iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33966

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Automatic Text Summarization for Extracting Key

Prabhat S. Mishra[1], Heer R. Shah[2], Rajni R. Yadav[3], Zubaida Khan[4]
[1, 2, 3]*Student,* [4]*Assistant Professor, Computer Engineering, K C College of Engineering Management Studies and Research Thane (east) Maharashtra, India - 400603.*

Abstract: ***The process of gaining and absorbing the knowledge from various sources is a time consuming process where people, mainly youth spend time surfing over the internet for relevant information. The proposed system mainly focuses on scraping the data from websites and providing the summary as well as keywords from the information extracted from various websites giving user flexibility to select the website of their choice. The proposed system for the text summarization and keyword extraction undergoes a sequence of steps starting from data extraction from website link, removal of outliers and irrelevant information, emphasizing on importance of particular data extracted from the website and creating summary of the extracted data. For selection of relevant information from the extracted data it is necessary to use natural language processing. The proposed project helps its users to reduce their surfing time and gives summary prepared from multiple website links and documents or keywords from a particular website or a document.***

I. INTRODUCTION

Summarization of any data plays a vital role in integrating central ideas in a meaningful way and to ignore irrelevant information. Keywords drawn out of the summary helps in underlining the main idea of the document. It saves adequate time for different domains of use cases like marketing, institutions, education, business etc. Data mining involves the process of generating new data by evaluating already existing large data sets. Classification, clustering, regression, association, outlier detection, prediction, tracking sequential patterns are the techniques used for data mining. The raw data is converted into useful information using these techniques.

Web mining is the procedure of one of the data mining techniques which emphasizes on the world wide web and its components as the primary source of data. It discovers patterns and evokes valid information from documents . It is used to find pattern in web pages and web documents by collecting and analyzing information to gain insight of the overall data. It aims to extract/mine useful information or knowledge from the web page content. Keyword extraction, being the most important, is a process of highlighting important words, phrases and expressions in a particular content. It is done using Natural Language Processing(NLP).

II. EXISTING SYSTEM

Some recent works in this area include text summarization in the medical domain (Afantenos et al., 2005), empirical methods in text summarization (Das and Martins, 2007), survey of extraction-based text summarization and text mining (Gupta and Lehal, 2009, 2010). In Louis et al. (2010), specific aspects of discourse that provide the strongest indication for text importance is analyzed.
They investigate both the graph structure of text provided by discourse relations and the semantic sense of the
these relations in the context of content selection for single document summarization of news.

III. LITERATURE REVIEW

Mining Topical Relevant Patterns for Multi document Summarization - Yutong Wu, Yang Gao, Yuefeng Li, Yue Xu, 2015 proposed a method with semantic and context based analysis for multi document summarization which measures the important information covered in a sentence. It also proposed a method which tries to reduce the sentence size and combines similar sentences to create new sentences.

A survey on Real-Time Accumulative Short Text Summarization on Comment Streams N. Vijay Kumar, Dr. M. Janga Reddy, 2017 performed a survey on real time accumulative short text summarization on comment streams, which uses an algorithm In-creSTS, which can incrementally update the clustering results to provide an effective summary. Thus these papers help in understanding how summarization of single and multiple documents, blogs and single website page is done and how noise is removed from sentences
An Integrated Approach to Web Document Summarization Using Semantic Similarity K .Vanisri, P. Ponnila, J. Jeejovetharaj, 2014 proposed an integrated approach to web document summarization using semantic similarity, which mainly focuses on ranking of clusters using the K-means clustering algorithm and enhances the quality of the obtained summary.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

IV. PROPOSED SYSTEM

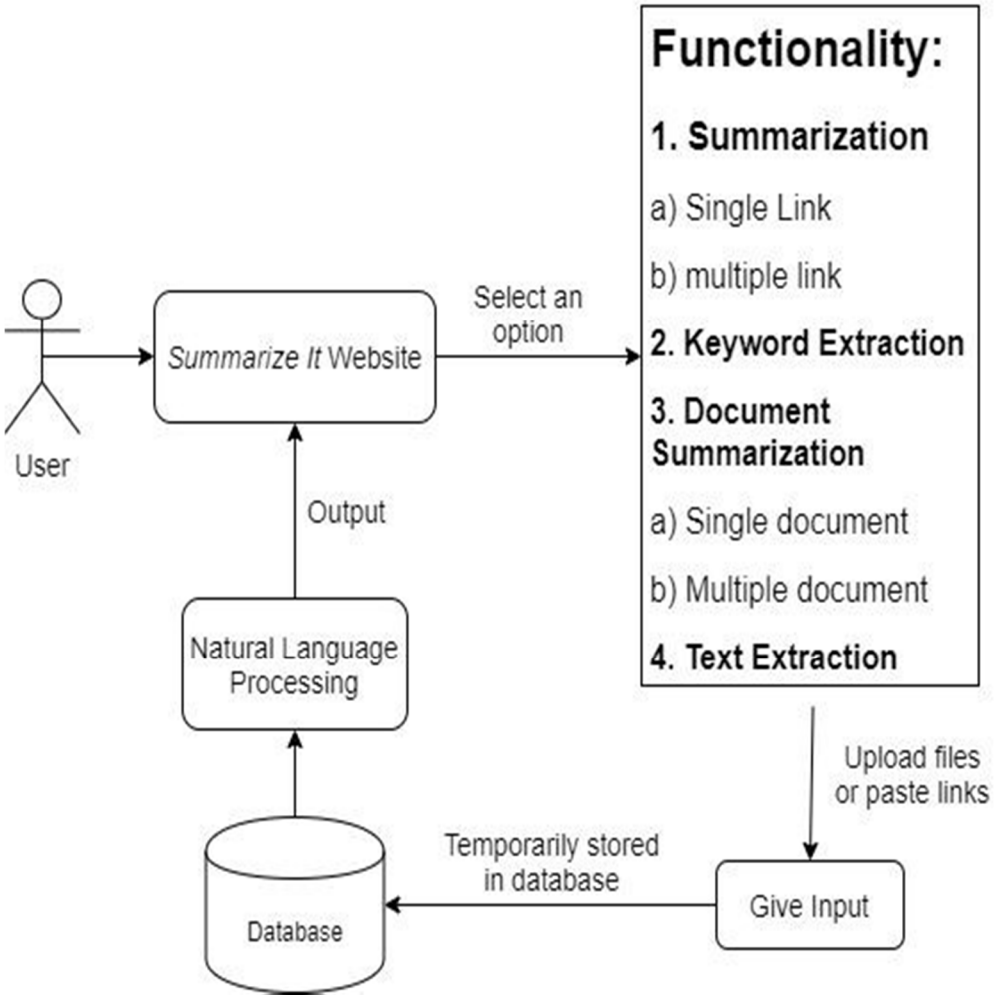


Fig.1 Proposed System Flowchart

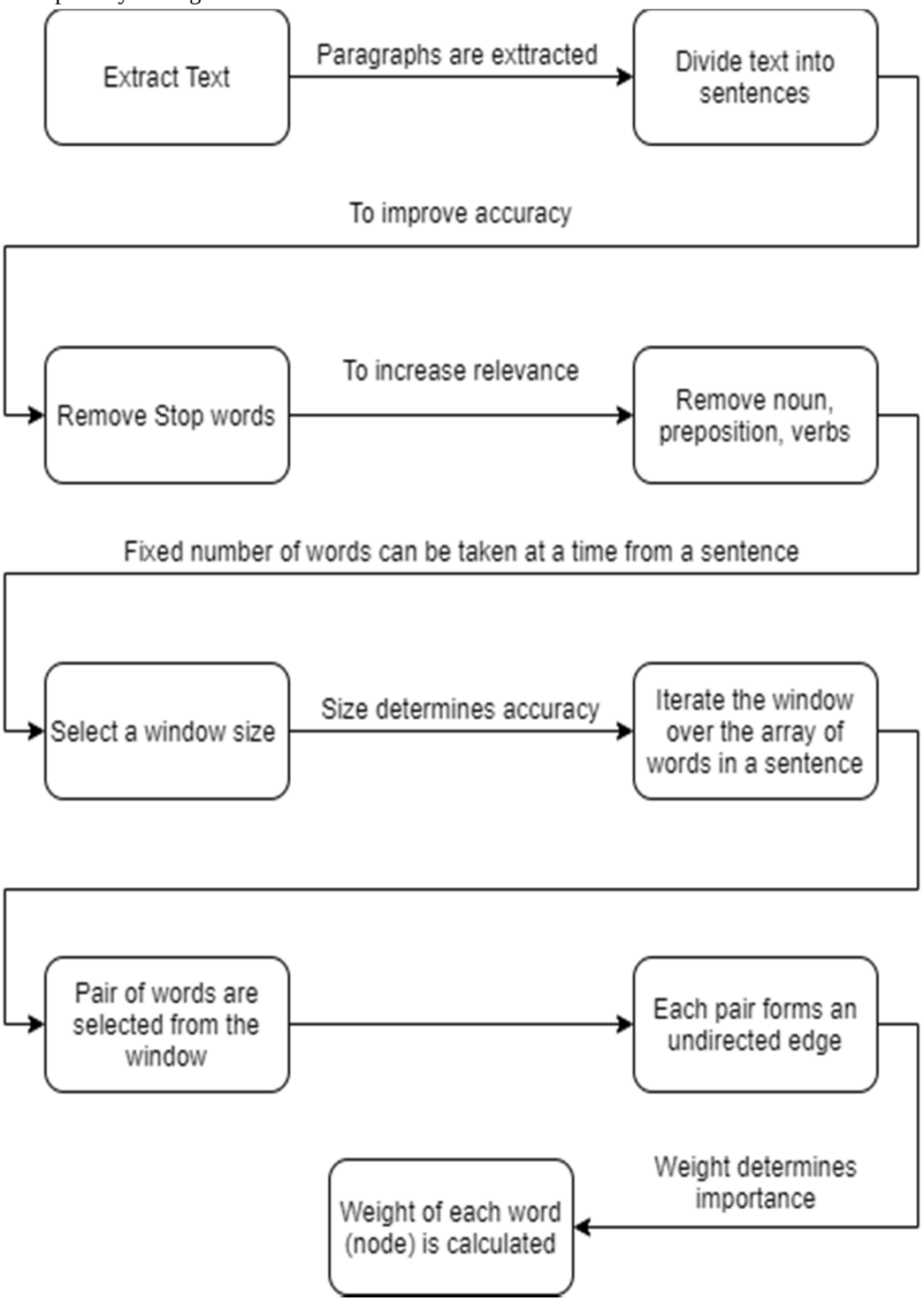
1) *Splitting Paragraphs Into Sentences:* When we split a paragraph into sentences, we do it on the basis of encountering a fullstop. This is very important for the further data cleaning process.
2) *Cleaning the Text:* After splitting paragraphs into sentences, the text need to be cleaned. For cleaning the data, all the special characters, numerals, stop words need to be deleted.
3) *Tokenization:* Tokenizing means breaking the sentences into words. These words are later retrieved in an array separated by commas.
4) *Counting Frequency:* After tokenizing the sentences into tokens, the next important step is to find the frequency of the words. Frequency is the number of times the word has occurred in the particular text. After finding the frequency, weighted frequency of the word is calculated. Weighted frequency is the ratio of the frequency of the word and the frequency of the word which has occurred the most number of times.The weighted frequency of the stopwords will be zero.
5) *Renewing The Words In The Sentences By The Weighted Frequency:* After calculating the weighted frequency, the words in the sentences are renewed with the weighted frequency. After this process, the sum of all the weighted frequencies of the words in the sentence is calculated. The sums hence obtained are stored in the array.
6) *Reverse Sorting:* The sums obtained in the above process are then sorted in the reverse order of their corresponding value. That is, we look for the sentences which have the highest sum. If two sentences have the same calculated sum, the sentence which appears first in the text is considered first. After selecting the sentence with the highest sum, the sentence with the second highest sum is then attached to the first sentenced to make the summary more relevant.

©IJRASET: All Rights are Reserved

V. METHODOLOGY

A. *Extraction*

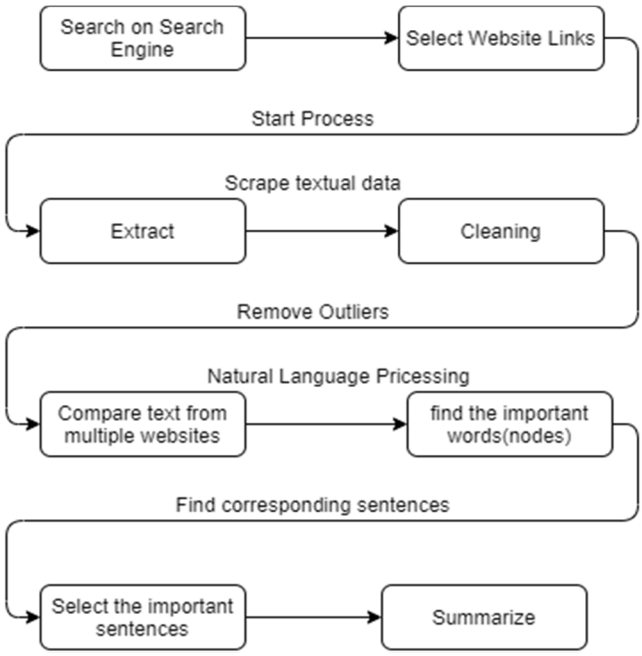
1) Website Text Extraction: For summarization and extraction of keywords it is necessary to extract the data first from the website websites. The websites links are temporarily stored as a buffer which this then provided to the system one at a time to extract.
2) Pdf and Word Data Text Extraction: Data can also be extracted from different sources like pdf and word files which requires uploading and temporarily storing the file for extraction.



©IJRASET: All Rights are Reserved

B. Summarization

Text summarization is creating a summarized data from the given text which has high relevance. Therefore it helps to reduce the user's effort to read and summarize a given text which could take hours differing from person to person. Summarization is done for creating a smaller version of the text which gives an overview of what the overall content conveys and saves time to read, understand and then infer from the complete text. Natural Language Processing (NLP) is used for the text summarization. For text summarization using NLP, Natural Language Toolkit (NLTK) library is used. NLTK module is used for Natural Language Processing(NLP). NLP is process of getting a computer to understand natural language and usually this in the form of written language and sometimes it can be in the form of spoken language. But usually spoken language gets converted to written language and then to numbers. Following are the steps taken for text summa



VI. CONCLUSION

The proposed system implements website link and document summarization using natural language processing which helps the users to save time. The user is given the liberty to choose multiple links of their choice from any search engine. Multi-document and multi-webpage summarization support enables user to use the functionality even more efficiently. It gives the summary of the individual links and also the combined summary of the links as per the user's requirement. This is what makes it different from the already existing systems. The keyword extraction feature also plays a vital role in providing the user with the gist of the complete document or website within seconds. The size of summary is thirty percent of the total extracted text in the first step. The functionality to add direct text as input for summarization helps the user to obatin summary of blog posts, any other post from social media sites or particular textual data which they want to summarize.

VII. ACKNOWLEDGEMENT

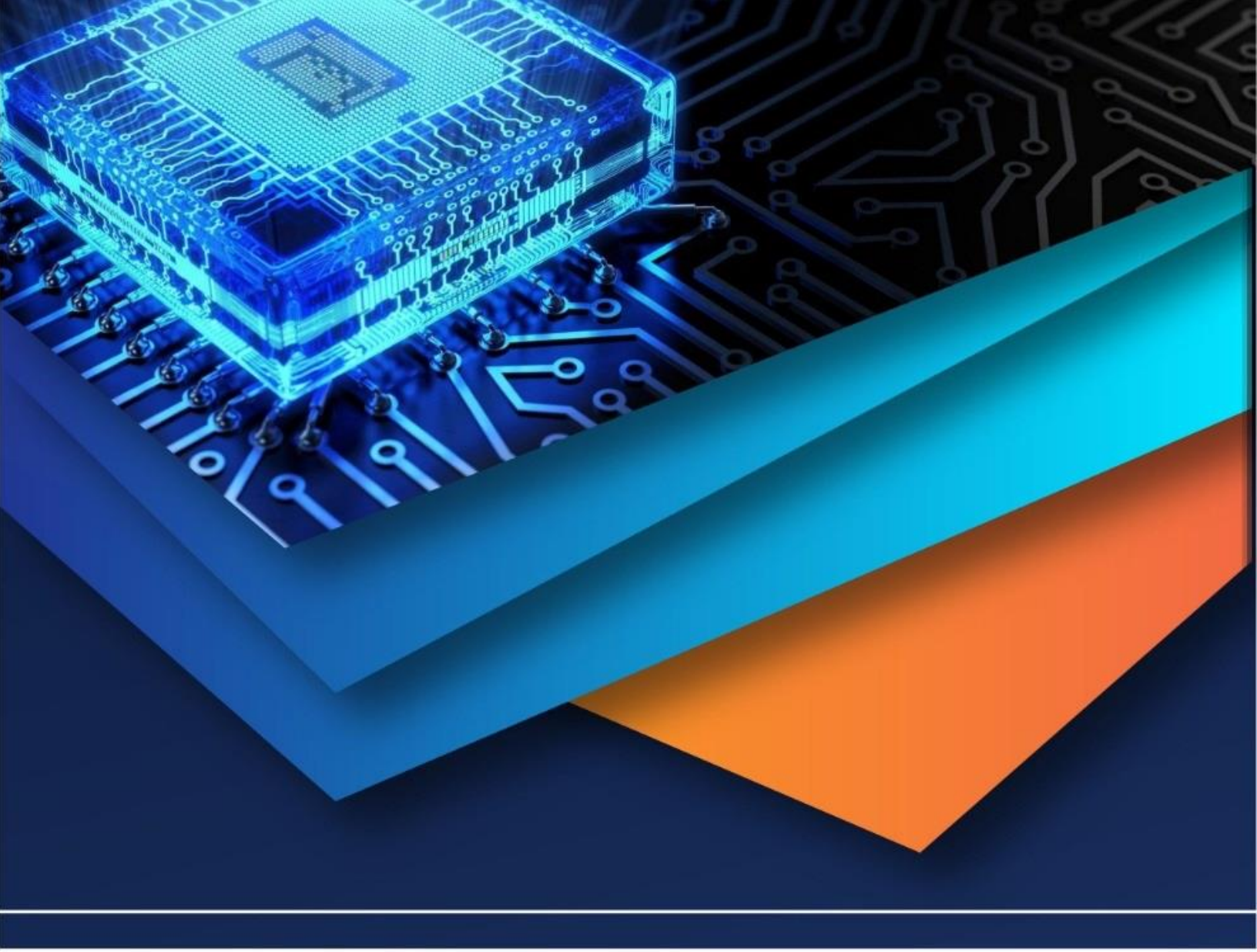
No project is ever complete without the guidance of those experts who have already traded this past before and hence become master of it and as a result, our leader. So we would like to take this opportunity to take all those individuals who have helped us in visualizing this project. We express our deep gratitude to our project guide Mrs. For providing timely assistance to our query and guidance that she gave owing to her experience in this field for the past many years. She had indeed been a lighthouse for us in this journey. We would also take this opportunity to thank our project coordinator Mr. For his guidance in selecting this project and also for providing us all this details on proper presentation of this project. We extend our sincerity appreciation to our entire Professor from K C College of Engineering Management Studies And Research .for their valuable inside and tip during the designing of the project. Their contributions have been valuable in so many ways that we find it difficult to acknowledge them individually. We are also grateful to our HOD Mrs. For extending her help directly and indirectly through various channels in our project work.

©IJRASET: All Rights are Reserved

REFERENCES

[1] ”Extractive Text Summarization Using Sentence Ranking” - J.N Madhuri, Ganesh Kumar.R., 2019.
[2] ”Automatic Text Summarization of News Articles” - Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R. B. Keskar, 2017.
[3] ”A Multi-document Summarization System Based On Genetic Algorithm” - Yan-xiang He, De-xi Liu, Dong-hong Ji3, Hua Yang, Chong Teng, 2006.
[4] ”Mining Topical Relevant Patterns for Multidocument Summarization” - Yutong Wu, Yang Gao, Yuefeng Li, Yue Xu, 2015.
[5] Study On Text Summarization Using Extractive Methods” - S.Mohamed Saleem, R.Krithiga, S.K.Rani, S.Celin Sindhya, 2015.
[6] ”A Summary On Extractive Text Summarization” - N. Moratanch , S. Chitrakala , 2017.
[7] ”An Integrated Approach to Web Document Summarization Using Semantic Similarity” - K .Vanisri, P. Ponnila, J. Jeejovetharaj, 2014.
[8] ”A survey on Real-Time Accumulative Short Text Summarization on Comment Streams” - N. Vijay Kumar, Dr.M.Janga Reddy, 2017.
[9] ”Automatic Keyword Extraction Using Textrank” - Papis Wongchaisuwat, 2019.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)