IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

A WEB APPLICATION FOR TITLE GENERATION AND SUMMARIZATION OF RESEARCH PAPERS

[1]Ranjitha B B, [2]Dr. Vijayalakshmi M N
[1]Student, [2]Associate Professor
[1]Department of Master of Computer Applications,
[1]RV College of Engineering®, Bengaluru, India

Abstract: According to Research and Development (R&D) statistics, the number of researchers in India's R&D sector is increasing at a 10.84 percent annual pace. Furthermore, according to US government report created with statistics compiled by the US National Science Foundation (NSF), India has published over 1.35 lakh scientific papers each year, making it the world's third largest publisher of science and engineering journals, accounting for 5.31 percent of total world publications in the region. The title of a paper is one of the most crucial aspect of a research paper. As a consequence, it's important for a researcher to entitle the paper in such a way that it's acknowledged. Researchers invest a large amount of time towards reviewing research papers to learn about the previous research works related to their subject. This paper presents, a web application for generating quality title for a research paper using deep learning neural network as well as presentable summary for research papers using Natural Language Processing (NLP).

***Index Terms* - Title generation, Text summarization, Natural Language Processing (NLP), Neural Networks.**

I. INTRODUCTION

A research paper begins with an original hypothesis or intent argument about a subject and then builds on it with evidence gathered from various sources. The aim of a research paper is to educate the readers so that they can draw on what others have said about a topic and resonate with the sources in order to thoughtfully provide a unique perspective on the subject. According to the survey, India's rise to third largest publisher is largely due to a remarkable double-digit growth rate over the last decade, from 2008 to 2018. Over the last ten years, global research productivity, as calculated by peer-reviewed Science and Engineering (S&E) journal articles and conference papers, has increased by about 4% per year. This rose at a 10.73 percent annual pace, and the nation now accounts for 5.31 percent of all science and engineering publications worldwide. Also, between the start of the pandemic and October 2020, more than 87,000 papers on coronavirus were written, according to a new report. [1]

The title of a paper is normally the first thing that is read and gives a general understanding of what the paper is about. People mostly decide whether or not to refer to the paper solely on the basis of its title, no matter how valuable the contents of the paper are. Therefore, it becomes the most significant component of the research paper. The more enticing the title, the more likely users are to cite it. As a result, it's important to choose a title that's simple, straightforward and accurately portray the entire research paper in a few words, and captures users' attention. Apart from this, to come up with a satisfactory title, researchers typically go through several rounds of revisions which is time consuming exercise leading to more confusions.

Hundreds of research papers are written every day on a various similar topics, as well as more researches are conducted. Researchers must refer to a greater number of research papers in order to learn about prior related studies. To refer to each paper, researchers must read the entire document, which is a time-consuming and repetitive process. Apart from them, there are large set of people who do research for various purposes (e.g.: literature survey as a part of project works).

The development of a web application to provide suggestions for titling a research paper by creating titles for a research paper and to produce a summary of research papers by summarizing the entire papers into a few sentences would be extremely useful and beneficial to users. In this paper, a web application is being proposed, which can be used for title generation and summarization purposes. It helps greatly in effectively titling the scientific papers, enabling researchers to publish their papers in widely known publications and earn more citations. Also, to comprehend the content of the research papers in a short period so that they can refer a greater number of research papers in less time and save their precious time.

II. Literature Survey

There are a number of approaches for extracting text from pdf files, as well as a wide range of natural language processing techniques and deep learning models for text summarization and text generation. Many authors have developed different methodologies with various specifications.

Text summary is the way to shorten a software text document to produce a summary of the main points of the original document. Text summarization is the process of reducing a text document using software to produce a description or abstract of the original document. The output type of a summarizer can be defined as follows: extractive, where important sentences from the input text are selected to form a summary; or interactive, where important sentences from the input text are selected to form a summary. [2]

Single document summarization techniques generate summaries of a single document's text, while multi-document summarization techniques generate summaries of multiple documents. Query-based summary models include text synopsis based on a particular field defined by the user's query. Generic summaries are mostly abstractive and concentrate on the general area of text input. Different methods for optimizing the proposed objective function were used for the purpose of summarizing the document. [3]

A model as a solution for text summarization that is based on the Extractive Approach, beginning with Natural Language Processing as the fundamental model is created. It's used to provide the appropriate description while maintaining the original text's context. As it is obvious, the extractive method selects different and distinct sentences/sections of the text/document before combining them to form a description. [4]

Since title generation is considered as one of the automated text summarization tasks, a neural network was proposed to generate titles from a body of posts. Since a title serves as an abstract for an article in this method, it is assumed that a title can be produced using an automated text summarization approach. An encoder and a decoder are the two components that make up a title generator. Recurrent neural networks are used to implement the encoder and decoder. [5]

A method for creating or discovering the core theme of a storyline or document in English without having to read the entire document. The ability to generate titles would aid newspaper editors and technical writers in quickly locating the central theme without having to read the entire post. The size of the database has a direct impact on the system's results. Python is a versatile language for language processing. The challenge of artificial intelligence is to create computational models and approaches to cognitive processes. [6]

The 'Layout-Aware PDF Text Extraction' system was introduced to make accurate text extraction from PDF files of research papers easier for text mining applications. In addition, the quality of the text extracted by LA-PDFText was compared to the text from PubMed Central's Open Access subset. The quality of this text was then compared to that of the text extracted by the PDF2Text method, which is widely used to extract text from PDF files. [7]

III. Proposed System

The proposed methodology is to develop web application for summarizing research papers using NLP techniques. The titles are generated using a Recurrent Neural Network with LSTM which is trained with a large corpus of titles to produce a title. Fig 1 shows the architecture diagram of the entire application. The methodology can be executed in two phases.

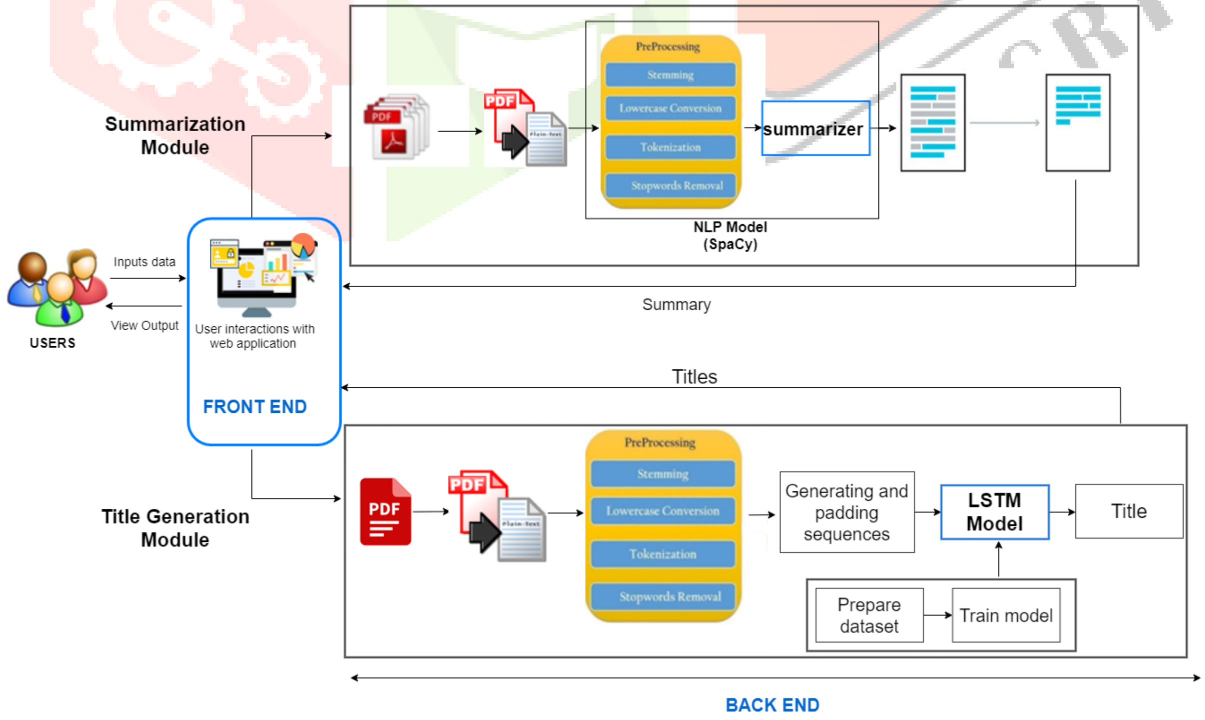

Fig 1: Architecture diagram of the proposed system

Phase I: SUMMARIZATION

In this phase summary of the research papers is carried out by collecting the different papers belonging to the same domain. At a time three papers are taken in pdf format to generate the summary. Then it extracts keywords from each paper to validate whether all the papers are of same subject.[8] Next, it extracts text from each research paper for parameter analysis.[9] The extracted text from the research papers are then combined and sent to the NLP summarizer. The summarizer involves two phases. The first phase is concerned about pre-processing of text which involves removal of punctuations, special characters, stop words and makes the entire text into lower case. Then the processed text is tokenized. In the second step, the tokenised text is used to calculate word frequencies which are normalized by dividing by the maximum frequency. After that, the sentences with high frequencies are identified and the most important sentences are taken to generate the summary. [4] The generated summary can be downloaded as a text file for later usage.

The summary of the papers can also be generated accepting the Uniform Resource Locator (URL) of the papers for non-downloadable research papers and any research articles as well. The text is extracted from the web page and summary of the paper is generated. [10].This phase is highly beneficial to researchers and saves their valuable time.

Phase II: TITLE GENERATION

This phase provides suggestions for titling a research paper by generating multiple titles, out of which the user can make choice. The model is developed based on LSTM networks. [11, 12] This model is trained over multiple large datasets [13, 14] containing titles of research papers, articles etc., on different domains belonging to different categories. Title is generated taking keywords or research paper .If user provides research paper as input, then the text is extracted from the file and pre-processed which involves lemmatization and identifies few important keywords from the research paper. These keywords are passed into the model and based on the keywords, the model generates titles for a research paper. The user can make use of any of the suggested titles and can also copy the title for using it. This is very useful for users as it saves time and clears up confusions on to what should be the title of the paper for researchers because it allows them to have a suitable title in less time, increasing their chances of receiving further citations through which their work gets recognized more, by producing efficient titles.

IV. CONCLUSION

This paper suggested a web application to assist researchers in entitling their research papers or articles by proposing multiple titles, making the research paper more appealing to read. This method helps in referring to a greater number of research papers in a shorter amount of time by producing a presentable description from multiple research papers. This simplifies the reading of research papers for those who read them on a regular basis by summarizing their findings. The proposed system significantly assists users in saving valuable time. As a result, this application is beneficial not only to researchers, but also to students, teachers, and learners.

REFERENCES

[1] Press Trust of India, India is world's third largest producer of scientific articles, following China and US: Report, INDIATODAY, January 2, 2020, https://www.indiatoday.in/education-today/latest-studies/story/india-is-world-s-third-largest-producer-of-scientific-articles-following-china-and-us-report-1633351-2020-01-02

[2] Ishitva Awasthi, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand and Prof. Piyush Kumar Soni, Natural Language Processing (NLP) based Text Summarization - A Survey Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021], IEEE Xplore Part Number: CFP21F70-ART; ISBN: 978-1-7281-8501-9, 2021

[3] Rahul, Surabhi Adhikari and Monika, NLP based Machine Learning Approaches for Text Summarization, Proceedings of the Fourth International Conference on Computing, Methodologies and Communication (ICCMC 2020), IEEE Xplore Part Number:CFP20K25-ART; ISBN:978-1-7281-4889-2, 2020

[4] Swaranjali Jugran, Ashish, Bhupendra Singh Tyagi, Mr. Vivek Anand Scse, Extractive Automatic text summarization using SpaCy in python & NLP, 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), ISBN:978-1-7281-7742-7, 20 April 2021

[5] Yuko Hayashi and Hidekazu Yanagimoto, Title Generation with Recurrent Neural Network, 5th IIAI International Congress on Advanced Applied Informatics, 2016

[6] Nandini Sethi , Prateek Agrawal , Vishu Madaan, Sanjay Kumar Singh And Anuj Kumar, Automated Title Generation In English Language Using Nlp, I J C T A, 9(11) 2016, Pp. 5159-5168, © International Science Press, 13 July 2019.

[7] Cartic Ramakrishnan, Abhishek Patnia, Layout-aware text extraction from full-text PDF of scientific articles, Ramakrishnan et al. Source Code for Biology and Medicine 2012, 7:7 http://www.scfbm.org/content/7/1/7, 2012

[8] Lucas Pluvinage, Extracting scientific results from research articles, Inria Publications, https://hal.inria.fr/hal-02956526, October 2, 2020

[9] Mike, Exporting Data From Pdfs With Python, Mouse Vs Python, May 3, 2018 https://www.blog.pythonlibrary.org/2018/05/03/exporting-data-from-pdfs-with-python/

[10] S Thivaharan, G Srivatsun, S Sarathambekai, A Survey On Python Libraries Used For Social Media Content Scraping, 2020 International Conference On Smart Electronics And Communication (ICOSEC), INSPEC ACCESSION NUMBER: 20051995, DOI: 10.1109/icosec49089.2020.9215357 publisher: IEEE 07 October 2020

[11] Sivasurya Santhanam, Context Based Text-Generation Using LSTM Networks, Conference: Artificial Intelligence International Conference – A2ic 2018, November 2018

[12] Ilya Sutskever, James Martens, Geoffrey E. Hinton, Generating Text With Recurrent Neural Networks,Conference: Proceedings Of The 28th International Conference On Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, January 2011

[13] Brian Maltzan, ARXIV, Cornell University Library, (18 April 2021), ARXIV Dataset, (Version 21) Kaggle.com

[14] PratirupGoswami, (15 April 2020), Articles Medium/Analyticsvidhya/TowardsDataScience, (Version 2) Kaggle.com