YouTube Transcript Summarization Using LSA

Mr.P.J.V.G.Prakasa Rao¹, S.Hyndhavi², T.Varshitha³, G.Sai Maha;akshmi⁴, T.Lavanya⁵

Associate Professor¹ CSE, LENDI INSTITUTE OF ENGINEERING AND TECHNOLOGY
Student²,³,⁴ CSE, LENDI INSTITUTE OF ENGINEERING AND TECHNOLOGY

ABSTRACT: YouTube Transcript Summarization extension combines machine learning and web development. In this project, we provide a reduced version of a video's documentation. YouTube transcript summarization using Latent Semantic Analysis (LSA) is a technique used to automatically generate concise and informative summaries of video content. LSA is a statistical technique that analyzes the relationships between words in a large corpus of text and identifies underlying latent semantic concepts. By applying LSA to a YouTube video transcript, it is possible to identify the most important concepts and extract key sentences that capture the essence of the video. The resulting summary can be useful for viewers who want to quickly understand the main points of a video without having to watch the entire content. In this abstract, we provide a brief overview of the process of YouTube transcript summarization using LSA and discuss its potential applications.

Keywords: Text Summarization, NLP, Spacy, NLTK, Chrome Extension, LSA

I. INTRODUCTION:

Every day, a sizable number of videos are produced and distributed online. In 2022, there will be an average of 2.47 billion YouTube users, a significant rise each year. Three hundred hours of videos are posted to YouTube every single minute. According to a Google survey, nearly one-third of YouTube users in India spend more than 48 hours each month on the platform. It is really challenging to watch videos that are longer than you anticipated. If we can locate the pertinent information we are looking for, sometimes our efforts could be more effective. Searching for videos that include the information we need can be frustrating and time-consuming due to the abundance of content on the Internet. The second most visited website worldwide is YouTube.

Python includes several packages that are highly beneficial for various purposes. Currently, a transcript API in the Python library makes it simpler to retrieve YouTube material. Using this benefit, we may easily retrieve and summarize the video's transcripts. Two of the most well-liked Natural Language Processing (NLP) tools for Python are NLTK and spaCy. Academics and researchers developed NLTK as a tool to aid in the development of sophisticated NLP functions. It virtually functions as an arsenal of NLP tools. As a service, on the other hand, SpaCy assists you in completing specific tasks. The text is initially tokenized using spaCy, which results in a Doc object.

After that, the document is processed through a number of phases, commonly known as the processing pipeline. The trained pipelines typically use a pipeline that comprises an entity recognizer, a tagger, a lemmatizer, and a parser. Natural Language Processing The area of computer science called "natural language processing" (NLP), and specifically the subfield of artificial intelligence (AI), focuses on instructing computers to understand both spoken and written language. Similar to how humans perceive text, the second-most-frequented website in the world is YouTube.

In order to represent human language using rules, natural language processing (NLP) models can be categorized into three main types: statistical models, machine learning models, and deep learning models. These innovations make it possible for computers to "completely understand" what is being said or written. By processing human language as text or auditory data, including the user’s intentions and sentiment, Natural language processing (NLP) is the technology behind computer programs that can translate text between languages, respond to spoken commands, and summarize large amounts of text quickly and in real-time.

To enable computers to understand the text and speech data they process, various NLP tasks involve breaking down a human text and voice data. Some of these tasks include:

- Speech recognition, often known as speech-to-text, is the procedure for precisely turning speech input into text. For any application that needs to respond to spoken instructions or queries, speech recognition is a necessary component.

Voice recognition is particularly difficult because to the way people speak, which is fast, muddled, with various accents and intonations, in a variety of dialects, and usually with poor grammar.

- Named Entity Recognition (NER) is a technique used to identify important words or phrases as entities. For instance, NER can identify "Vizag" as a location or "Sam" as a person's name.

- Co-reference resolution involves determining whether two words refer to the same entity or concept. This is commonly used to identify the referent of pronouns, such as "He" referring to "Sam." However, it can also involve recognizing the...
use of metaphors or idioms, such as when "Rock" refers to a physically strong and powerful person rather than a literal stone.

- Sentiment analysis examines the text for any emotional components, such as actions, emotions, mockery, perplexity, and suspicion.

II. LITERATURE SURVEY:

Ishitva Awasthi, Kuntal Gupta, Prajbot Singh Bhojal, Anand, and Piyush Kumar[1]:
Text summarization techniques include abstract and extractive methods. The ability to interpret phrases based on their linguistic and statistical properties is one of the benefits. The drawback is that different situations benefit from different summarising techniques. Which approach is more effective cannot be determined.

Adhika Pramita, Supriadi Rustad, Abdul Shukur, and Affandy[2]:
In 2020, it was published. Text summaries and systematic review methods have been employed. The disadvantage of the fuzzy-based method is that it is insufficient for semantic issues. The procedures used by the extractive industries need to be improved in a number of ways.

Hardik Pradeep, Meghana Naik, and Parth Rajesh Dedhia[3]:
In 2020, the book was released. There have been several applications for Pointer Mechanism, Encoder-Decoder, and Seq2Seq. The disadvantage is that if the model is given many papers, it will not function.

Aniqa Dilawari and Muhammad Usman Ghani Khan[4]:
They made use of the RCNN deep neural network model and the multi-line video description. The issue is that it merely highlights how brief the overview is. Memory capacity and time limits are not taken into account.

The LSA Natural Language Processing algorithm, which needs less computing resources and doesn't need training data, was suggested in this study.

Rand Abdulwahid Albeer, Huda F. Al-Shahad[6]:
It was suggested that the TF-IDF algorithm was employed to summarise the YouTube video's transcript based on its key sentences. With this experiment, it was demonstrated that the method utilised for extraction summary is appropriate. By employing this technique, TF-IDF demonstrated its strength in producing the value that determines which word inside the text is crucial.

Atluri Naga Sai Sri Vybhavi; Laggisetti Valli Saroja[7]:
It is suggested that the developed model receives user-provided video links and the needed summary duration as input and outputs a condensed transcript. The results show that when compared to other suggested procedures, the final translated text was obtained faster. In addition, the final text accurately and consistently conveys the main idea of the video.

III. METHODOLOGY:

3.1

LSA represents the documents and the terms as vectors in a high-dimensional space, and then applies a mathematical technique called Singular Value Decomposition (SVD) to reduce the dimensionality of the space. This process results in a new representation of the documents and terms that captures the underlying relationships between them. The reduced-dimensional representation obtained by LSA can be used to perform tasks such as document classification, document similarity ranking, and information retrieval. It has been used in various applications, including search engines, text classification, and recommendation systems.
Algorithm:

**Preprocessing:** Preprocess the text data by removing stop words, stemming or lemmatizing, and converting the text into a numerical representation such as a TF-IDF matrix.

**Sentence splitting:** Split the text into individual sentences

**Sentence ranking:** Rank the sentences based on their relevance to the summary. This can be done using various techniques such as TF-IDF, graph-based algorithms, or deep learning models.

**Summary generation:** Select the top-ranked sentences to create the summary. The length of the summary can be specified beforehand or determined automatically.

**Deployment:** Deploy the summarization model as a web application, chrome extension for users to summarize text.

### IV. Experimental Results and Analysis:

#### 4.1 Libraries used:

**Google Trans:** Google Translate API is implemented by google trans, a free and unlimited Python utility. It calls functions like detect and translate using the Google Translation Ajax API. Main features:

- Quick and dependable; it makes use of the same servers as the translate.google.com
- Automatic language recognition
- Translations in volumes
- Adaptable service Pooling URL connections is a benefit of utilizing requests.
- HTTP/2 assistance

**SpaCy:** Modern features for rapid, high-volume natural language processing are provided by the free, open-source Python package called spaCy. It supports the creation of models and production tools that can help the functionality of chat bots, document analysis, and all other kinds of text analysis. Software for use in production is spaCy's primary focus. Additionally, SpaCy enables deep learning workflows that let users connect statistical models that have been built using well-known machine learning frameworks. For named entity identification, dependency parsing, text categorization, and part-of-speech tagging, SpaCy uses convolutional neural network models (NER). Prebuilt statistical neural network models are available for 23 different languages that can carry out these tasks, additionally a multilingual NER model. Due to increased support for tokenization for more than Sixty-Five different languages, users can easily train custom models on their own datasets.

**NLTK:** A Python toolkit for working with NLP is called NLTK (Natural Language Toolkit). It offers a there are many different text-processing libraries and test datasets available. With NLTK, various activities can be carried out, including tokenizing and visualizing parse trees. The field of natural language processing aims to make natural human language understandable to computer programs (NLP). You could study a lot of unstructured data that includes text that humans can read. The data must be preprocessed before programatically exploring it. using NITK, we can find a text to analyze, get your work ready for analysis, examine your
content, and, based on your analysis create visualizations.

**Latent Semantic Analysis (LSA):** LSA, which stands for Latent Semantic Analysis, is one of the foundational techniques used in topic modeling. The core idea is to take a matrix of documents and terms and try to decompose it into separate two matrices – A document-topic matrix A topic-term matrix. Latent semantic analysis (LSA) is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis). A matrix containing word counts per document (rows) represent unique words and columns represent each document is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns. Documents are then compared by cosine similarity between any two columns. Values close to 1 represent very similar documents while values close to 0 represent very dissimilar documents.

**RESULTS**
Comparing the Transcript and Summary

V. CONCLUSION:

With the YouTube Video Summarization tool project, we've devised a video summarizing system based on natural language processing (NLP) that condenses YouTube video transcripts while maintaining their essential elements. Our goal was to create a mechanism that assists in saving time. People can access their preferred material in several languages. Downloadable transcripts are very helpful for Braille because they may be printed on paper using automated Braille styluses. In English, there is also real-time text-to-speech conversion accessible. Our goal in starting this project is to draw attention to the importance of providing tools that make it easier for users to maneuver the complicated internet. The work of summarization becomes challenging in this since the entire text must be understood by the computer in order for it to provide a summary with wholly new words. Its meaning is the same as the original text. RNN and LSTM are two future methodologies that would be incorporated to learn the words and to store their appropriate meanings; encoders decoders & sequence-to-sequence model have to be used to produce efficient summary using this methodology.

Here, the model has to be trained with many words & their synonyms, one word replacing many words & the correct usage of each word. We hope to improve the concept further and create an automatic text summarizer in the future.

REFERENCES:

1. Natural Language Processing (NLP) based Text Summarization –
2. ASurveyhttps://ieeexplore.ieee.org/abstract/document/9358703?casa_token=qQ1uWklaOp0AAA AA:1Rm8FZ3ACyVu1YbYeABia6rM7p5kFKWQpJwQCaE5E9BipORB4yHWIMumZ458nc6yp6YJNnw
ets1649174683.pdf
7. Extractive Automatic Text Summarization using SpaCy in Python & NLP
8. https://ieeexplore.ieee.org/abstract/document/9404712?casa_token=aUorZySSJwcAAAA AA:1Rm9S006ZQNzj_B_eMM7e4MXlalBg7xZkhpVQETJnd1ainwcY1cG1cYbRdqkX3P7K8IL9tl suppose