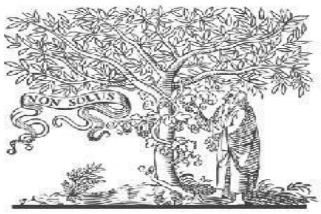
International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 14th Jul 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue 07

DOI: 10.48047/IJIEMR/V11/ISSUE 07/06

Title **AUTOMATIC TEXT SUMMARIZATION AND KEYWORD EXTRACTION USING NATURAL LANGUAGE PROCESSING**

Volume 11, ISSUE 07, Pages: 38-43

Paper Authors
Sai Pranavi Chitti, Neha Reddy Vantari,Shynitha Muthyam,
Aditi Vakeel,

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

AUTOMATIC TEXT SUMMARIZATION AND KEYWORD EXTRACTION USING NATURAL LANGUAGE PROCESSING

Sai Pranavi Chitti, BTech, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, pranavic2@gmail.com
Neha Reddy Vantari, BTech, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, nehareddyvantari8@gmail.com
Shynitha Muthyam, BTech, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, shynithavarma01@gmail.com
Aditi Vakeel, BTech, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, aditivakeel@gmail.com

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 the 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 a website link, removal of outliers and irrelevant information, emphasizing on the importance of particular data extracted from the website and creating a summary of the extracted data. For the 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.

Keywords:*Sign language, Hand gesture, Feature extraction, Gesture recognition.*

1. 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 undeclining 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 emphasize 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 a pattern in web pages and web documents by collecting and analyzing information to gain insight into 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).

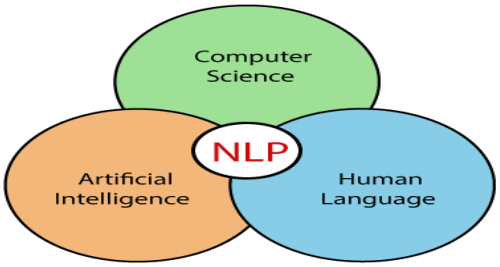


Fig.1: NLP structure

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A text summarization technique [1] based on a ranking algorithm which gives the important sentences which are collected to form an audio summary. There are two techniques for doing the text summarization: 1) Abstractive 2) Extractive This is done by using extractive text summarization which helps in important sentence selection using the linguistic or statistical features. PrakharSethi et.al.proposed a summarization algorithm[2] which uses lexical chains and thesaurus to generate a text summary of the news. It compared various summaries of the content and hence the scoring parameters. The sentence is given the score based on repeated nouns in the article. Thus, the sentence with the maximum score is considered in the summary of the article. Hua Yuan et.al.proposed a summarization method for tourism blogs[3] which focuses on removing the noise efficiently and was tested over a Chinese tourism blog. Yutong Wu et.al.Proposed a method with semantic and context-based analysis for multidocument summarization [4] which measures the important information covered in a sentence.Yan-Xiang He et.al. Proposed a method for multi-document summarization[5] which tries to reduce the sentence size and combines similar sentences to create new sentences. DragomirRadev et.al. proposed a centroid- based summarization method[6] for multiple documents which finds the most important words from the documents and then selects the sentences that largely represent the context of the document. N. Moratanch, S. Chitrakala et.al. surveyed various methods for text summarization.

2. PROPOSED SYSTEM

PrakharSethi et.al.proposed a summarization algorithm[2] which uses lexical chains and thesaurus to generate a text summary of the news. It compared various summaries of the content and hence the scoring parameters. The sentence is given the score based on repeated nouns in the article. Thus, the sentence with the maximum score is considered in the summary of the article. Hua Yuan et.al.proposed a summarization method for tourism blogs[3] which focuses on removing the noise efficiently and was tested over a Chinese tourism blog.

Disadvantages:

1. User suffering time.

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 the 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 a website link, removal of outliers and irrelevant information, emphasizing on the importance of particular data extracted from the website and creating a summary of the extracted data. For the selection of relevant information from the extracted data, it is necessary to use natural language processing.

Advantages:

1. 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.

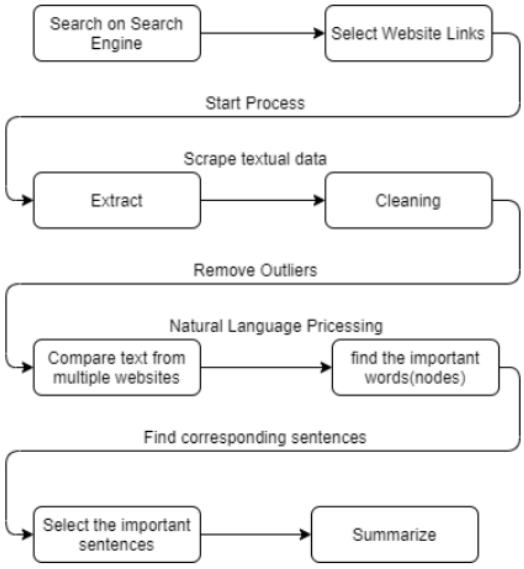


Fig.2: System architecture

The proposed system should be designed in such a manner that it will alleviate these limitations. The

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

problem definition of our proposed system is as follows:

• To extract textual data from any link chosen by the user and display its summary. Websites might contain irrelevant data and information which is not important. This project emphasizes on removing outliers to display a summary of only important data and also to remove redundant information gathered from multiple websites.

• To summarize textual data from multiple web pages at once. Existing summarizing software, browser extensions can either summarize multiple documents at once or a single web page. This project mainly focuses on summarizing multiple web pages and documents.

• To let the user decide if a summary of one link or multiple links is required and to give the choice to produce either a combined summary or a separate summary of the chosen website links.

• To implement keyword extraction as functionality to help the user to know the context of the textual information acquired from the link or document quickly.

• To implement multi-document summarization to generate summary out of documents present in different formats such as pdf and word.

The proposed system is built using the Flask framework and MySQL database at the back-end. Figure 2 shows the working diagram of the system for summarization. When the user enters a website link or uploads a document for summarization, the application starts the process and scrapes all the textual data from the link(s) or the document(s) using BeautifulSoup text mining python module. Cleaning stepinvolves removal of outliers and stop-words which is done using NLTK(Natural Language Toolkit) package available for python. Tokenization of sentences involves each word toget stored in an array where sentence acts as a node and each node has a weight assigned to it. Each sentence is a node for Textrank algorithm. Weight calculation is same as implemented in keyword extraction and is debriefed in implementation. HeapQ algorithm used in natural language processing retrieves the thirty percent that is three-tenth of the total number of sentences of the extracted text. This forms the summary of the extracted textual data. User has the option to choose whether the required summary should be a combined summary of all the documents or links given as an input or a separate summary of each input link or document is required by the user. In the case of combined summary, textual data from each link or document is combined first and it is summarized.

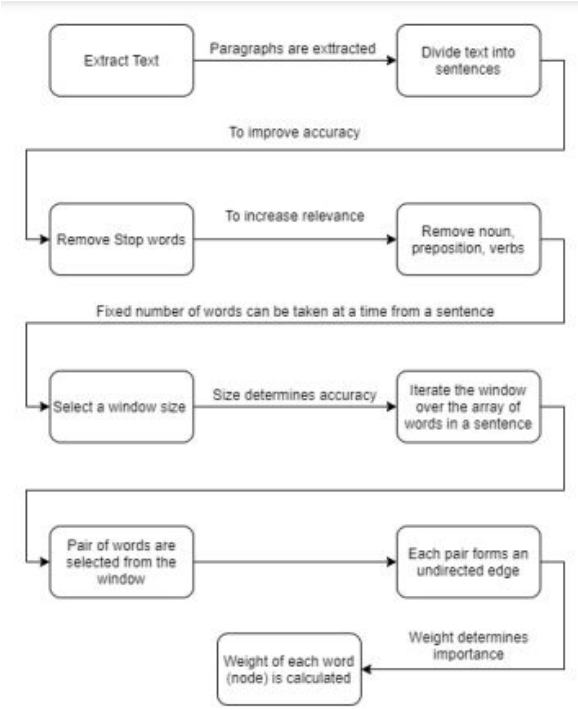


Fig.3: Keyword extraction

3. RELATED WORK

Extractive TextSummarization Using Sentence Ranking

Automatic Text summarization is the technique to identify the most useful and necessary information in a text. It has two approaches 1) Abstractive text summarization and 2) Extractive text summarization. An extractive text summarization means an important information or sentence are extracted from the given text file or original document. In this paper, a novel

statistical method to perform an extractive text summarization on single document is demonstrated. The method extraction of sentences, which gives the idea of the input text in a short form, is presented. Sentences are ranked by assigning weights and they are ranked based on their weights. Highly ranked sentences are extracted from the input document so it extracts important sentences which directs to a high-quality summary of the input document and store summary as audio.

Automatic Text Summarization of News Articles

Text Summarization has always been an area of active interest in the academia. In recent times, even though several techniques have being developed for automatic text summarization, efficiency is still a concern. Given the increase in size and number of documents available online, an efficient automatic news summarizer is the need of the hour. In this paper, we propose a technique of text summarization which focuses on the problem of identifying the most important portions of the text and producing coherent summaries. In our methodology, we donot require full semantic interpretation of the text, instead we create a summary using a model of topic progression in the text derived from lexical chains. We present an optimized and efficient algorithm to generate text summary using lexical chains and using the WordNet thesaurus. Further, we also overcome the limitations of the lexical chain approach to generate a good summary by implementing pronoun resolution and by suggesting new scoring techniques to leverage the structure of news articles.

Towards Summarizing Popular Information form massive Tourism Blogs

In this work, we propose a research method to summarize popular information from massive tourism blog data. First, we crawl blog contents from website and segment each of them into a semantic word vector separately. Then, we select the geographical terms in each word vector into a corresponding geographical term vector and present a new method to explore the hot tourism locations and, especially, their frequent sequential relations from a set of geographical term vectors. Third, we propose a novel word vector subdividing method to collect the local features for each hot location, and introduce the metric of max-confidence to identify the Things of Interest (ToI) associated to the location from the collected data. We illustrate the benefits of this approach by applying it to a Chinese online tourism blog data set. The experiment results show that the proposed method can be used to explore the hot locations, as well as their sequential relations and corresponding ToI, efficiently.

Mining Topical Relevant Patterns for Multidocument Summarization

Multi-document summarization addressing the problem of information overload has been widely utilized in the various real-world applications. Most of existing approaches adopt term-based representation for documents which limit the performance of multi-document summarization systems. In this paper, we proposed a novel pattern-based topic model (PBTMSum) for the task of the multi-document summarization. PBTMSum combining pattern mining techniques with LDA topic modelling could generate discriminative and semantic rich representations for topics and documents so that the most representative and non-redundant sentences can be selected to form a succinct and informative summary. Extensive experiments are conducted on the data of document understanding conference (DUC) 2007. The results prove the effectiveness and efficiency of our proposed approach.

4. IMPLEMENTATION

Proposed system emphasizes on providing:

Summarization of not only a single website link but multiple links through Textrank algorithm. User may input single or multiple website links and the proposed system visits the link(s) and scrapes all the textual data from the website(s).

Summarization of not only single but multiple documents(pdf file or word file) through Textrank algorithm. User may upload single or multiple documents which are temporarily stored until the proposed system scrapes the data from the documents.

Keyword extraction is also implemented with the help of Textrank algorithm. User may input a website

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

link or upload a document and keywords will be extracted. A total number of keywords extracted depends upon the size of the document.

The proposed system also gives the flexibility to the user to input direct text in the text area through which keywords can be extracted or text can be summarized based on the user's choice.

5. EXPERIMENTAL RESULTS

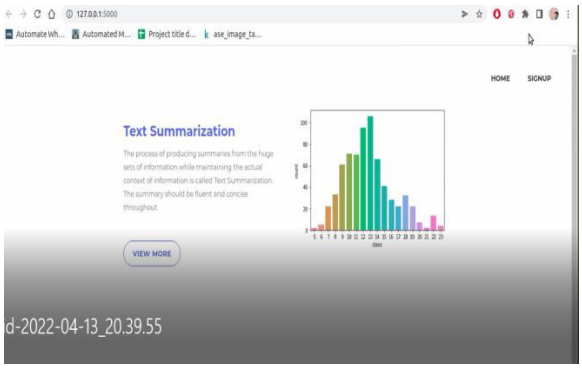


Fig.4:Home screen

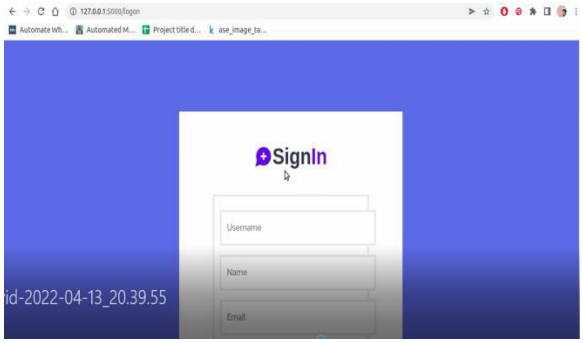


Fig.5: Signup

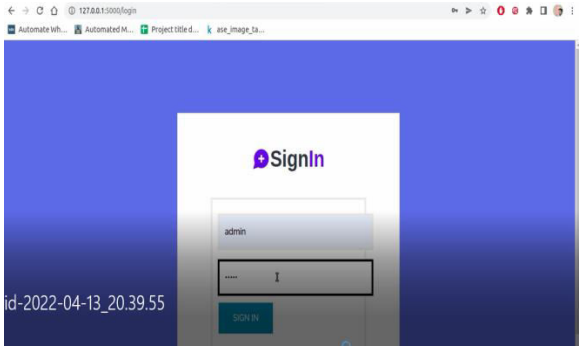


Fig.6: Signin

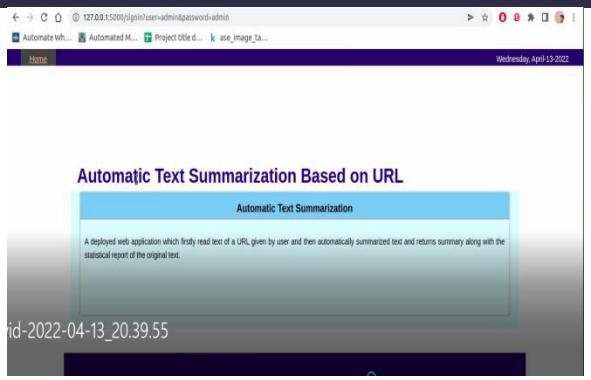


Fig.7: Main screen

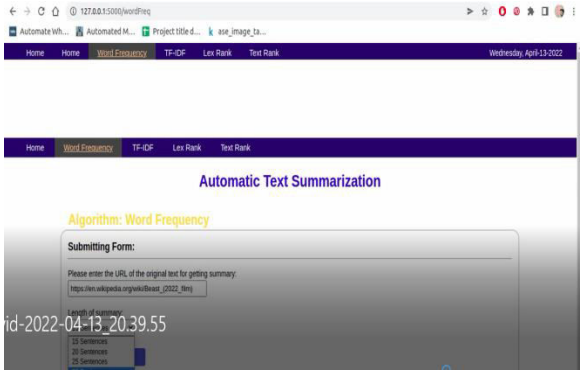


Fig.8: Input screen

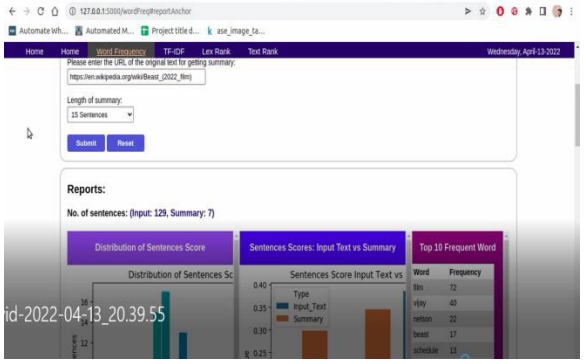


Fig.9: output

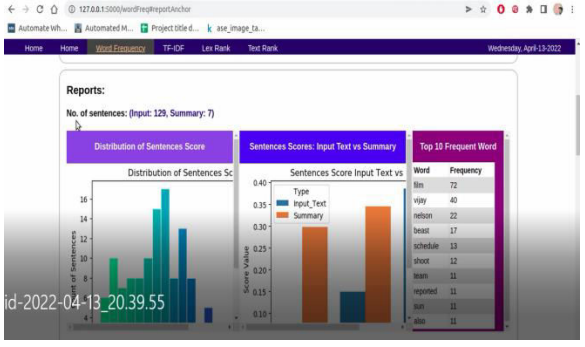


Fig.10: output

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

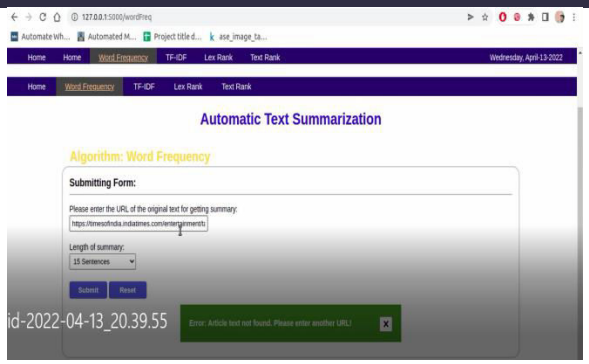


Fig.11: Output

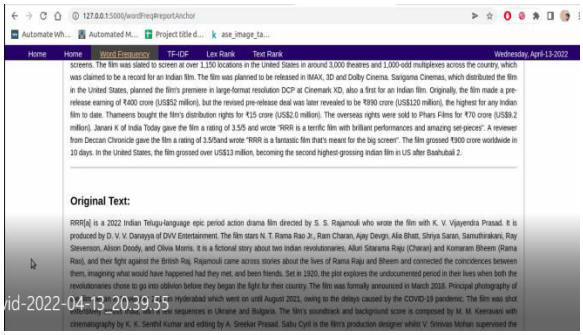


Fig.12:Output

6. 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 the 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 the 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 obtain summary of blog posts, any other post from social media sites or particular textual data which they want to summarize.

REFERENCES

[1] "Extractive TextSummarization Using Sentence Ranking" - J.N Madhurt, Ganesh Kumar.R., 2019.

[2] "Automatic Text Summarization of News Articles" - PrakharSethi, Sameer Sonawane, SaumitraKhanwalker, R. B. Keskar, 2017.

[3] "Towards Summarizing Popular Information form massive Tourism Blogs" -Hua Yuan,HualinXu,YuQian,Kia Ye, 2016.

[4] "Mining Topical Relevant Patterns for Multidocument Summarization" - Yutong Wu, Yang Gao, Yuefeng Li, Yue Xu, 2015.

[5] "A Multi-document Summarization System Based On Genetic Algorithm" - Yan-xiang He, De-xi Liu, Dong-hong Ji3, Hua Yang, Chong Teng, 2006.

[6] "A Scalable Multi-document Centroid-based Summarizer" - DragomirRadev , Timothy Allison, Matthew Craig , StankoDimitrov , Omer Kareem , Michael Topper , Adam Winkel , and Jin Y, 2004.

[7] "A Summary On Extractive TextSummarization" - N. Moratanch , S. Chitrakala , 2017.

[8] "Review Paper On Extractive TextSummarization" - ArpitaSahoo, Dr. Ajit Kumar Nayak, 2018

[9] " Study On Text Summarization Using Extractive Methods" - S.MohamedSaleem, R.Krithiga, S.K.Rani, S.CelinSindhya, 2015.

[10] "An Integrated Approach to Web Document Summarization Using Semantic Similarity" - K .Vanisri, P. Ponnila, J. Jeejovetharaj, 2014.