JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TEXT SUMMARIZATION USING NLP

Aameyaa Dhumal, Sneha Joshi, Piyush Sharma, Shantanu Mane, Prof. Pooja Wale

Department of Computer Engineering, Indira College of Engineering and Management, Parandwadi, Pune

ABSTRACT

Data mining is a field that has seen significant evolution in recent years as a result of enormous breakthroughs in software and hardware technologies. As technology advances, more types of data become available, which is especially useful in the case of text data. The software and hardware platforms that power social networks and the internet have aided in the rapid production of massive data stores. Structured data is typically maintained by a database system, whereas text data is typically managed by a search engine due to the lack of structures. With the help of a keyword query, the search engine allows the online user to find the essential information from the gathered works. Text summary is the practice of extracting the most relevant information from a source document in order to create an abridged version for a specific job.

Keywords: *NLP(natural language processing), extractive, abstractive, encoding, decoding*

1. INTRODUCTION

In today's world, a massive amount of data is being generated on the internet every day. As a result, a better mechanism for extracting important information quickly and effectively is required. Text summarization is one of the strategies for finding the most significant and meaningful information in a document or set of linked documents and compressing it into a shorter version while maintaining the overall meaning. It cuts down on the time it takes to read an entire page and solves the space issue that comes with keeping big amounts of data. There are two techniques to Automatic Text Summarization. 1) Abstractive text summarising and 2) Extractive text summarization are two types of text summarization. The term "extractive text summarization" refers to the extraction of key information or sentences from a text file or source document. An extractive text summarizing approach selects interesting informative sentences based on linguistic or statistical criteria. An abstractive text summary will attempt to comprehend the input or original file and re-generate the output in a few words by recognizing the input file's core concept. That extracted text has been mentioned in a number of academic studies. Text summarization is widely utilized in a variety of sectors, including science, medicine, law, and engineering. Researchers have concentrated on creating summaries of doctor's prescriptions, which have proven to be extremely beneficial to patients. Long news stories have also been summarized so that readers can get a lot of information on a variety of topics in a short amount of time. For the past five years, we've discussed the numerous strategies utilized in text summarizing in this document. Machine learning (ML), neural networks (NNs), reinforcement learning, sequence to sequence modeling, and fuzzy logic were determined to be the most popular methods. Similarly, for the goal of text summarization, numerous optimization methods have been applied to maximize the specified objective function.

2. LITERATURE SURVEY

[1] Text mining and text summarization have a lot more in common than you would think. Established summarizing systems should be built and classified based on the kind of input text, based on the differences in requirements summary with respect to input text. The concept of text mining and its relationship with text summarization are discussed initially in this work. Following that, a study of some of the summary methodologies and their key parameters for extracting dominant phrases was conducted, as well as the main stages of the summarizing process and the most critical extraction criteria. Finally, the most basic proposed evaluation methods are taken into account.

[2] Text summarizing techniques have been altered by the use of linguistics to advanced machine learning models; this work investigates summarization approaches as well as contemporary state-of-the-art models for single and multi-document summary. This survey aims to conduct a comprehensive investigation using machine learning, modern graph and evolutionary based methods, from feature representation to sentence selection and summary creation. The whole study will assist researchers in properly handling enormous amounts of data while developing effective Natural Language Processing apps. Finally, this research identifies common abstractive mechanisms and observations that will be useful in the research.

[3] A novel statistical strategy for extracting text summarization on a single document is demonstrated in this research. The method of sentence extraction is provided, which gives the idea of the input text in a concise way. Sentences are graded by assigning weights to them and then ranking them according to those weights. Highly scored sentences are taken from the input document, allowing it to extract essential sentences that lead to a high-quality summary of the input material, which may then be saved as audio.

[4] This paper discusses numerous methods for generating summaries of large books. Various articles have been examined for distinct text summarizing approaches that have been employed in the past. Abstractive (ABS) or Extractive (EXT) summaries of text documents are the most common outputs of the methods presented in this study. Techniques for query-based summarization are also presented. The majority of the work is devoted to the structured and semantic approaches to text document summarization. The CNN corpus, DUC2000, single and multiple text documents, and other datasets were utilized to test the summaries produced by these models. We investigated these strategies, as well as their trends, accomplishments, previous work, and future potential in text summarization and other domains.

[5] The suggested system primarily focuses on scraping data from websites and presenting a summary as well as keywords from the information taken from multiple websites, allowing the user to choose their preferred website. Starting with data extraction from a website link, removing outliers and irrelevant information, emphasizing the importance of particular data extracted from the website, and creating a summary of the extracted data, the proposed system for text summarization and keyword extraction goes through a series of steps. Natural language processing is required for the selection of relevant information from the extracted data.

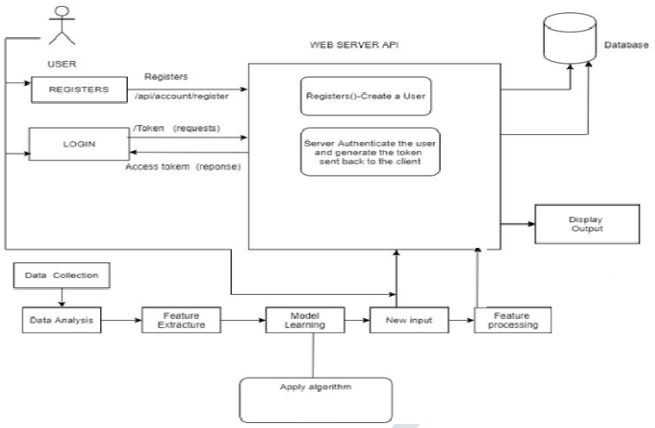
The suggested software assists users in reducing their browsing time by providing a summary compiled from several website links and documents.

3. SYSTEM OVERVIEW

A summary is a text created from one or more texts that provides a major amount of the information contained in the original text and is less than half the length of the original text .Text summarizing is the process of extracting the most significant information from a source (or sources) in order to create an abridged version for a certain user (or users) and task (or tasks). Automatic Text Summarization is what we term it when this is done by a computer, i.e. automatically. Despite the fact that text summarizing has typically concentrated on text input, multimedia material such as photos, video, or music, as well as on-line information or hypertexts, can also be used as input to the summary process. Furthermore, we can discuss summarizing a single paper or a number of

them. The technique is called as Multi-document Summarization (MDS) in this situation, and the source documents might be in a single language (monolingual) or multiple languages (translingual or multilingual).

4. SYSTEM ARCHITECTURE



GUI

The GUI (Graphical User Interface) is the user interface via which the user will login and submit his text. The GUI provides a user interface for interacting with the database. It serves as a connector and communicator, connecting the database and facilitating data flow between the GUI and the database.

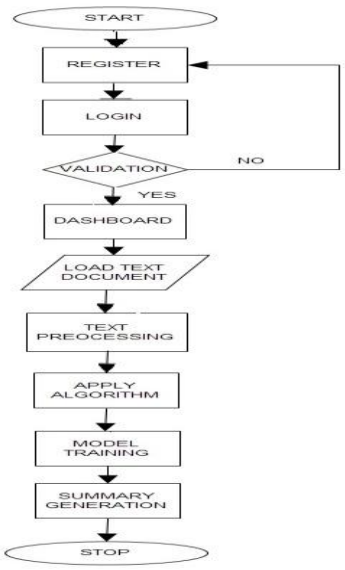
Processing block:

The processing block is where our project's actual processing takes place. This block connects the GUI to the database, acting as both a connector and a communicator. It connects the database and facilitates data flow between the GUI and the database. Its primary goal is to process information from text of user in order to save it in a structured format and database. After storing this data, the system will generate an output via a web application.

Database:

The database responsible for data storage. This layer holds all of the required for the project's processing. The information received from the user in the text format.

5. FLOWCHART



Text summarizing presents a number of issues, including text identification, interpretation, and summary generation, as well as analysis of the resulting summary. Identifying important phrases in the document and exploiting them to uncover relevant information to add in the summary are critical jobs in extraction-based summarizing.

The steps involved in creating the text summary are as follows:Data cleansing, which includes the removal of special characters, numeric values, stop words, and punctuation. Tokenization – Tokenization is the process of creating tokens (Word tokens and Sentence tokens). Determine the frequency of each word. For each sentence, calculate the weighted frequency.Create a summary by selecting the top-weighted sentences.

7. ALGORITHMS

SUMY-

Sumy can generate an extracted summary. That is, it tries to extract the most important sentences from the document(s) and combine them into a shorter text. Another method is to make an abstractive summary, however this requires understanding the topic and creating new condensed text from it. Sumy's existing capabilities do not allow for this.

NLTK-

Natural language processing (NLP) is a field that focuses on making computer algorithms understand natural human language. Natural

Language Toolkit, or NLTK, is a Python package that can be used for NLP. Unstructured data with human-readable text makes up a large portion of the data you could be examining. You must first preprocess the data before you can analyse it programmatically.

SPACY-

spaCy is an open-source software library designed in the Python and Cython programming languages for advanced natural language processing. Matthew Honnibal and Ines Montani, the founders of the software business Explosion, are the main developers of the library, which is released under the MIT licence. Unlike NLTK, which is widely used for teaching and research, spaCy concentrates on creating production-ready software. SpaCy also supports deep learning workflows, which allow statistical models created using popular machine learning frameworks like TensorFlow to be connected.

GENSIM-

Gensim, which stands for "Generate Similar," is a well-known open source natural language processing (NLP) framework for unsupervised topic modelling. It performs a variety of complex tasks using top academic models and contemporary statistical machine learning, including as Corpora is a tool for creating document or word vectors. Gensim, written in Python and Cython, is designed to handle enormous text collections using data streaming and incremental online algorithms, in addition to the above hard tasks. This distinguishes it from machine learning software that focuses solely on in-memory processing.

6. CONCLUSION:

Automatic text summarization is a method that allows individuals to achieve a quantum leap in productivity by reducing the sheer volume of information they encounter on a daily basis. This not only allows people to reduce the amount of reading they must do, but it also allows them to read and comprehend previously neglected literary works. Text summarising is a rapidly growing area, with specialized tools being created to handle increasingly concentrated summary jobs. Users are expanding the use case of this technology as open-source software and word embedding packages become more widely available.

.REFERENCES

[1]Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean," Dist ribut ed Representations of Words and Phrases and their Compositionality," arXiv:1310.4546v1 [cs.CL], 2013.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representat ions in Vector Space," arXiv:1301.3781v3 [cs.CL], 2013.

[3] Shi Ziyan " The Design and Implementation of Domain-specific Text Summarizat ion System based on Co-reference Resolution Algorithm" 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery.

[4] Paul Gigioli, Nikhita Sagar, Anand Rao, Joseph Voyles" Domain-Aware Abst ract ive Text Summarizat ion for Medical Document s" published during 2018 IEEE BIBM.

[5] Niladri Chat terjee, Amol Mittal and Shubham Goyal's " Single Document Ext ract ive Text Summarizat ion Using Genetic Algorithms" (2012)

[6] Amol Tandel, Brijesh Modi, Priyasha Gupta, Shreya Wagle and Sujata Khedkar's "Mult i-document text summarization - A survey" 2016 Internat ional Conference on Data Mining and Advanced Comput ing (SAPIENCE).

[7] Adit ya Jain, Divij Bhatia, Manish K T hakur's "Extractive Text Summarizat ion using Word Vect or Embedding"(2017).

[8] Canasai Kruengkrai and Chuleerat Jaruskulchai "Generic T ext Summarizat ion Using Local and Global Properties of Sentences"(2003).

[9] Nithin Raphal, Hemanta Duwarah and Philemon Daniel "Survey on Abstractive Text Summarization" 2018 International Conference on Communicat ion and Signal Processing (ICCSP).

[10] Yang Wei and Yang Zhizhuo "Query based Summarization using topic background knowledge" 2017 13t h International Conference on Natural Computat ion, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD).