IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

Tool to Summerize Text

Sneha Wankar[1], Gudiya Prasad[2], Achal Ragit[3], Achal Waghmare[4], Prof. Anand Donald[5]
Students, Department of Computer Science and Engineering[1,2,3,4]
Guide, Department of Computer Science and Engineering[5]
Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur, India

Abstract: *Text summarization is the process of making a synopsis from a given text document while keeping the important information and meaning of it. Automatic summarization has become an essential method for accurately locating significant information in vast amounts of text in a short amount of time and with minimal effort. In this project, we propose to implement a web application that can summarize a text or a Wikipedia link.*
We have additionally been given an opportunity to compare different methods of summarization. Problem Statement - The tremendous abundance of material available on the internet has produced an odd paradox: people are immersed in information, yet they are yearning for wisdom. It is tough to keep up with the internet's daily production of billions of articles. Is there a method to absorb information more effectively in this case without increasing reading time?
We are proposing for the above problem a Text Summarizer web app using NLP and NLTK libraries.

Keywords: Automatic summarization, Extractive, Natural Language Processing, frequency-based

I. INTRODUCTION

Text summarization simply means the technique ofshortening long pieces of text. The intention is to create a systematic and fluent summary having only the main points outlined in the document. With sucha big amount of data circulating in the digital space, there is need to develop machine learning algorithms that can automatically shorten longer texts and deliver accurate summaries that can fluently pass the intended messages in very less time.

Fundamentals : The following terms will be used for the discussion of automatic text summarization :

- **Extractive text summarization :** Extractive technique involves retaining the key phrases from the source document and combining them to make a summary.
- The extraction is made according to the defined metric without making any changes to the texts.
- **Abstractive text summarization :** This algorithm creates new phrases and sentences that represent the most useful information of the original text. Abstractive text summarization overcomes the grammar inconsistencies of the extractive method and provides a much condensed summary.
- **Text preprocessing:** transforms the text into a more digestible form so that machine learning algorithms can perform better.
- **Tokenization :** Tokenization is about splitting strings of text into smaller pieces, or tokens . Paragraphs can be tokenized into sentences (sentence tokenization) and sentences can be tokenized into words(word tokenization).
- **Stemming :** It is the process of reducing a word to its word stem that affixes to suffixes and prefixesor to the roots of words
- **Lemmatization :** Lemmatization converts a word into its lemma (root form). It usually refers todoing things properly with the use of a vocabulary and morphological analysis of words. Lemmatization is a technique in NLP that breaks aword in its root form, within the boundaries of linguistics.
- **TF-IDF** (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying twometrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

II. LITERATURE REVIEW

Introduction :

Literature surveys provide brief overviews or a summary of the current research on topics. Literature surveys are used in ensuring that the used experiments, methodologies and experiments offer reliability and validity in the research being conducted. They are useful invalidating or providing proof and also provides a base of moving a research idea forward on what researchers have done and exciting avenues that it opens for investigation during future work in the field. In this literature survey research papers and articles on automatic text summarization were referred through for the project.

Literature Review:

The following research papers and reports were used for the topic of Text Summarization System for English Language:

Automatic Text Summarization Using Natural Language Processing:

Authors Pratibha Devihosur, Naseer R implemented an automatic text summarization mechanism based on an unsupervised learning system. The significance of the generated summary was assessed with the assistance of Simplified Lesk calculation along with an online semantic lexicon WordNet Based on their evaluation the algorithm provides best summarized outcome ranging from 25-50 percent with respect to the source data. In this project they also focussed on ambiguous words because a specific word may have distinctive significance in various setting. Hence they tried to inculcate the principle of word sense disambiguation to decide the right feeling of a word utilized as part of a specific setting.

Text Summarization using Natural Language Processing:

In this paper authors Ankit Kumar, Zixin Luo and Ming Xu created an end to end web application which can take an article as input and generate a summary. The model was trained using deep learning approach and trained on Juniper's datasets Juniper is a corporate organization that develops and markets networking devices. In order to provide a better customer experience, Juniper Networks maintains large datasets of articles wherein each of these articles can be long and verbose. Hence these datasets were used to train the text summarization model. The model built used abstractive summarization technique and significantly generated excellent human readable sentences from given inputs. However, it did not always generate summaries capturing all the important information in the input documents.

Text Summarization Techniques: A Brief Survey:

In this survey Mehdi Allahyari, Elizabeth D. Trippe, Saeid Safaei and others study the main approaches to automatic text summarization and also review the different processes for summarization and describe the effectiveness and shortcomings of different methods. Topics like the impact of context in summarization and semantic analysis are also mentioned.

NLP Based Text Summarization Using Semantic Analysis:

In this project paper authors Harsh Desai, Dhairya Pawar, Geet Agrawal reviewed the different methods for text summarization and provided a novel technique generating the summarization of domain specific text by using Semantic Analysis for text summarization.

Text Summarization: An Overview:

In this research paper author Samrat Babar provides an analysis about the meaning of text summarization in natural language processing and their types along with the technical and mathematical analysis of text summarization in detail. This paper basically is a documentation for all the information required to study as well refer through on the topic of automatic text summarization.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

III. SYSTEM REQUIREMENT SOFTWARE REQUIREMENT

Software-Python IDE and its libraries, Jupyter-Notebook,GoogleColab, VScode.

HARDWARE DESCRIPTION:

Modern Operating System:
Windows 7 or 10
Mac OS X 10.11 or higher, 64-bit
4 GB RAM
x86 64-bit CPU (Intel / AMDarchitecture

ARCHITECTURE OF PROJECT

The architecture for a text summarization tool with semanticanalysis typically involves several key components. Here's a high-level overview::

Input Module:

Text Ingestion: Accepts the input text data that needs to besummarized.

Preprocessing:

Tokenization: Breaks down the text into individual words ortokens.
Cleaning: Removes irrelevant characters, stop words, and performs basic text cleaning operations.

Semantic Analysis:

Named Entity Recognition (NER): Identifies entities (suchas people, organizations, locations) in the text.
Part-of-Speech (POS) Tagging: Assigns grammatical partsof speech to each word.
Dependency Parsing: Analyzes the grammatical structure tounderstand relationships between words.

Semantic Representation:

Word Embeddings: Converts words into numerical vectors,capturing semantic relationships.
Sentence Embeddings: Represents entire sentences in asemantic space.

Information Extraction:

Key phrase Extraction: Identifies important phrases orkeywords.
Entity Extraction: Extracts relevant entities from the text.

Summarization Model:

Abstractive or Extractive Summarization: Chooses between generating a summary in its own words (abstractive) or selecting important sentences from the original text (extractive).

Post-processing:

Fluent Summary Generation: Ensures that the generatedsummary is coherent and grammatically correct.
Length Control: Manages the length of the summary.

Output Module:

Generated Summary: Provides the final summarized output.

Evaluation Module:

Quality Assessment: Assesses the quality of the generated summary using metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

User Interface (Optional):

Integration with Applications:

If applicable, integrates with user interfaces or other applications where summaries are needed.

Feedback Loop (Optional):

User Feedback: Incorporates user feedback tocontinuously improve the summarization model. It's important to note that the specific architecture can vary based on the chosen approach (abstractive or extractive), the complexity of semantic analysis,and the desired use cases. Additionally, machinelearning frameworks and libraries likeTensorFlow or PyTorch may be utilized for building andtraining models in the semantic analysis andsummarization stages

IV. ADVANTAGES & DISADVANTAGES

Advantages:

- Time Efficiency: Summarization tools in NLP can process large volumes of text quickly, saving time compared to manual summarization.
- Consistency: Automated tools provide consistent results, reducing variability that may occur with human summarizers.
- Handling Large Datasets: NLP summarization tools excel at handling large datasets, making them valuable for processing extensive amounts of information.
- Objective Output: Automation reduces the impact of subjective biases that might be present in human summarization.

Disadvantages:

- Loss of Nuance: Automated tools may struggle to capture the subtle nuances and context that a human summarizer can grasp.
- Complexity Handling: Handling complex sentences, ambiguous language, or diverse writing styles can be challenging forautomated summarization systems.
- Domain Specificity: Some tools might not perform well in certain specialized domains where domain-specific knowledge is crucial.
- Evaluation Challenges: Assessing the quality of an automated summary can be challenging, as it often depends on the specific requirements of the user.

In summary, while NLP summarization tools offer efficiency and consistency, theymay lack the nuanced understanding that human summarizers bring, and their performance can vary based on the complexity and specificity of the text.

V. CONCLUSION

With the ever-growing text data, text summarization seems to have the potential for reducing the reading time by showing summaries of the text documents that capture the key points in the original documents. Automatic text summarization is an old challenge but the current research direction diverts towards emerging trends in biomedicine, product review, education domains, emails, and blogs.

This is due to the fact that there is information overload in these areas, especially on the World Wide Web. Automated summarization is an important area in NLP (Natural Language Processing) research. It consists of automatically creating a summary of one or more texts. The purpose of extractive document summarization is to automatically select several indicative sentences, passages, or paragraphs from the original document. Text summarization approaches based on NLP have, to an extent, succeeded in making an effective summary of a document.

Both extractive and abstractive methods have been researched. Most summarization techniques are based on extractive methods. As with time the internet is growing at a very fast rate and with it data and information are also increasing. It will be difficult for humans to summarize large amounts of data. Thus there is a need for automatic text summarization because of this huge amount of data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

BIBLIOGRAPHY

[1] G. Erkan, D. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research 22, 2004.

[2] Selvani Deepthi Kavala, Dr. Radhika Y, "Extractive Text Summarization Using Modified Sighing and Sentence Symmetric Feature Methods", I.J. Modern Education and Computer Science, 2015.

[3] H.P.Luhn, "The Automatic Creation of Literature Abstracts". IBM Journal of Research and Development, 1958.

[4] H.P. Edmundson, "New Methods in Automatic Extracting", Journal of the Association for Computing Machinery, April 1969.

[5] A.Das, M.Marko, A.Probst, M.A.Portal, C.Gershenson —Neural Net Model For Featured Word Extraction‖, 2002.

[6] Jagadeesh J, Prasad Pingali, Vasudeva Varma, "Sentence Extraction based single Document Summarization", Workshop on Document Summarization, 19th and 20th March 2005, IIIT Allahabad.

[7] Arman Kiani B, M. R. Akbarzadeh —Automatic Text Summarization Using: Hybrid Fuzzy GA-GP‖, IEEE International Conference on Fuzzy Systems. July 16-21, 2006.

[8] R. Mihalcea, and P. Tarau, "Textrank: Bringing order into texts,". In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004.

[9] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to- sequence RNNs and beyond,". In Computational Natural Language Learning, 2016