**Name**: Deep Kumawat

**Enrollment:**21C21041

**Branch**: B.Tech -CSE(Sem-5)

**Title: Youtube Video Summarization**

**Description** :

In today's digital age, YouTube has become a massive repository of video content, with millions of hours of new content uploaded every day. Users often face the challenge of finding the ideal video that they need. But lately a trend of misleading titles and/or thumbnails has risen which nowadays is termed as clickbait. This makes the user’s experience with respect to finding the desired video to watch really annoying at times. The goal of this project is to develop an automated YouTube video summarization system that can generate concise and informative summaries for videos on the platform. This system will use natural language processing (NLP) and machine learning techniques to extract key information from videos and present it in a summarized form.

**Need for the topic:**

The need for a YouTube video summarization system arises from several key considerations and challenges:

* Information Overload: YouTube is one of the largest video-sharing platforms globally, with an immense volume of new content uploaded daily. As a result, users often face information overload and have difficulty finding the specific information they seek within lengthy videos.
* Time Constraints: Many users have limited time available for consuming video content. They may not have the patience or time to watch entire videos to extract relevant information.
* Anti Clickbait: Viewers can read the brief summary if they are not sure about the subject of the video just by the thumbnail and title of the video.
* Accessibility: Some users may have accessibility needs that require content to be presented in alternative formats, such as text-based summaries, to make videos more accessible to them.
* Educational Benefits: Educational institutions, trainers, and students can benefit from video summarization to quickly identify relevant content within educational videos, enhancing the learning experience.
* Content Curation: Content creators, marketers, and curators can use video summaries to provide highlights and previews of videos, increasing engagement and viewer retention.
* Efficient Search: A video summarization system can improve search capabilities by enabling users to search for specific information within video summaries.
* Machine Learning Opportunities: Developing advanced natural language processing and machine learning techniques for video summarization presents opportunities for innovation and research.
* User Experience: Improving the user experience on YouTube by reducing the time and effort required to extract valuable information from videos can enhance user satisfaction and retention.

**Key Objectives:**

* Data Collection: Collect a diverse dataset of YouTube videos from various channels and categories. The dataset should include videos of different lengths and topics to ensure the system's versatility.
* Transcript Generation: Automatically transcribe the video's spoken content to text. This transcript will serve as the basis for summarization.
* Text Summarization: Implement text summarization techniques, such as extractive or abstractive summarization, to create concise and coherent summaries from the video transcript. We have experimented with both types and have decided that abstractive summarization would be a good first step but we plan to create another version with extractive summarization.
* Keyphrase Extraction: Extract key phrases and keywords from the video content to enhance the summary's accuracy and relevance.
* User Interaction: Develop a user-friendly interface that allows users to input a YouTube video URL or search for videos of interest. The system should provide summarized content and allow users to watch the full video if desired.
* Scalability: Ensure that the system can handle a large number of videos and process them efficiently.

**Expected Outcomes:**

* A fully functional YouTube video summarization system capable of summarizing a wide range of videos from the platform.
* Improved user experience by reducing the time and effort required to extract valuable information from YouTube videos.
* Enhanced search capabilities, as users can now quickly identify relevant videos based on the quality of their summaries.

**Benefits:**

* The project addresses the growing need for efficient content consumption on YouTube, benefiting both content creators and viewers.
* It showcases the capabilities of advanced NLP and machine learning techniques in real-world applications.
* The system can potentially be used for educational purposes, content curation, and information retrieval beyond YouTube.

**Technical Stack:**

* Natural Language Processing (NLP) library NLTK.

**Process :**

Pseudo code :

**IMPORT** libraries pytube, youtube\_transcript\_api, nltk

**GET** Youtube video url input from user

**EXTRACT** video id from url. Example,

**url:** https://www.youtube.com/watch?v=xC-c7E5PK0Y

**id:** xC-c7E5PK0Y

**Get** the transcript

**Save** it in transcript.txt file.

**Remove** stop words

stopWords = set(stopwords.words("english"))

**tokenizing** words

words = word\_tokenize(inputText)

**Reducing** the words to it's root form using PorterStemmer

stem = PorterStemmer()

**Creating** dictionary for the word frequency table

frequencyTable = dict()

**Create** a dictionary containing (sentence, sentence\_scores) as (key, value) pairs. This helps us quantify how important a sentence is. The more the frequency of the words inside that sentence the more its score. But it’s also inversely proportional to the number of words in it. Formula : (sum of all the frequencies of the words of a sentence) / (number of words)

**Printing** the sentences whose score is higher than a certain threshold. We initially use the average score of the sentences as the threshold but it can be changed to get a more or less number of sentences in the summary.

**Detail of the algorithm :**

` Imports necessary libraries, including pytube for YouTube video handling, youtube\_transcript\_api for transcript retrieval, and nltk for natural language processing.

Takes input from the user for a YouTube video ID.

Retrieves the transcript of the specified YouTube video.

Creates a text file to save the video transcript, naming it based on the video ID.

Utilizes the generated text file in another script called transcriptSummary.py (not shown in the provided code).

Removes common English stop words from the transcript using NLTK's stopwords list.

Tokenizes the text into individual words.

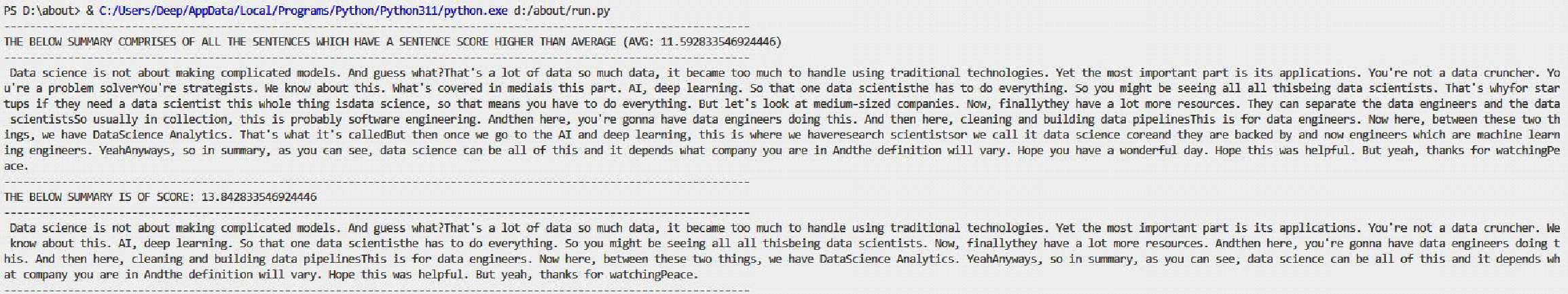
Applies word stemming using the Porter Stemmer algorithm.

Initializes an empty dictionary to store word frequencies for later use.

Calls unspecified functions related to text summarization and sentence scoring, although the actual implementation of these functions is not provided.

**Output :**

**Screenshot:**



**Text form:**

-------------------------------------------------------------------------------------------------------------------

THE BELOW SUMMARY COMPRISES OF ALL THE SENTENCES WHICH HAVE A SENTENCE SCORE HIGHER THAN AVERAGE (AVG: 11.592833546924446)

-------------------------------------------------------------------------------------------------------------------

Data science is not about making complicated models. And guess what?That's a lot of data so much data, it became too much to handle using traditional technologies. Yet the most important part is its applications. You're not a data cruncher. You're a problem solverYou're strategists. We know about this. What's covered in mediais this part. AI, deep learning. So that one data scientisthe has to do everything. So you might be seeing all all thisbeing data scientists. That's whyfor startups if they need a data scientist this whole thing isdata science, so that means you have to do everything. But let's look at medium-sized companies. Now, finallythey have a lot more resources. They can separate the data engineers and the data scientistsSo usually in collection, this is probably software engineering. Andthen here, you're gonna have data engineers doing this. And then here, cleaning and building data pipelinesThis is for data engineers. Now here, between these two things, we have DataScience Analytics. That's what it's calledBut then once we go to the AI and deep learning, this is where we haveresearch scientistsor we call it data science coreand they are backed by and now engineers which are machine learning engineers. YeahAnyways, so in summary, as you can see, data science can be all of this and it depends what company you are in Andthe definition will vary. Hope you have a wonderful day. Hope this was helpful. But yeah, thanks for watchingPeace.

-------------------------------------------------------------------------------------------------------------------

THE BELOW SUMMARY IS OF SCORE: 13.842833546924446

-------------------------------------------------------------------------------------------------------------------

Data science is not about making complicated models. And guess what?That's a lot of data so much data, it became too much to handle using traditional technologies. Yet the most important part is its applications. You're not a data cruncher. We know about this. AI, deep learning. So that one data scientisthe has to do everything. So you might be seeing all all thisbeing data scientists. Now, finallythey have a lot more resources. Andthen here, you're gonna have data engineers doing this. And then here, cleaning and building data pipelinesThis is for data engineers. Now here, between these two things, we have DataScience Analytics. YeahAnyways, so in summary, as you can see, data science can be all of this and it depends what company you are in Andthe definition will vary. Hope this was helpful. But yeah, thanks for watchingPeace.

-------------------------------------------------------------------------------------------------------------------