End-End Summarization Tool using NLP

A Major Project Synopsis Submitted To

Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal Towards Partial Fulfillment for the Award of Bachelor of Engineering (Information Technology)

Under the Supervision of

Prof. Kapil Sahu

Submitted By
PRAYAG GAVSHINDE
RAGHAV AGRAWAL
RISHABH RATHORE
RISHI SOMANI
SAMBHAV JAIN

Department of Information Technology
Acropolis Institute of Technology & Research, Indore

1. Abstract

Machine learning is a branch of artificial intelligence concerned with the creation and study of systems that can learn from data. In this project, Automatic text summarization is summarizing the given paragraph using natural language processing and machine learning. There has been an explosion in the amount of text data from a variety of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. In this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods.

.

2. Introduction of the Project:

In the modern Internet age, textual data is ever increasing. Need some way to condense this data while preserving the information and meaning. We need to summarize textual data for that. Text summarization is the process of automatically generating natural language summaries from an input document while retaining the important points. It would help in easy and fast retrieval of information. There are two prominent types of summarization algorithms.

- **Extractive summarization -** systems form summaries by copying parts of the source text through some measure of importance and then combine those parts/sentences to render a summary. The importance of a sentence is based on linguistic and statistical features.
- **Abstractive summarization -** systems generate new phrases, possibly rephrasing or using words that were not in the original text. Naturally abstractive approaches are harder. For a perfect abstractive summary, the model has to first truly understand the document and then try to express that understanding, in short, possibly using new words and phrases. Much harder than extracting. Has complex capabilities like generalization, paraphrasing, and incorporating real-world knowledge. The majority of the work has traditionally focused on extractive approaches due to the ease of defining hard-coded rules to select important sentences rather than generate new ones. Also, it promises grammatically correct and coherent

summaries. But they often don't summarize long and complex texts well as they are very restrictive.

Possible current uses of summarization that the Project address

- People need to learn much from texts. But they tend to want to spend less time while doing this.
- It aims to solve this problem by supplying them the summaries of the text from which they want to gain information.
- The goals of this project are that these summaries will be as important as possible in the aspect of the texts' intention.
- Supplying the user, with a smooth and clear interface.
- Configuring a fast replying server system.

3. Objective:

The objective of the project is to understand the concepts of natural language processing and create a tool for text summarization. The concern in automatic summarization is increasing broadly so the manual work is removed. The project concentrates on creating a tool that automatically summarizes the document.

- Automatic text summarizing by providing top sentences with the highest score in the document to save time.
- Helping users to provide a proper caption or description for the image

4. Scope:

- It provides sensitivity to the client and adapts well to future

summarization techniques.

- It considers a complete document or article instead of fixed-length content.
- It increases security and control.
- It reduces IT Administration costs.

5. Study of Existing System:

No.	Existing system/website/software	Features	Disadvantages	Limitations/Gaps
1.	TLDRthis (website)	Summarizing any article through URL or pasting text	summarizing documents by uploading them from the local system is not possible.	cannot summarize pdfs. only applicable for articles or blogs through links.
2.	Resoomer(Website)	Summarize text for everyone By copy-paste content. Options for downloading generated summary in pdf of any doctype.	summarizing documents by uploading them from the local system is not possible.	cannot summarize pdfs. only applicable for text content.

6. Project Description:

The project is well suited for extracting summaries of each kind. The major functionality that the project enjoys is.

- **Document Summarization**: Any pdf, doc, or text files like research papers, chapter notes, or anything if you upload it will find the best sentences with the highest dominating score and generate a short and sweet summary.
- **The article, Blog summary through URL**: If you want to read a summary of any online article available on any website, just provide a link. You will get important details like the author, date, and article summary.
- **Topic Modelling**: With the advent of big data and Machine Learning along with Natural Language Processing, it has become the need for an hour to extract a certain topic or a collection of topics that the document is about. Think when you

have to analyze or go through thousands of documents and categorize them under 10 – 15 buckets. How tedious will it be? Thanks to Topic Modeling where instead of manually going through numerous documents, with the help of Natural Language Processing and Text Mining, each document can be categorized under a certain topic.

- **Image Captioning**: Sometimes It feels tedious to describe any Image. We will integrate a model which will describe and provide a perfect description of an Image.

ALGORITHM SPECIFICATION:-

INPUT:- Large Chunks of Text/articles/Web Page URLs

OUTPUT:- Summarized Text and Audio File of Summarized Text

STEP1:- Concatenate all the contained in the articles

STEP2:- Entire concatenated Text is Split into individual Sentences.

STEP3:- Find Vector Representation(Word Embeddings) for each since several are sentences by using Glove Algorithm.

STEP4:- Similarities Between sentence vectors are calculated and stored in a matrix using Cosine Similarity

STEP5:- Convert the similarity matrix into Graph using Page Rank Algorithm.

STEP6:- Find a certain number of top-ranked sentences using the page rank algorithm to form a summary.

STEP7:- Convert the summarized text into an audio file using Google Text To Speech API.

7. Resources and Limitations :

FUNCTIONAL REQUIREMENTS:-

- Large Chunks of Text.
- Web Page Urls.
- Text Rank Algorithm.
- Word2vec Representation(Glove Algorithm)
- Similarity Matrix(Cosine Similarity).

NON FUNCTIONAL REQUIREMENTS:-

- Reliability.
- Performance
- Usability.
- Platform independent
- Supportability.

8. Conclusion :

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 a number of 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.

9. Bibliography:

[1].Anagnostopoulos, A., Broder, A.Z., Gabrilovich, E., Josifovski, V., Riedel, L.: Just-in-time contextual advertising. In: CIKM ’07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 331–340. ACM, New York, NY, USA (2007). DOI http://doi.acm.org/10.1145/1321440.1321488

[2] A new sentence similarity measure and sentence based extractive technique on Automatic Text summarization

[3] Ta Hahn, Udo, and Inderjeet Mani. "The challenges of automatic summarization." Computer 33.11 (2000): 29-36. 3.

[4]Vishal Gupta, Gurpreet Singh Lehal,” A Survey of Text Summarization Extractive Techniques.”JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010

[5] Josef Steinberger, KarelJežek, “Using Latent Semantic Analysis In-Text Summarization and Summary Evaluation”, Department of Computer Science and Engineering, UniverzitníCZ-306 14 Plzeň.

[6] Dipanjan Das, Andre F.T. Martins, “A Survey on Automatic Text Summarization”, Language Technologies Institute, Carnegie Mellon University, November 2007.