G H RAISONI COLLEGE OF ENGINEERING AND MANAGEMENT, PUNE
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Youtube transcript summarizer

Presentation

Team

Pratik Rajesh jade

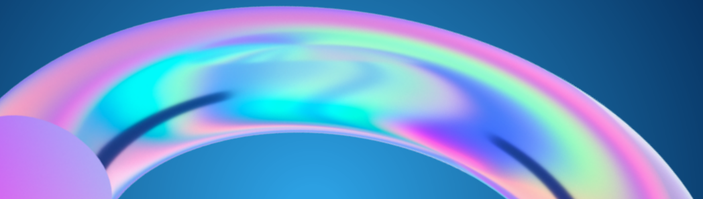
Ty AI A70

pratik.jade.cs@ghrcem.raisoni.net

MOTIVATION

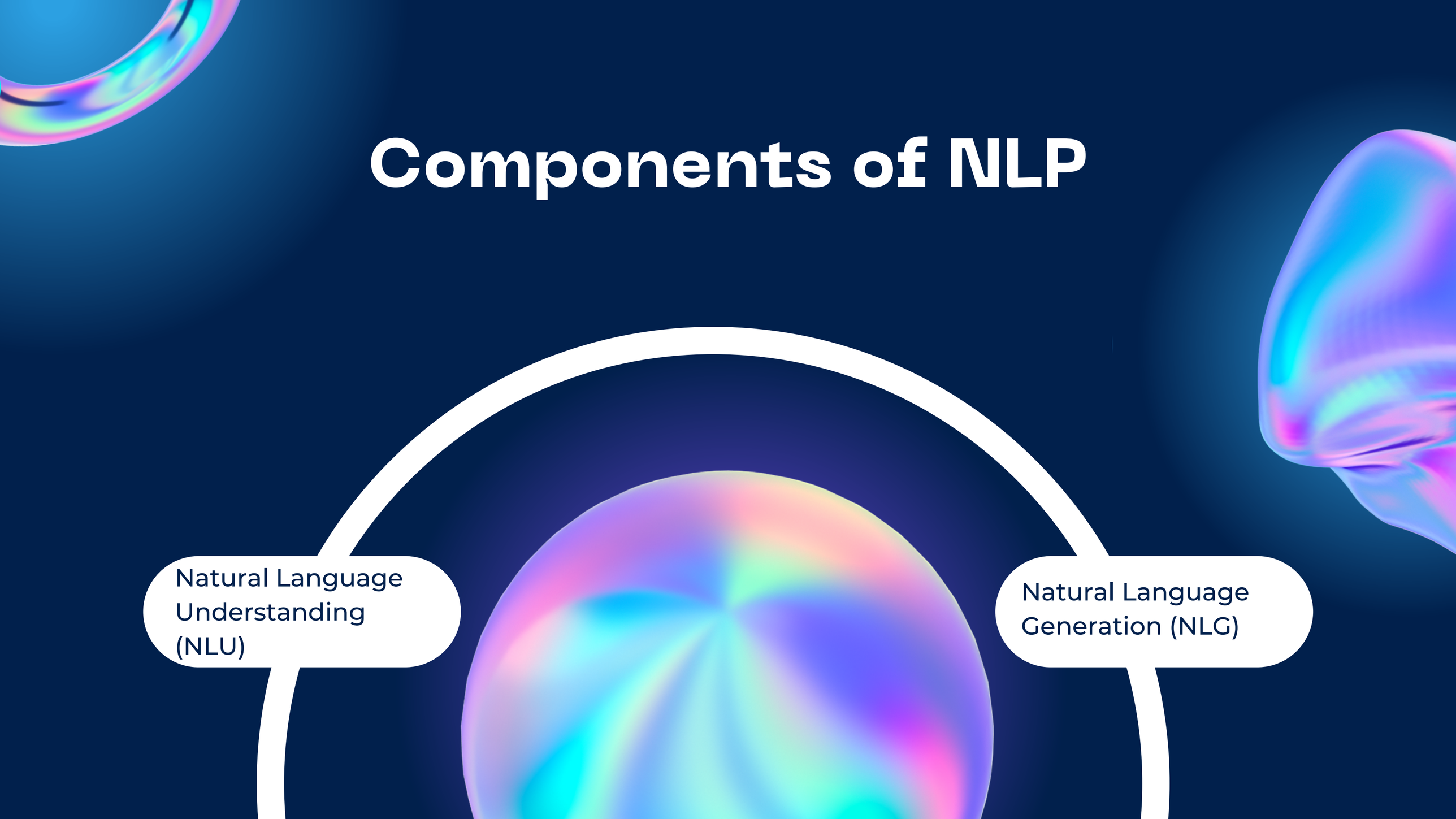
Have you ever imagined getting a short summary of a big youtube tutorial or video for quick reading before watching the video, definitely this will help you to save a lot of your time by getting a quick understanding or summarization about the video in a short time. In this project, a YouTube Summarizer which will summarize the content(subtitle) of the youtube video. For many videos, the main content of the videos is only 50-60% of the total length, so our youtube summarizer will summarize the content of the video by keeping all the important points and making it short and easily understandable. This will be useful in getting the summary of several lecture videos easily.

Introduction to NLP



- NLP stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence.
- It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages.
- It helps developers to organize knowledge for performing tasks such as translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction, and topic segmentation.

Components of NLP



Natural Language Understanding (NLU)

- Natural Language Understanding (NLU) helps the machine to understand and analyse human language by extracting the metadata from content such as concepts, entities, keywords, emotion, relations, and semantic roles.
- NLU mainly used in Business applications to understand the customer's problem in both spoken and written language.

Natural Language Generation (NLG)

- Natural Language Generation (NLG) acts as a translator that converts the computerized data into natural language representation.
- It mainly involves Text planning, Sentence planning, and Text Realization.

what is summarization

- Summarization is the technique of making short, understandable notes for a given large text document without excluding the important contents of the passage
- There are 2 types of summarization in NLP, extractive summarization and abstractive summarization.

Types of Summarization –

extractive summarization

- the system will extract the important paragraphs and contents from the given passage and combine these extracted paragraphs to create the summarized text.

Abstractive summarization

- the system will create a summary based on the given passage with its own words. This is more complex than extractive summarization.

NLTK

- NLTK is a toolkit build for working with NLP in Python. It provides us various text processing libraries with a lot of test datasets. A variety of tasks can be performed using NLTK such as tokenizing, parse tree visualization, etc...
- Tokenizing, you can conveniently split up text by word or by sentence. This will allow you to work with smaller pieces of text that are still relatively coherent and meaningful even outside of the context of the rest of the text. It’s your first step in turning unstructured data into structured data, which is easier to analyze.

Types of tokenization

Tokenizing by word:

- Words are like the atoms of natural language. They’re the smallest unit of meaning that still makes sense on its own. Tokenizing your text by word allows you to identify words that come up particularly often. For example, if you were analyzing a group of job ads, then you might find that the word “Python” comes up often. That could suggest high demand for Python knowledge, but you’d need to look deeper to know more.

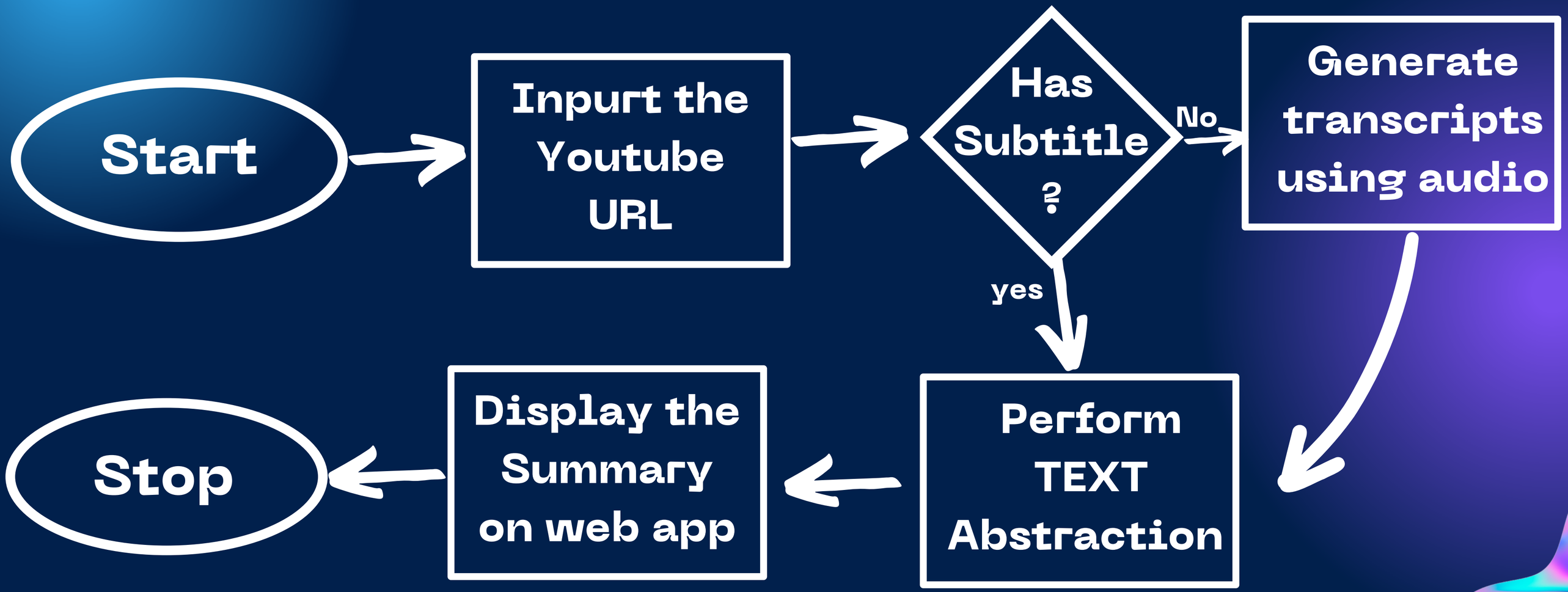
Tokenizing by sentence

- When you tokenize by sentence, you can analyze how those words relate to one another and see more context. Are there a lot of negative words around the word “Python” because the hiring manager doesn’t like Python? Are there more terms from the domain of herpetology than the domain of software development, suggesting that you may be dealing with an entirely different kind of python than you were expecting?

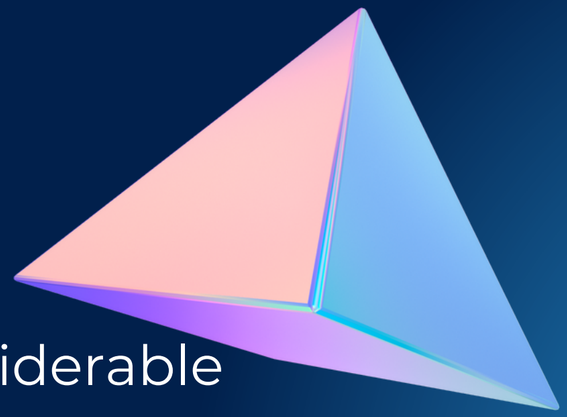
youtube_transcript_api

- Python provides a large set of APIs for the developer to choose from. Each and every service provided by Google has an associated API. Being one of them, YouTube Transcript API is very simple to use provides various features.
- This is an python API which allows you to get the transcripts/subtitles for a given YouTube video. It also works for automatically generated subtitles, supports translating subtitles and it does not require a headless browser, like other selenium based solutions do!

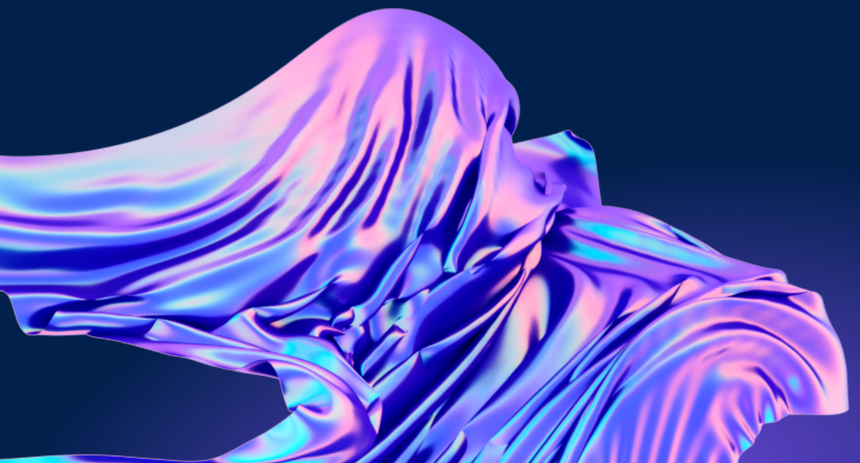
How its work


Start
Inpurt the Youtube URL
Has Subtitle ?
No
yes
Generate transcripts using audio
Perform TEXT Abstraction
Display the Summary on web app
Stop

Conclusion



Recently, video summarization has attracted considerable interest from researchers and as a result, various algorithms and techniques have been proposed.This project is to provide a web app or a chrome extension that can be used tosummarize the YouTube video content and extract important information from those patterns by using state-ofthe-art Natural Language Processing methods for abstractive text summarization and Machine Learning for classification.



Thank You...

A70 Pratik Jade
https://github.com/pratikjade
pratik.jade.cs@ghrcem.raisoni.net