical Character recognition (OCR)

- Python Programming

Python Anywhere is an online Integrated Development Environment (IDE) and Web hosting service based on the Python programming provides in

browser access to server-based Python and Bash Command-line interfaces, along with a code editor with Syntax highlighting. One striking different between Python Anywhere and the usual Python Cloud Computing solution that we know of, is that you can totally work on it online using internet browser in developing your Python application

III. METHODOLOGY

This research adopted Rapid Application Development (RAD), also known as Rapid Application Building (RAB), is both a common term used to refer to innovative software development methods, as well as the name for the rapid development approach of Terry Barraclough. RAD software development strategies generally place less focus on scheduling and more emphasis on an adaptive system. In contrast to or sometimes even instead of design specifications, models are often used.

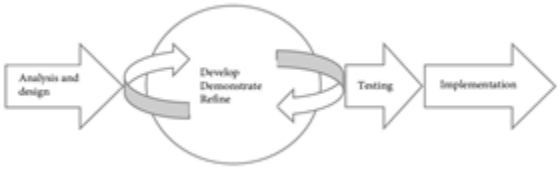


Fig. 4.: Model of RAD

The logical data modeling outcome is a data model with entities that describes the extraction, transforming and Load (record information about), attributes (entity facts) and relationships (entity-to-entity associations) as shown in Fig. 5.

Fig. 5: Entity-Entity- Relationship

Fig. 6 presented information flow modeling on this paper explores processes (activities transferring information from one type to another), data storage (data keeping areas), external entities (sending data to a system or receiving data from a system) and data flows (routes through which data may flow.

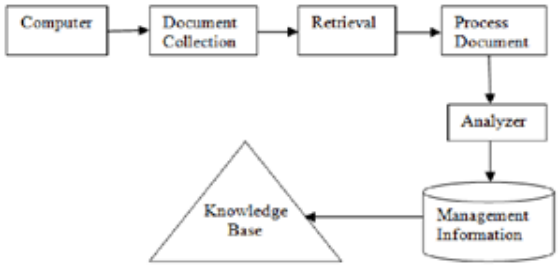


Fig. 6: Data flow Diagram of Information Extraction

The process methods of data gathering were to enable proper understands the performance efficiency of using python in word extraction. The facts behind this choice is the elaborate community python has built overtime which gave rise to thousands of extensions and libraries that makes working with python much easier in this research work

- Analysis

Considering system problems, and decomposition of the large word files to be extracted into its components. A closer look at the elite work done by [8] as contained in Fig. 7 was deeply dependent upon Textual Semantic Annotation and so poses a serious negative result. To solve this [8] focus on reducing their subject domain.

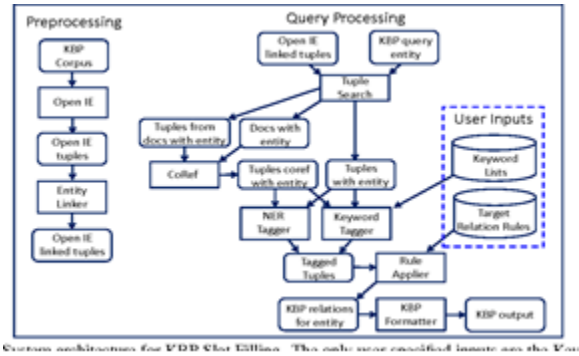


Fig. 7: High Level Model of the Proposed System [8]

IV. IMPLEMENTATION AND RESULTS

i. System Algorithm

This paper proposed a multi-algorithm dimension to obtain different result at each phase of the program. This system needs to extract keywords from different pages of a structured binary file like word and PDF, to achieve this; the research developed a parsing algorithm that will receive an input keywords and checks through every line of the file. This is called

parsing of files and at each turn, the system returns related keywords from each page. Python, regular expression module is very useful to detect pattern in a sentence or in the whole document that why it was recommended and used. The module imported regular expression object simply called ***re.compile()*** where ***re*** is a substitute of ***regexes.***

Algorithm of the operation of the system is given

Start program():
User login verification()
If login details == True
Open extraction panel ()
Else:
Print message() == "Invalid Login Details"
Select file you want to extract()
Select extract option ()
Enter keyword as KEY
Click on Extract Now
Set KEY as KEY object
While REGEXS imported {
Select file as FILE;
For KEY as KEY in extract ()
Set PATTERN with REGEXES
KEY.get() as KEY_EXTRACT
Re.compile(KEY)
}
SET VAR1 as RESULT
SET newfile = new file(KEY)
WHERE newfile not empty:
Newfile = RESULT()
RESET VAR1;
Close re.compile()
End Program()

ii. System Implementation

The research was implemented using Python programming languages and MySQL database management system.

iii. Results and Discussion

The view result module allows the user access to view extracted pages of the input file. This outcome displays as text on the screen until the user decides to carry specific actions such as encryption which is the security assurance earlier stated as one of the objectives of the study. The extracted words are presented in Fig. 8.

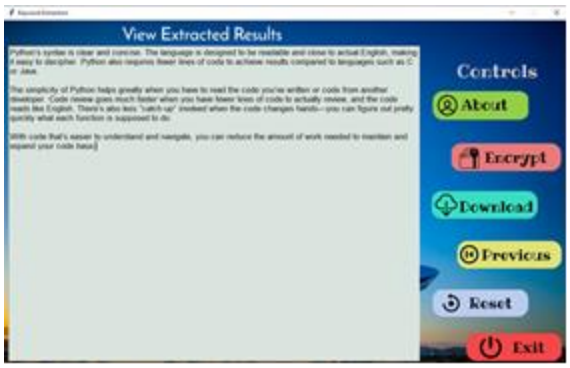


Fig. 8: Extracted Output.

Since the computers are electronic devices, they cannot accept data in human readable format of speech and manuscripts; this emphasizes the necessity for input devices which serve as a medium through which data can be presented to the computer in a way that it is convertible into machine-based readable forms. This input devices read data supplied by the user convert them into machine sensible form and produce output using corresponding output devices. Input devices would consist of keyboard and mouse while output devices would include monitors and printers.

In having access to the extraction application, the user authentication details will be requested as shown in Fig. 9. It requires username and password.

Fig. 9: User Authentication

The input design seeks to deploy the best strategy to facilitate data entry into the computer system accurately and in real time. The input design was developed in a format to accept all possible inputs necessary for the software to perform optimally without omission or errors. The application allows selection of extracted files that is uploaded for extraction process this is achieved by deploying input

devices with well-defined interfaces Fig. 10 presented the uploaded window for extraction.

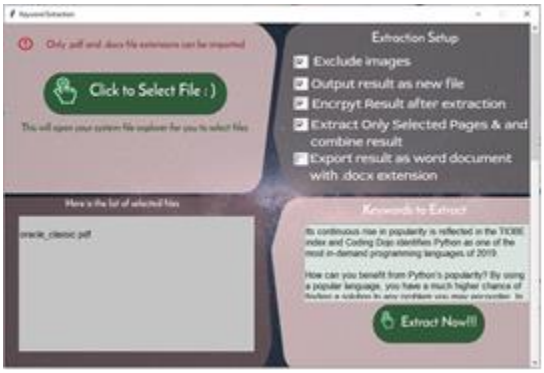


Fig. 10: Extraction Process

CONCLUSION

Based on findings during the study, we observed the need for every researcher to have this system as a support tool alongside any method that is chosen to be operated. When it comes to manual approach, is no doubt that paper reviews can take weeks running to months, but time frame issues could be mitigated using the approach the proposed study system offers will enable extraction using keywords. The result we got from system testing, deduced that the keyword extraction produced was efficiency to extract the relevant areas need for researches at a reduced time. The positive result signifies that the software will work effectively and efficiently for the purpose it was created. The research is recommends that; adopted and utilized in the areas of application for university researchers and private sectors. Also, the paper possesses promising area of research as the quality of the result could be improved. The improvement to the method and reducing time of information Extraction from files containing limitless pages. The system proposes a more efficient process that takes care of possible issues that could arise in the process; reducing redundancy and helping researchers derive the intended benefits of the research process.

REFERENCES

[1] Xiaofeng L., Zhiming Z.,(2019). Unsupervised Approaches for Textual Semantic Annotation, A Survey. *https://dl.acm.org/doi/pdf/10.1145/3324473* University of Amsterdam=

[2] Gupta, V., Lehal, G.S.(2009): A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell. 1(1), 60–

[3] Navathe, S.B., Ramez, E (2000).: Data warehousing and data mining. Fundam. Database Syst., 841– 872

[4] Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan(2016) Using Text Mining Techniques for Extracting Information from Research Articles Using Text Mining Techniques for Extracting Information from Research Articles Chapter in Studies in Computational Intelligence · January 2018 DOI: 10.1007/978-3-319-67056-0_18

[5] Zhang, Y., Chen, M., Liu, L.(2015): A review on text mining. 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 681–685.

[6] M.Geetha, R C Pooja, J. Swetha, N. Nivedha, T. Daniya (2020) Implementation of Text Recognition and Text Extraction on Formatted Bills using Deep Learning International Journal of Control and Automation Vol. 13, No. 2, pp. 646 – 651

[7] T. Gnana Prakash K. Anusha (2017) Text Extraction from Image using Python International Journal of Trend in Scientific Research and Development (IJTSRD) International Open Access Journal, ISSN 2456-6470

[8] [8] Xiaofeng L., Zhiming Z.,(2019). Unsupervised Approaches for Textual Semantic Annotation, A Survey. *https://dl.acm.org/doi/pdf/10.1145/3324473* University of Amsterdam

[9] Sonit S., (2018). Natural Language Processing for Information Extraction. *Department of Computing, Faculty of Science and Engineering, Macquarie University, Australia.* arXiv:1807.02383v1 [cs.CL]

[10] Chengyao L., Huihua L., Yuanxing D., Yunliang C., (2016). Corpus based part-of-speech tagging. *International Journal of speech Technology*.https://www.readcube.com/articles/10.1007/s10772-016-9356-2

[11] Kong F., Xiaoming Z., (2010). Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). *Association for Computational Linguistics.* https://aclanthology.org/2021.acl-short.pdf

[12] Ly Papa A.,; Pedrinaci, C., Domingue J., (2012). Automated information extraction from web APIs documentation. *The Open University's repository of research publications and other research outputs.* http://oro.open.ac.uk/34934/1/webApiDocProcessing.pdf

[13] Adnan., Akbar., (2018). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of big data.* https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0254-8

[14] Information Extraction (2020). *Information ExtractionAbstract*https://en.wikipedia.org/wiki/Information_extraction