A Project report

on

YOUTUBE TRANSCRIPT SUMMARIZER

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

In

Computer Science and Engineering

Submitted by

A Aravind	19H51A0595
K Rohan	19H51A05A6
K Bharath	19H51A05K7

Under the esteemed guidance of
Mr. Muthu Krishnan
Assistant Professor CSE Department

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING AND TECHNOLOGY
(An Autonomous Institution under UGC & JNTUH, Approved by AICTE, Permanently Affiliated to JNTUH, Accredited by NBA.)
KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.
2019- 2023

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Mini Project-2 report entitled " **YOUTUBE TRANSCRIPT SUMMARIZER**"being submitted by **A.ARAVIND(19H51A0595),K.ROHAN(19H51A05A6),K.BHARATH(19H51A05K7)** in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr.Muthu Krishnan
Assistant Professor
Dept of CSE

Dr. S Siva Skandha
Associate Professor and HOD
Dept of CSE

Submitted for viva voice Examination held on _____

External Examiner

Acknowledgment

With great pleasure I want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr. Muthu Krishnan** Dept of Computer Science and Engineering for his valuable suggestions and guidance during the execution of this project work.

We would like to thank **Dr. S Siva Skandha**, Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and technology, for his constant support and motivation in carrying out the project work successfully.

I am highly indebted to **Dr. V A Narayana**, Principal CMRCET for giving permission to carry out this project in a successful and fruitful way.

I would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering for their co-operation.

Finally, I express my sincere thanks to **Mr. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care. I sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

SIGNATURE

A.ARAVIND 19H51A0595

K.ROHAN 19H51A05A6

K.BHARATH 19H51A05K7

DECLARATION

We hereby declare that results embodied in this Report of Projection **"YOUTUBE TRANSCRIPT SUMMARIZER"** are from the work carried out by using partial fulfillment of the requirements for the award of B. Tech degree. We have not submitted this report to any other university/institute for the award of any other degree.

SIGNATURE:

A.ARAVIND 19H51A0595

K.ROHAN 19H51A05A6

K.BHARATH 19H51A05K7

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Enormous number of video recordings are being created and shared on the Internet throughout the day. In today's daily lifestyle, it has become really difficult to spend time watching such Videos which may have a longer duration than expected and sometimes our efforts may become futile if we couldn't find relevant information out of it as most of the videos are been uploading to grab the attention of the viewers by tricking or misleading the viewers by thumbnails, advertisements, etc. As the number of users of YouTube are increasing rapidly from year to year, this may directly impact the number of videos to be created. In greed for the number of views, the chance that the creators of the videos may give wrong information on the original content of the video is probably very high. This may waste the valuable time and resources of the user. To further improve the user interaction with the summarizer summarizes the extracted text and gives result while comparing with the audio.

CHAPTER 1
INTRODUCTION

INTRODUCTION

Huge number of video recordings are being created and shared on the Internet throughout the day. The number of YouTube users in 2020 was approximately 2.3 billion, and has been rapidly increasing every year. For every single minute, watch time of 300 hours of videos are uploaded to YouTube. Almost one-third of the YouTube viewers in India access videos on their mobiles and spend .

It has become really difficult to spend time watching such videos which may have longer duration than expected and sometimes our efforts may become futile if we couldn't find relevant information out of it which we are in search of. It is frustrating and time consuming to search for the videos that contain the information we are actually looking for. For instance, there are many videos available online in which the speaker talks for a long time on a given topic, but it is hard to find the content the speaker wants to convey to the audience unless we watch the entire video.Python has various packages which are very helpful. Currently accessing the YouTube content has been made easier through the API in the python library like the transcripts of the videos etc. By taking this advantage we directly access the transcripts of the video and summarize to view them to the user.

This can be implemented by using Hugging face transformer which is one of the text summarization techniques. To Make User understandable we use google collab so that the user can understand the code if he is the developer or if the user isn't a developer then he gets the summarized output. This summarized text is generated by the hugging face transformer package. The YouTube videos are usually summarized through manual descriptions and thumbnails. YouTube is the second most visited website worldwide.

The range of videos on YouTube includes short films, music videos, feature films, documentaries, audio recordings, corporate sponsored movie trailers, live streams, vlogs, and many other contents from popular YouTubers. users watch more than one billion hours of video every day. This project proposes the usage of a transformer package for summarizing the transcripts of the video, thereby providing a meaningful and germane summary of the video.

CHAPTER 2
BACKGROUND WORK

2. BACKGROUND WORK

2.1 Existing Solution:

In 2021, "Natural Language Processing (NLP) based Text Summarization - A Survey" was published by Ishitva Awasthi, Kuntal Gupta, Prajbot Singh Bhojal, Anand, Piyush kumar. The techniques used are Extractive and abstract methods for summarizing texts. The advantages are Based on linguistic and statistical characteris tics, the implications of sentences are calculated.

"Review of automatic text summarization techniques & methods" is developed by AdhikaPramita, Supriadi Rustad, Abdul Shukur, Affandy. It was published in 2020. They have used Text summarization, Systematic review techniques.

"Study on Abstractive Text Summarization Techniques" is developed by Parth Rajesh Dedhia, Hardik Pradeep, Meghana Naik. It was published in 2020. They have used Seq2Seq, Encoder-Decoder, and Pointer Mechanism.

"Abstractive Summarization of video sequences" is developed by AniqaDilawari, Muhammad Usman Ghani Khan. They have used multi-line video description, RCNN deep neural network model.

2.2 Demerits in existing Solution:

- The disadvantage is each type of summarization technique
- is useful in different situations. One cannot say which technique is more promising .
- The drawback is that the Fuzzy based approach is weak in semantic problems.
- Many loopholes have to be improved in extractive methods .
- The drawback is that the current model does not work if multiple documents are passed to the model.
- The drawback is It focuses only on the conciseness of the summary. Memory efficiency and time constraints are not under consideration.

CHAPTER 3
PROPOSED SYSTEM

3.1 Proposed model:

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.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. This model is designed by using python language,hugging face transformers,youtube_transcript_api.

3.2 Software Requirements:

Operating Systems	: Windows 10 and Above,Any Linux destro
Technology	: Web browser (chrome, fire fox)
Ide	: Google collab
Libraries	: transformers, youtube_transcript_api.

3.3 Software Features:

3.3.1 Google Collab

Google Colaboratory, popularly known as Colab, is a web IDE for python that was released by Google in 2017. Colab is an excellent tool for data scientists to execute Machine Learning and Deep Learning projects with cloud storage capabilities.Colab is basically a cloud-based Jupyter notebook environment that requires no setup. What's more, it provides its users free access to high compute resources such as GPUs and TPUs that are essential to training models quickly and more efficiently.

The main parts of our element are as follows:

Code: This is a input box where we have write the code or if we want to install any packages require we can download them from this input box.

Run Button: To the left side of the input text box of google collab their will be a run button in hover mode ,we can run the code while clicking the button

The Output: The output will be displayed in the new text area in the collab itself.

Advantages:

i)Similar to working on a Google Sheet, multiple developers can co-code on the same workbook simultaneously. This is particularly helpful if it is a group project, right? You can also share your work with your fellow developers.

ii)When you use Jupyter Notebook as your working environment, all your work is stored on the local machine. Opting for Colab, on the other hand, lets you access your workbooks seamlessly from any other device, as they are all saved in your Google Drive.

3.3.2 Python & it's Libraries:

Python is one of the most mention-worthy programming languages in today's world. It ranks among the fastest-growing programming language in the world. It is versatile, flexible, extremely effective, easy to use and develop. It has a very active community as well. This means that the best minds in the field will provide enough support to adopt this new language. It is used in numerous organizations due to its multiple programming paradigm support and its performance of automatic memory management. Due to its comprehensive standard library, Python is also often referred to as a battery-included language.

ADVANTAGES OF Python:

Easy to use and learn: For beginners, Python is straightforward to use. It is a high-level programming language, and its syntax is like the English language. These reasons make the language easy to learn and adapt to. Compared to Java and C, in Python, the same task can be performed using fewer lines of code. As a result of its easy learning, the principles in Python can be executed faster compared to other languages.

Extensive library: Python provides the user with a vast library. Python's standard library is massive, and just about every function one needs to perform is available in its library. This is because it has a hugely supportive community and corporate sponsorship. External libraries are not used by users while working with Python.

Flexibility: This language is very flexible, and hence it allows the user to try new things. The users can develop new sorts of the application using Python programming language. The language does not restrict the user from trying something different. Other programming languages do not provide this type of flexibility and freedom, and hence Python is more preferred in these matters.

3.3.3 transformers:

transformers in Python can be used to clean, reduce, expand or generate features. The fit method learns parameters from a training set and the transform method applies transformations to unseen data.There are several predefined transformers available in different Python packages, that allow us to easily apply different transformations on our data sets.Depending on the problem being addressed, custom transformers can help overcome some of the incapability or challenges posed by predefined transformers.

We used pipelines for text summerization in our project where we create a model to summarize the extracted text.we can even do text classification also with pipeline if we want.A “pipeline” denotes a series of concatenated data transformations. Every step in a pipeline feeds from the previous step and data flows through the pipeline from beginning to end. This helps optimize the process of model building, thus allowing more time to understand the underlying data.

3.3.4 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 module is used for getting the captions/subtitles from a YouTube Video.
It is used to extract the transcript in the youtube vedio through the vedio id.

CHAPTER-4

DESIGNING

4.1 CREATING A MODEL

We used pipelines for text summerization of transformers in our project where we create a model to summarize the extracted text. we can even do text classification also with pipeline if we want. A “pipeline” denotes a series of concatenated data transformations. Every step in a pipeline feeds from the previous step and data flows through the pipeline from beginning to end. This helps optimize the process of model building, thus allowing more time to understand the underlying data.

4.2 GET TRANSCRIPTS

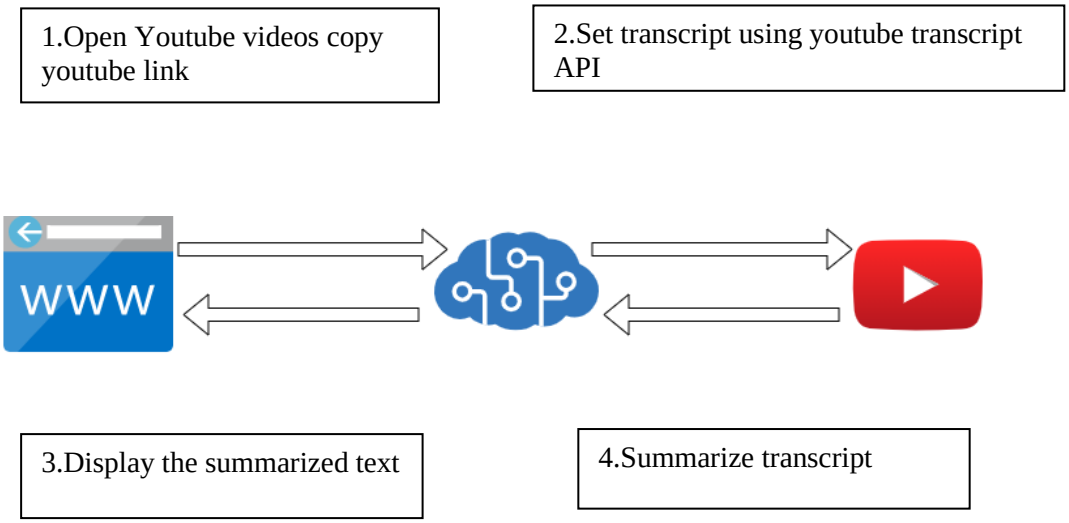
We use youtube_transcript_api library in python to extract the transcript in the youtube video through video id where to extract the trascript we revoke the function call of get_transcript(). It returns us a list of dictionaries in which each dictionary contains 3 key-value pair inside it, the first one being the content, the second one being the time instant from which the caption sentence/phrase start to be spoken and the third one being the duration in seconds that is taken to speak the sentence or phrase completely.

4.3 PERFORM TEXT SUMMARIZATIOZ:

Text summarization is the task of shortening longer text into a precise summary that preserves key information content and overall meaning. Thereare two different approaches that are widely used for text summarization:Extractive Summarization: In this the model identifies the important sentences and phrases from the original text and only outputs the necessary part.

Abstractive Summarization: The model produces a completely different text that is shorter than the original, it generates new sentences in a new form. In this project, we will use transformers for this approach. In this system, we will use the HuggingFace transformers library in Python to perform abstractive text summarization on the transcript obtained thorugh the youtube_transcript_api.get_transcript(). So extracting text we create an empty string so sum up all the extracted text to check whether its working or not.If it’s working we simply use pipeline() to summerize the extracted text and then we print them.

4.4 Use Case Diagram:



4.5 GOOGLE COLLAB:

Colab is an excellent tool for data scientists to execute Machine Learning and Deep Learning projects with cloud storage capabilities.Colab is basically a cloud-based Jupyter notebook environment that requires no setup. What's more, it provides its users free access to high compute resources such as GPUs and TPUs that are essential to training models quickly and more efficiently.

4.6 DISPLAY SUMMARIZED TEXT:

After extracting the transcript and summarizing the text we make only few iterations to not make runtime error,so we make count of the total length of iterations and divide it by the 10^x based on the length of the extracted text and then print it .The Output will be in Google collab if we use jupyter google collab .

4.7 Source Code

```
from transformers import pipeline
    from youtube_transcript_api import YouTubeTranscriptApi
    youtube_video = "https://www.youtube.com/watch?v=A4OmtyaBHFE"
video_id = youtube_video.split("=")[1]
transcript = YouTubeTranscriptApi.get_transcript(video_id)
result = ""
for i in transcript:
result += ' ' + i['text']
n=len(result)
summarizer = pipeline('summarization')
num_iters = int(n/1000)
summarized_text = []
for i in range(0, num_iters + 1):
start = 0
start = i * 1000
end = (i + 1) * 1000
print("input text \n" + result[start:end])
out = summarizer(result[start:end])
out = out[0]
out = out['summary_text']
print("Summarized text\n"+out)
summarized_text.append(out)
```

CHAPTER 5
IMPLEMENTATION

5 IMPLEMENTATION:

Implementation is the stage in the project where the theoretical design is turned into a working system and is giving confidence on the new system for the users that it will work efficiently and effectively. It involve careful planning, investigation of the current system and its constraints on implementation, design of methods to achieve the change over, an evaluation of change over methods. Apart from planning major task of preparing the implementation are education and training of users. The implementation process begins with preparing a plan for the implementation of the system. According to this plan, the activities are to be carried out, discussions made regarding the equipment and resources and the additional equipment has to be acquired to implement the new system. In network backup system no additional resources are needed.

Implementation is the final and the most important phase. The most critical stage in achieving a successful new system is giving the users confidence that the new system will work and be effective. The system can be implemented only after thorough testing is done and if it is found to be working according to the specification. This method also offers the greatest security since the old system can take over if the errors are found or inability to handle certain type of transactions while using the new system.

5.1 Security and Maintenance:

Maintenance involves the software industry captive, typing up system resources .It means restoring something to its original condition. Maintenance follows conversion to the extent that changes are necessary to maintain satisfactory operations relative to changes in the user's environment. Maintenance often includes minor enhancements or corrections to problems that surface in the system's operation. Maintenance is also done based on fixing the problems reported, changing the interface with other software or hardware enhancing the software. Any system developed should be secured and protected against possible hazards. Security measures are provided to prevent unauthorized access of the database at various levels. An uninterrupted power supply should be so that the power failure or voltage fluctuations will not erase the data in the files. Password protection and simple procedures to prevent the unauthorized access are provided to the users .The system allows the user to enter the system only through proper user name and password

CHAPTER 6

RESULTS AND DISCUSSION

6.1Install required packages: To install libraries or packages use “pip install <package name>” in gnome console or use “!pip install -q <package name>” in google collab.

```
[ ]  !pip install -q transformers
```

```
|████████████████████████████████| 2.9 MB 2.5 MB/s
|████████████████████████████████| 3.3 MB 32.4 MB/s
|████████████████████████████████| 52 kB 1.3 MB/s
|████████████████████████████████| 636 kB 69.3 MB/s
|████████████████████████████████| 895 kB 67.4 MB/s
```

```
[ ]  !pip install -q youtube_transcript_api
```

6.2 Get Video Id:To get video id we need to find a youtube link first and then by simply using split function we get video id.

```
video_id = youtube_video.split("=")[1]
```

```
video_id
```

```
'A4OmtyaBHFE'
```

6.3 Extract transcript: Using youtubetranscriptapi we extract the transcript through the youtube link

```
[ ]  YouTubeTranscriptApi.get_transcript(video_id)
     transcript = YouTubeTranscriptApi.get_transcript(video_id)
```

```
[ ]  transcript[0:5]
```

```
[{'duration': 4.96,
  'start': 1.52,
  'text': "for germany it's the end of an era"},
 {'duration': 5.279,
  'start': 4.4,
  'text': "and as europe's biggest economy there"},
 {'duration': 4.72, 'start': 6.48, 'text': 'are some huge challenges ahead'},
 {'duration': 3.681, 'start': 9.679, 'text': 'from its increasingly complex'},
 {'duration': 3.12,
  'start': 11.2,
  'text': 'relationship with china to climate'}]
```

6.4 Summerizer: Using transformers we invoke the pipeline and summerize the transcript

```
[ ]  summarizer = pipeline('summarization')
```

```
No model was supplied, defaulted to sshleifer/distilbart-cnn-12-6 (https://huggingface.co/sshleifer/distilbart-cnn-12-6)
Downloading: 100% 1.76k/1.76k [00:00<00:00, 28.0kB/s]
Downloading: 100% 1.14G/1.14G [00:29<00:00, 44.9MB/s]
Downloading: 100% 26.0/26.0 [00:00<00:00, 453B/s]
Downloading: 100% 878k/878k [00:00<00:00, 726kB/s]
Downloading: 100% 446k/446k [00:00<00:00, 376kB/s]
```

6.5 Summerized Sentences:

```
#print(summarized_text)
```

```
input text
 for germany it's the end of an era and as europe's biggest economy there are some huge challenges ahead from its increasingly complex relationship with china to clima
Summarized text
 For germany it's the end of an era and as europe's biggest economy there are some huge challenges ahead . There's one german industry that reveals a lot about the cou
input text
g is incredibly important in germany it represents around 18 percent of gdp and of course the car sector represents a very large part of that it's not quite you know w
Summarized text
 After over half a century of global dominance german car makers are now playing catch-up . In 2000 seven of the world's most valuable companies were german but in the
input text
 trading partner [Music] in 1998 german trade with china was worth 16 billion dollars now it's worth over 180 billion but there's a fear that germany could have become
Summarized text
 In 1998 german trade with china was worth 16 billion dollars dollars now it's worth over 180 billion dollars . There's a fear that germany could have become too relia
input text
acturing digitalization this is not traditionally a country that has been closely associated with digital innovation it's going to be a big challenge across the whole
Summarized text
 Two-thirds of german businesses said they still regularly used a fax machine and only one in three german companies have a central digital strategy . The next governm
input text
ve to do a hell of a lot more to translate the rhetoric on digitalization into the reality of fostering genuine change in this country which has been severely lacking
Summarized text
 Germany's car industry is a microcosm for how the country will need to radically shift traditional forms of power to be replaced for cars . oliver zipsa took over as
input text
olid business foundation climate change has forced this shift to a greener future but it was american not german engineering that first realized it an electric car car
Summarized text
 Sono motors is a startup based in munich that is producing an electric vehicle that is equipped with solar cells . The scion can share power with the grid and other c
input text
ilt in [Music] their claim is that they're the only car company that thinks the industry should be selling fewer cars their approach could upend the industry and their
Summarized text
 Sono have vision for a greener future but germany's politics may make it hard to realize . The next coalition is probably going to be made up of three different parti
input text
 well suited to manage this extraordinary industrial transition that we're going to have to bring about in the years ahead so that's going to mean a very very big role
Summarized text
 Angela Merkel's new leader should plan for a workforce in need of upskilling [Music] We're going to have a massive ramp up of jobs in the battery manufacturing sector
input text
ormation but that's left germany's new leaders with a lot to do germany has to do better right we have those very big companies those very big players here but they ar
Your min length is set to 56, but you input length is only 20. You might consider decreasing min length manually, e.g. summarizer('...', min length=10)
```

CHAPTER 7

CONCLUSION AND FUTURE WORK

CONCLUSION:

This project has proposed a YouTube Transcript summarizer. The system takes the input YouTube video id through split function and when the user clicks the run button the code extract the transcripts of that video with the help of python API. The accessed transcripts are then summarized with the transformers package. Then the summarized text is shown in the google collab itself. This project helps the users a lot by saving their valuable time and resources. This helps us to get the gist of the video without watching the whole video. It also helps the user to identify the unusual and unhealthy content so that it may not disturb their viewing experience.

REFERENCES

- Python programminghttps://www.geeksforgeeks.org/python-downloading-captions-from-youtube/
- https://www.analyticsvidhya.com/blog/2022/01/youtube-summariser-mini-nlp-project/
- https://github.com/jdepoix/youtube-transcript-api
- https://www.geeksforgeeks.org/getting-started-with-transformers/