METADATA EXTRACTION FROM SCIENTIFIC PDF

[1]Vishal Rambhau Kardile, [2]Shivam Nanagir Gosavi, [3]Bhushan Gokul Jadhav, [4]Vaibhav Rajendra Joshi

MET's INSTITUTE OF ENGINEERING, NASHIK

Abstract: With the availability of World Wide Web in every corner of the world these days, the amount of information on the internet is growing at an exponential rate. However, given the hectic schedule of people and the immense amount of information available, there is increase in need for information abstraction or summarization. Be it browsing through the seemingly endless pages of terms and conditions on an important official document or kicking back and flipping through an intriguing eBook- reading is quite an undeniable and inescapable part of our everyday lives. However, reading anything demands our complete undivided attention making it nearly impossible for us to multitask. This Online PDF to Audio Converter and Translator was created by using Python (Django) can instantly convert any PDF text into audio. Along with reading any PDF document out loud, this application can also translate and vocalize any text into up to five languages. Text summarization presents the user a shorter version of text with only vital information and thus helps him to understand the text in shorter amount of time. The goal of this project is to condense the documents or reports into a shorter version and preserve important contents convert that summarized text into audio for better understanding of the user. Also projects convert the generated summery to the audio for better understanding.

***Keywords*: Python, NPL, PDF Extraction, audio converter, machine learning**

INTRODUCTION

Natural Language Processing (NLP) is an area of application and research that explores how computers can be used to understand and manipulate natural language speech or text to do useful things. The foundation of NLP lie in a number of disciplines, namely, computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence robotics, and psychology. NLP researchers aim to gather knowledge on how human beings use and manipulate natural languages to perform desired tasks so that appropriate tools and techniques can be developed. Applications of NLP include a number of fields of study such as multilingual and cross-language information retrieval (CLIR), machine transaction, natural language, text processing and summarization, user interfaces, speech recognition, artificial intelligence and expert systems. Text-to-speech and related read audio tools are being widely implemented in an attempt to assist students' reading comprehension skills.PDF to the audio system is a screen reader application designed and constructed for an effective audio communication system. PDFs were designed to present and exchange documents reliably, PDFs are an open standard document format used globally, maintained by the International Organization for Standardization (ISO). The document format is one of the most convenient methods for electronic communication, and also for the exchange of information. Hence, there is a need to make it more accessible to readers on-screen through audio. PDF documents are designed and structured to contain links and buttons, form fields, audio or sounds, video, and business logic. The PDF to the audio system will power text on screens to read aloud (speak) with support for many languages [2]. The PDF to Audio Converter project provides an alternative to access the PDF books for the blind, lazy, 1 Metadata extraction from scientific pdf readers, and others. Using this PDF to Audio Converter the user will be able to listen to his favorite PDF and can do their daily routine. The following application can be used to convert text from PDF to audio using Python predefined libraries

LITURATURE SURVEY

• "An approach to sentence-selection-based text summarization", Fang Chen; Kesong Han; Guilin Chen is a author of this paper, this paper published in 2016. This paper presented an We introduced a newly developed text summarization system. It supports both Chinese and English, while this paper focuses on Chinese processing. We apply 6 word level features and 3 sentence level features to weigh each word and sentence. We also describe two new techniques, one is for processing the topic sensitive word feature and another is for processing the sentence length feature. Primary subjective evaluation shows that these approaches are effective and efficient, and performance of the system is promising.[3]

• "Automatic Text Summarization Using Hybrid Fuzzy GA-GP" is paper of A. KianiB; M.R. Akbarzadeh-T , 2015 A novel technique is proposed for summarizing text using a combination of Genetic Algorithms (GA) and Genetic Programming (GP) to optimize rule sets and membership functions of fuzzy systems. The novelty of the proposed algorithm is that fuzzy system is optimized for extractive based text 3 Metadata extraction from scientific pdf summarizing. In this method GP is used for structural part and GA for the string part (Membership functions). The goal is to develop an optimal intelligent system to extract important sentences in the texts by reducing the redundancy of data. The method is applied in 3 test documents and compared with the standard fuzzy systems as well as two other commercial summarizers: Microsoft word and Copernic Summarizer. Simulations demonstrate several significant improvements with the proposed approach.[5]

• "Generic text summarization using local and global properties of sentences" C. Kruengkrai; C. Jaruskulchai; 2015. In this paper described The paper With the proliferation of text data on the World-Wide Web, the development of methods for automatically summarizing these data becomes more important. Here, we propose a practical approach for extracting the most relevant sentences from the original document to form a summary. The idea of our approach is to exploit both the local and global properties of sentences. The local property can be considered as clusters of significant words within each sentence, while the global property can be thought of as relations of all sentences in the document. These two properties are combined to get a single measure reflecting the in formativeness of sentences. Experimental results show that our approach compares favorably to a commercial text summarizer.[2]

• "A Review on Optical Character Recognition and Text to Speech Conversion" Swati Vikas Kodgire; 2013. The application depending on image and voice with a parallel functioning is suitable to assist physically challenged people. So that dependability of a challenged person is decreased to a improved level. Image acquisition based text reader can help visually challenged people to manage the handheld objects in day to day life. Initially steps involves capturing of image, distinguishing image with text portion and residual regions, image pre-processing on region of interest, after the extraction of characters and words, conversion of text to speech is done. To splinter text from a document it is obligatory to discover all the possible manuscript text regions. Text detection, line detection, character identification, feature extraction, training of extracted features are the steps in sequence that are executed.[1]

AIM & OBJECTIVES

• Easy to clear the idea: Instantly Reading the entire article, breaking it and separating the important ideas from the original text takes time and effort.
• Important facts: This highly improves productiveness as it quicken surfing process, Does Not Miss Important Facts.
• Improves Quality: Some software summarizes not only documents but also web pages

MOTIVATION

A system for the summarization of single documents. The system produces multi as well as single document summaries using data mining techniques for identifying common terms across the set of documents.

SYSTEM ARCHITECTURE

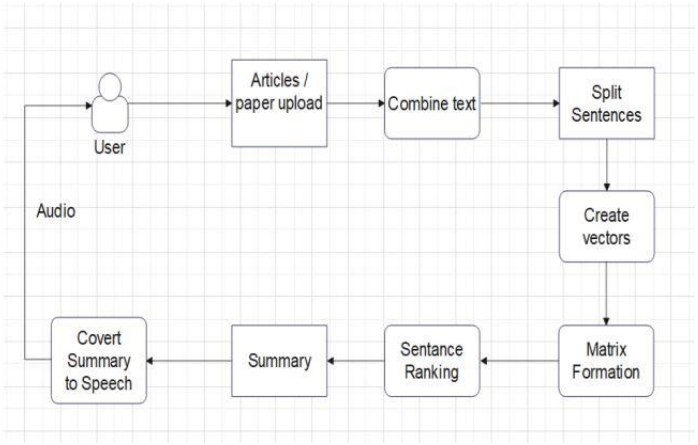

Fig -1: System Architecture Diagram

APPLICATION

• This feature will help mostly for the disabled persons like the blind and handicap.
• Teachers and school librarians may also use these findings as a rationale for adding audiobooks to the list of reading strategies used successfully with struggling readers.
• Those who participated in the studies and on audiobook usage of English Language Learners usually.

FUNCTIONAL & NON-FUNCTIONAL REQUIREMENTS

Functional requirements:

• The System should be able to retrieve the results stored on database by using quick retrived process.
• The system application of modules must able to encrypt the data and decrypt it whenever needed.

Nonfunctional Requirements

• There should be minimal lag between taking of the processing and result
• The processing should be as efficient with maximum accuracy.
• The system should give valid result for positive as well as negative test cases.
• Usability: The ease with which the system can be learned, managed or used. Usability gives the measure of how much user friendly the system is.
• Reliability: The degree to which the system must work for users. It also refers to the mean time between failures, means what can be the maximum down time. MET's Institute of Engineering 15 Metadata extraction from scientific pdf
• Performance: Performance specifications typically refer to response time, transaction throughput, and capacity. They deal with response time, which means the time taken by the system to load, reload, screen open and refresh times etc.
• Scalability: It refers to the ability of the proposed software application to increase the number of users or applications associated with the product.
• Open standard: t ensures the viability and future expansion of the system, all offered development tools, server software, as well as, the application are based on open templates and are available under the terms of the General Public License

Functional requirements

- Registration
- User Login
- Creation of database: Users Mandatory Information

Design Constraints:

1. Database
2. Operating System
3. Web-Based Non-functional Requirements

Security:

1. User Identification
2. Login ID
3. Modification

Performance Requirement:

1. Response Time
2. Capacity
3. User Interface
4. Maintainability
5. Availability

SYSTEM REQUIREMENTS

Software Used:

Operating System: Windows xp/7/8/10 2.
Programming Language: Python
Software Version: Python 4.4 4.

Tools: Anaconda/pycharm
Front End: Python

Hardware Used:

- I3 processor or above
- 150 GB Hard Disk or above
- 4 GB RAM or above

CONCLUSION

The Conclusion of this project is that the client will get an web application that will execute on client side and get the summary of the input document as per clients requirement. The automatic generated summary is useful for the client to understand the core concept of the document with in few lines instead of reading whole document. It was seen that this code performs really well in reading straightforward PDF text files. Should enable users to select the desired PDF and convert it to audio and display text in, so the user can understand that particular text has been read. Should enable students with reading disabilities. The success of this research project is significant given the broad use of audiobooks in literacy and library programs across the United States. Teachers and school librarians may also use these findings as a rationale for adding audiobooks to the list of reading strategies used successfully with struggling readers.

REFERENCES

[1] Pankaj Gupta, Vijay Shankar Pendhluri, Ishant Vats, "Summarizing text by ranking text units according to shallow linguistic features", Feb. 13 16, 2011 ICACT, 2011.
[2] Rajesh S. Prasad, U. V. Kulkarni, Jayashree R. Prasad, "Connectionist Approach to Generic Text Summarization,", World Academy of Science, Engineering and Technology 55,2009.
[3] R. S. Prasad, U. V. Kulkarni, J. R. Prasad, "A Novel Evolutionary Connectionist Text Summarizer (ECTS),", 2009,IEEE Xplore.
[4] Rajesh Shardanand Prasad, Uday. V. Kulkarni, "Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization", Journal of Computer Science 6 (11): 1366-1376, 2010 ISSN 1549-3636, 2010 Science Publications.
[5] Ranjit Bose "Natural Language Processing: Current state and future directions", International Journal of the Computer, the Internet and Management Vol. 121, January – April, 2004.
[6] Natural Language Processing Techniques Applied in Information Retrieval-Analysis and Implementation in Python, TulikaNarang, International Journal of Innovations Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 5, Issue 4 April 2016.
[7] Pdf. (2021, March 08). Retrieved March 09, 2021, from https://en.wikipedia.org/wiki/PDF
[8] 7 ways Audio books benefit students who struggle with reading. (n.d.). Retrieved March 09, 2021, from: :https://learningally.org/Solutions-for School/7-Ways-Audio books-Benefit-Students WhoStruggleWith-Reading