International Journal of Research in Engineering, Science and Management
Volume-2, Issue-3, March-2019
www.ijresm.com | ISSN (Online): 2581-5792

Text Extraction and Metadata Analysis of PDF Documents: A Survey

Aarti Mete[1], Priyanka Kanthale[2], Pritam Bhaye[3], Rahul Subhedar[4], Shafali Gupta[5]
[1,2,3,4]*Student, Department of Computer Engineering, R. M. D. Sinhgad College of Engineering, Pune, India*
[5]*Assistant Professor, Dept. of Computer Engineering, R. M. D. Sinhgad College of Engineering, Pune, India*

***Abstract*: Most of the data is in the form of text. It is represented in the form PDFs and Web Pages. Text mining is a very significant step of Knowledge discovery process. Text Mining extracts hidden information from the documents that are not in structured form to the one with semi - structured or standard form. Text mining processes are used to extract relevant data from Documents, like Classification, Clustering, and Information Extraction. Metadata is data (information) that is supposed to provide information from other data. The metadata of PDFs consists of the basic information about the PDFs. The frameworks of text mining techniques as well as its tools for Extraction of text and metadata have been discussed.**

***Keywords*: Text mining framework, Text mining tools, Information extraction, PyPDF, JSON.**

1. Introduction

Text Mining is also termed as "knowledge discovery" which discovers the undiscovered data by extracting it from different PDFs. It is somewhat similar to data mining tools. Not all data can be stored in database while, we have text everywhere. We cannot analyze the data that is stored in co - operate database. Since unstructured data is more complex, higher efficiency techniques are required. Here comes the role of the tools for extraction of data that make the work of extraction easier. As text mining is extraction of useful information from text data it can be termed as Data Discovery. The problem is discovering patterns and trends out of massive data, which is a great challenge. The main objective of text mining is to discover relevant trend and patterns from database. Collecting information is easy, but finding relevant information from the collected information on demand can be difficult and also the problem is in the result as it is not important as per the users need. So the aim of text mining is to find the relevant information which is not known and yet not written down.

Text mining process is started with document collection from different PDFs. It would extract a PDF document and then pre-process it by checking format. Then the PDF document goes through the process of text analysis, which is also known as text analysis phase. Text Analysis is semantic analysis which derives information from text.

Sometimes text analysis techniques for a particular PDF document are repeated until relevant information is extracted. The resulting information is then stored in management information system which has abundant amount of knowledge for the user system.

There are various tools available for the extraction of the PDF. PyPDF tool is the one those tools. It is a pure python library built as a PDF toolkit, which is capable of doing various tasks such as extraction information (Metadata: Title, Author, etc.), page by page splitting of document, merging document page by page, page cropping, multiple page into a single page merging, PDF file encryption and decryption. As it is a python concept it executes on various python IDEs on any external libraries. It allows Document manipulation in storage. It is better than file streams on string input and output objects.

Metadata is a processed data that provides high level information about other data. The types of metadata are descriptive metadata, structural metadata, administrative metadata, referential metadata, statistical metadata and many more. Metadata is also known as the processed data that provides information about many aspects of data. It is used for summarization the basic information about data. This can make working with specific information easier.

Json is the data type in python which is used to store the fields of metadata of the PDFs. JSON is a light mass processed data exchange format. This makes it easy for Analysis. It is also easy for machines to parse and generate useful information n. It is a text format which is completely language independent. It also uses conventions. These conventions are similar to in C, C++, C#, Python, etc.

2. Literature review

The author of [1] paper discusses techniques used in text mining. To teach computers how to analyze, understand and generate text, technologies are produced by natural language processing. There are various technologies. The roles that they play in text mining are discussed in the following sections. The types of situations where each technology is useful are the areas of interest. But the paper does not tell much about the techniques of text mining but only the survey.

The author of paper [2] show us that Information extraction is first step for computer. It helps to analyze unstructured text. It is done by identifying key phrases. It is also done sometimes by identifying relationships within text. To do this task process of pattern matching is used. It is useful to look for pre fixed

series in text. Information extraction task includes tokenization, etc. It can also have part-of-speech assignment, identification of named entities, sentence segmentation, etc. Firstly, phrases and sentences are parsed. Then, semantically they are interpreted and are then required pieces of information are entered into the database.

Categorization automatically assigns one or more category to free text document. Categorization is supervised learning method because it is based on input output examples to classify new documents. Predefined classes are assigned to the text documents based on their content. The text categorization process consists of pre-processing and other various significant processes. The various Clustering and Classification methods can be used in order to find groups of documents. These groups are with similar content. The outcome of clustering is typically a partition called clusters numbered ask. These each clusters consists of a numerous documents. The contents of the documents within one cluster are more similar to each element in that cluster. Between the clusters more dissimilar elements are present. Then the quality of clustering is considered as better.

The author of paper [3] tells the representation of individual documents or groups of documents text flags are used to show document category and to show density colors are used. Visual data are at high use for any further analysis. Text mining puts large text and documental sources in visual levels. The user can interact with the document by zooming and scaling. Text summarization is to reduce the detail of a PDF files. It is better than retaining significant data and basic semantics. Text summarization is helpful. It checks whether or not a lengthy document satisfies the user's requirement. It is important to read for further information.

3. Text mining basic stages

Usually the Text mining process has to undergo various steps. It is a very deep concept having numerous sub processing stages. Here are some main stage sin text mining. Processing the text includes things such as classification, visualization, extraction. The steps in the Text mining are below:

A. Stage I: Pre-processing Step

This step is the foremost step in the text mining process. It mainly deals with the removal of noises and other non-useful things. This step has to be carried out before carrying out any further step.

B. Stage II: Applications of techniques

After the Pre-processing, the appropriate algorithms are applied over the extracted clean step such as the classification algorithms, clustering algorithms, etc.

C. Stage III: Text Evaluation/Analysis

As the main output to expect is the Knowledge that is important for the further study or survey. Thus the proper analysis of this data can be done to have processed data called as Knowledge.

4. Proposed architecture

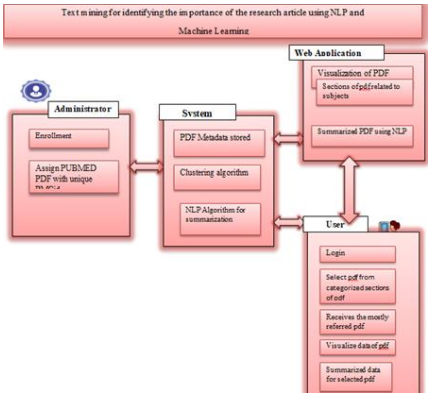

Fig. 1. Proposed architecture

5. Conclusion

Text mining or extraction is an interdisciplinary field which comes under the machine learning and artificial intelligence domain and draws on information retrieval, Data Mining. There are some problems that occur while extracting the text like to extract explicit and implicit concepts using NLP techniques. The main objective of this is to get information into larger quantities of text data that is processed or is yet to be analyzed further. Text mining in this is discussed with various techniques and usage. Extraction of unstructured text information to structures information, the technique of information extraction is used. One of the explored tools, we preferred PyPDF. We use text mining for extracting the PDF file. The contents are extracted from the PDF file and analysis is carried out. Also, we extract the metadata of the PDF using text mining.

References

[1] Manasi Joshi, Varunakshi Bhojane, "Literature Survey on Text Mining Techniques," vol. 3, no. 2, 2013.
[2] Vaishali Bhujade, N. J. Janwe, "Knowledge Discovery in Text Mining Techniques using Association Rule Extraction", 2011 International Conference on Computational Intelligence and Communication System, 2011
[3] F. S. Gharehchopogh, Z. A. Khalifelu "Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing"., IEEE, 2011.
[4] M. Sukanya, S. Biruntha "Techniques on Text Mining" 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2012.
[5] W. He, "Examining students online interaction in a live video streaming environment using data mining and text mining," Computers in Human Behavior, vol. 29, no. 1, pp. 90–102, 2013.
[6] Divya Nasa "Text Mining Techniques- A Survey" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.
[7] Mahesh T. R, Suresh M. B, M. Vinayababu "Text Mining: Advancements, Challenges and Future Directions" International Journal of Reviews in Computing 2009-2010 lJRIC & LLS.
[8] S. Anzaroot and A. McCallum. A New Dataset for Fine-Grained Citation Field Extraction. In ICML Workshop (PEER), 2013.
[9] Apache. PdfBox. hps://pdox.apache.org/, 2017.
[10] Ø. R. Berg. PDFExtract. hps://github.com/oyvindberg/PDFExtract/, 2011.