IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

TEXT SUMMARIZATION

[1]Kanan Gulati, [2]Sumit, [3]Mohit Saxena
[1]Student, [2]Student,[3]Assistant Professor
Department of Information Technology,
Ajeenkya DY Patil University, Pune, India.

***Abstract*:**
Over the past two years, taking short notes from lectures has proven to be an important tool for writing down key words and phrases that best represent context. However, many methods now use legacy methods, produce output or require hours of manual editing to produce good results. Recently, new machine learning architectures provide inferential summarization mechanisms by clustering the output embeddings of deep learning models.

In this project, we will use many to many sequence models using the abstractive text summarization techniques to predict content analysis. We have implemented frequency based and position based summary in which few test cases to examine it, later on we have implemented using the you-tube video transcript summarization and Hindi text- summarization techniques. Using the tracking system, we will focus on the specific content while checking the content of the article.

We've used two classes of Attention Mechanisms:

Global Attention: In Global attention, all the hidden states of every time step from the encoder model are used to generate the context vector. Local Attention: Here few of the hidden states which are there from the encoder model generates the context vector.

Index terms—Text Summarization, NLP, Machine Learning, NLTK, Stopwords, Attention Mechanism

1.INTRODUCTION

When it comes to automated text writing, there are two different types: abstraction and extractive Abstract text summarization, it better imitates human summarization because it uses words instead of text, summarizes important points, and is usually small (Genest &amp; Lapalme, 2011). Although this method is desirable and has gone through many research papers, since it simulates how humans collect data, it is difficult to automate, requiring large numbers of GPUs for days of deep learning, or complex algorithms and rules with limited generalizability. traditional NLP methods. With this challenge in mind, the tutorial introduces the content using written content.

In general, remove short written text using the original format of the text, sentence, or paragraph, and leave content using only elements from the material. To start using the service, simply use the phrase to explain. Transcripts of video lectures are available in many MOOC contexts, but finding the most important information in each lecture can be difficult .Currently, many attempts have been made to solve this problem, but almost all solutions use the normal process of handling messages, which should be monitored regularly due to the possibility of being beautiful. Because the most important tools in the course content are needed, the course content provides a RESTful API and a command line (CLI) tool that can provide summary for each map to demonstrate that the application can be extended.

The following sections examine the background and related work of the lecture summaries, the methods used in the design of the service, the results and evaluation of the model, and an abstract example showing how these compare to the most used.

Natural language processing is an integral area of computer science in which machine learning and computational linguistics are broadly used. The area of NLP involves making computer systems to perform meaningful tasks

International Research Journal IJNRD Research Through Innovation

with the natural and human understandable language.

Thus the input can be speech, text or image where output of an NLP system can be processed Speech as well as Written Text.

2. RELATED WORK:

2.1 Summarization Improvements

Since deep learning algorithms were not widely used in 2007, experimenters try to incorporate rhetorical information into their lessons to help develop their problem-solving skills (Zhang, Chan, &amp; Fung, 2007).This leads to normal performance while creating a product, concluding that the process is implicit but must be learned (Zhang, Chan, &amp; Fung, 2007). Six years later, an engineer/planner created a commercial product called "OpenEssayist" that displays content and highlights from student essays to help students complete tasks (Van Labeke, Whitelock, Field, Pulman, &amp; Richardson, 2013).

This product includes various content selection algorithms such as TextRank for keywords and content extraction (Van Labeke, Whitelock, Field, Pulman, &amp; Richardson, 2013). This is not as in previous studies, but to help students learn important topics, sentences, etc. from an article.

It will demonstrate the educational value of automated content aggregation that provides With a good start, the algorithm called TextRank uses the suggestion of the conversation's content.

Researchers Balasubramanian, Doraisamy, and Kanakarajan used this to create a similar application that uses the Naive Bayes algorithm to determine which sentences and content of a lecture or slide are the most descriptive (Balasubramanian, Doraisamy, &amp; Kanakarajan, 2016).

In recent documentation there have been several attempts to make class content without class labels. Two popular ideas are to extract text from whiteboards or slides and then use that information to create content. In a research project, the authors developed a tool that uses deep learning to extract written content from plain text and convert it to text for additional content (Kota, Davila, Stone, Setlur, &amp; Govindaraju, 2018).

2.2 Moving towards deep learning

While no deep learning has been done on the doctrines themselves, this is one of the first studies to use some form of deep learning to extract data for sermons.

In a project focused on extracting information from slides, the authors used video and audio tools to extract content, then used TF-IDF to Extract content and terms for final summary (Shimada, Okubo, Yin, &amp; Ogata, 2018).

In Kota et al, the authors used several methods in the case for initial extraction, but finally chose the NLP technique for the final content.

3. OBJECTIVE:

The main objective of this project is to automatically generate a shorter, condensed version of a longer piece of text while retaining its most important information. The goal is to provide the reader with a quick and comprehensive overview of the original text without having to read the entire document.

Text summarization can be applied to a wide range of applications, such as news articles, research papers, legal documents, and social media posts. Summarization can be performed either extractively or abstractively.

Extractive summarization helps in identifying the most important sentences or phrases from the original text andform a summary. It relies on most statistical techniques which helps in identifying.

Abstractive summarization, on the other hand, involves generating a summary by paraphrasing and synthesizing the most iportant information from the original text. This approach is more challenging as it requires a deep understanding of the content and context of the text, and often involves using advanced NLP techniques such as natural language generation.

The objective of text summarization using NLP is to save time and effort for readers by providing a condensed version of the text while preserving its most important information. It can also help in information retrieval and knowledge management, by enabling users to quickly identify relevant information from a large corpus of documents.

4. Methodology:

4.1. Data Collection: We have collected data from various sites where we have implemented frequency based and position based summary in which few test cases to examine it, later on we have implemented using the you-tube video transcript summarization and Hindi text- summarization techniques. Using the tracking system, we will focus on the specific content while checking the content of the article.

4.1 Data Pre Processing:

Once the data was collected, We cleaned and pre-processed it to remove noise and irrelevant information Following steps were performed:

1.Text cleaning: This involves removing noise from the raw text data, such as punctuation marks, special characters, and stop words.

2.Tokenization: This involves splitting the text data into individual words or phrases, known as tokens. This step makes it easier to analyze the text data.

3.Stop words removal: This involves removing common words such as "and," "the," "a," etc., which do not add any significant meaning to the text data.

4.Stemming/Lemmatization: This involves reducing words to their root forms, known as stems or lemmas. This step helps to reduce the number of words that need to be processed and improves the accuracy of the text summarization.

5.Part of Speech (POS) tagging: This involves identifying the part of speech for each word in the text data. This step helps to identify important keywords and phrases that should be included in the summary.

6.Named Entity Recognition (NER): This involves identifying and categorizing named entities such as people, places, organizations, etc., in the text data. This step helps to identify important information that should be included in the summary.

4.2 Sentence Selection:

In extractive summarization, the next step is to identify the most important sentences from the original text. Various techniques are applied such as:

1. Frequency analysis: This involves selecting sentences that contain the most frequently occurring keywords or phrases.

2.Clustering: This involves grouping sentences based on their similarity and selecting representative sentences from each cluster.

3.Graph-based algorithms: This involves representing the sentences as nodes in a graph and selecting the most important sentences based on their centrality in the graph.

The selected sentences are then combined to create the summary.

4.3 Sentence Compression:

In some cases, the selected sentences may be too long for a summary. In such cases, sentence compression techniques can be used to shorten the sentences while preserving their meaning. Some popular sentence compression techniques include:

1. Deleting non-essential words: This involves removing words that do not contribute to the meaning of the sentence.
2. Merging similar sentences: This involves combining two or more sentences that convey similar information.
3. Splitting long sentences: This involves breaking down long sentences into shorter ones while preserving their meaning.

4.4 Sentence Synthesis:

In abstractive summarization, the next step is to synthesize new sentences that capture the essence of the original text. Techniques:

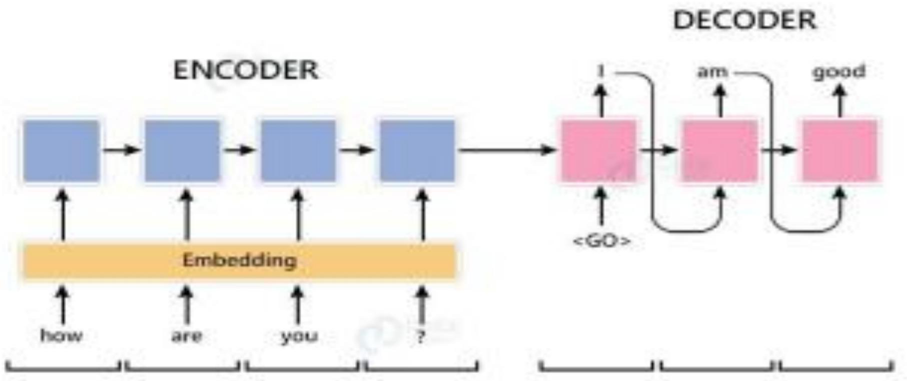
1. Deep learning-based models: These models use neural networks to generate new sentences based on the content of the original text.
2. Rule-based methods: These methods use a set of predefined rules to generate new sentences based on the content of the original text.
3. Natural language generation (NLG) techniques: These techniques use a combination of deep learning-based models and rule-based methods to generate new sentences.

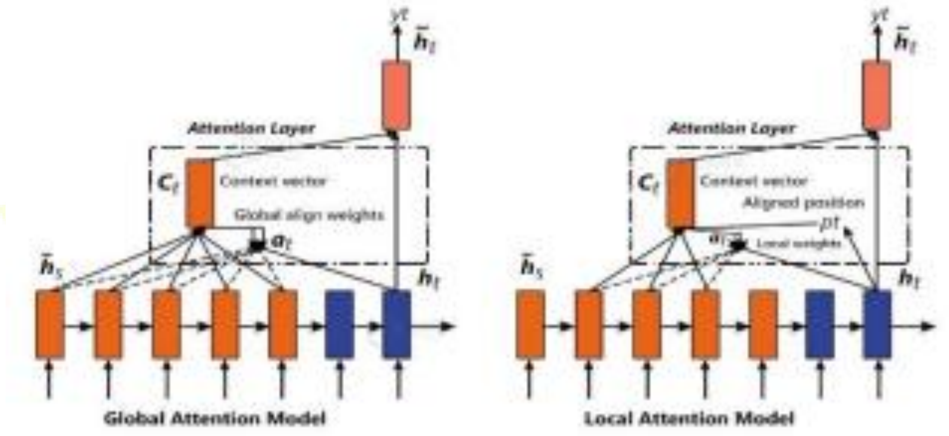
4.5 Evaluation:

The final step in text summarization using NLP is to evaluate the quality of the generated summary. Evaluation metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy) are commonly used to measure the similarity between the generated summary and the original text. The higher the ROUGE or BLEU score, the better the quality of the generated summary.

The methodology of text summarization using NLP involves preprocessing the text data, selecting or synthesizing important sentences, compressing or synthesizing sentences as needed, and evaluating the quality of the generated summary. The choice of approach (extractive or abstractive) depends on the specific requirements.

5. WORK FLOW:

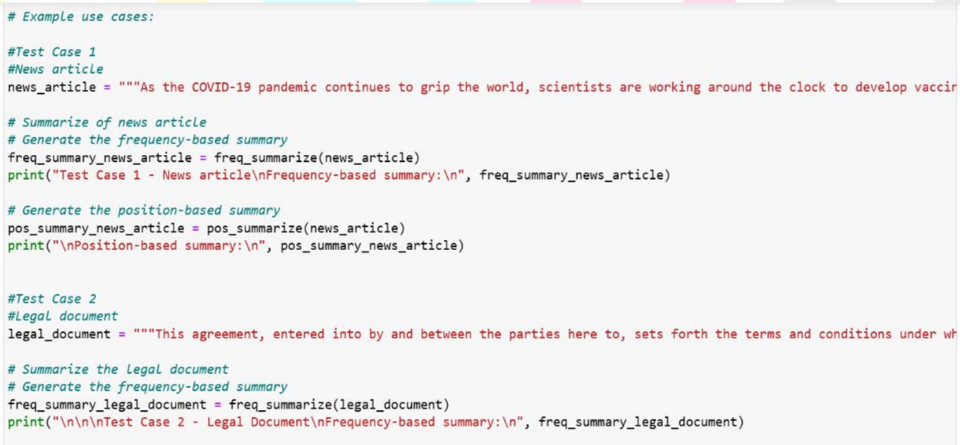


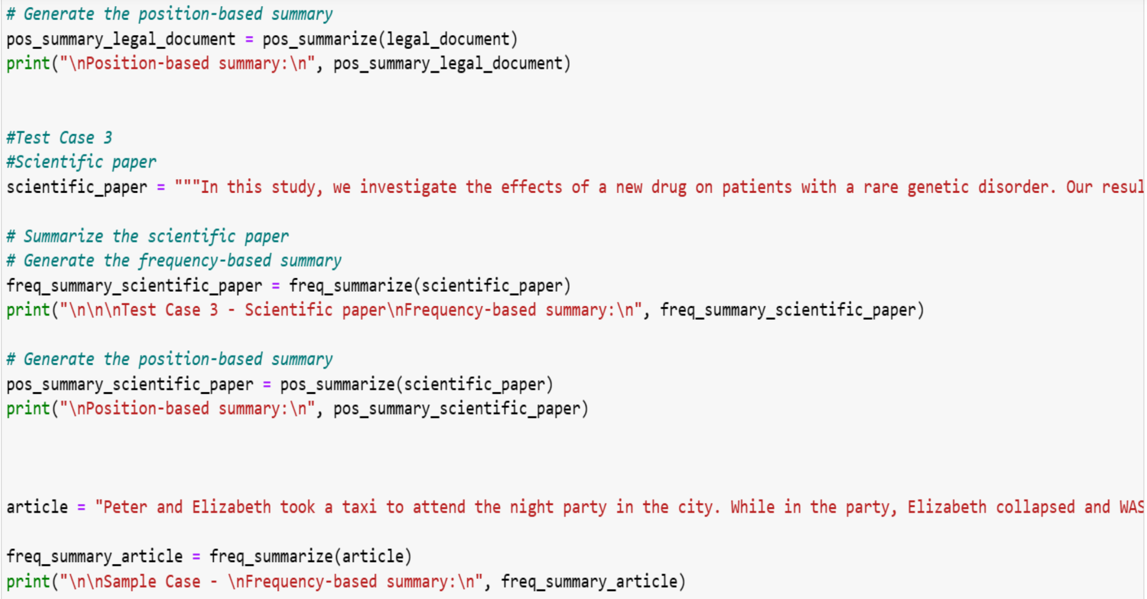


International Research Journal

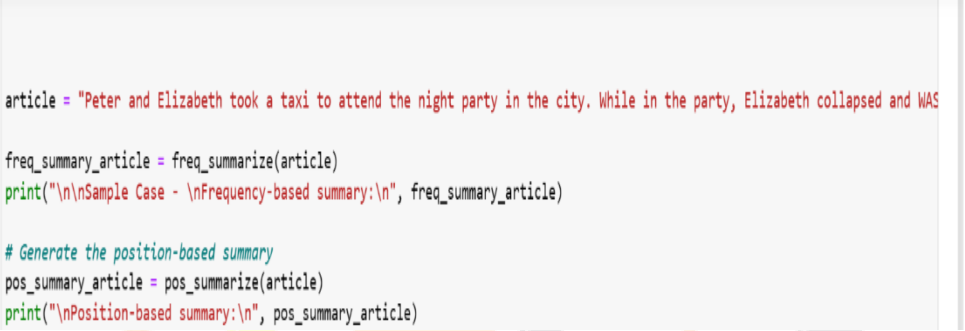
6. OBSERVATION:

6.1 Frequency and Position Based Summary:


```
# Example use cases:

#Test Case 1
#News article
news_article = """As the COVID-19 pandemic continues to grip the world, scientists are working around the clock to develop vaccir

# Summarize of news article
# Generate the frequency-based summary
freq_summary_news_article = freq_summarize(news_article)
print("Test Case 1 - News article\nFrequency-based summary:\n", freq_summary_news_article)

# Generate the position-based summary
pos_summary_news_article = pos_summarize(news_article)
print("\nPosition-based summary:\n", pos_summary_news_article)


#Test Case 2
#Legal document
legal_document = """This agreement, entered into by and between the parties here to, sets forth the terms and conditions under wh

# Summarize the legal document
# Generate the frequency-based summary
freq_summary_legal_document = freq_summarize(legal_document)
print("\n\n\nTest Case 2 - Legal Document\nFrequency-based summary:\n", freq_summary_legal_document)
```

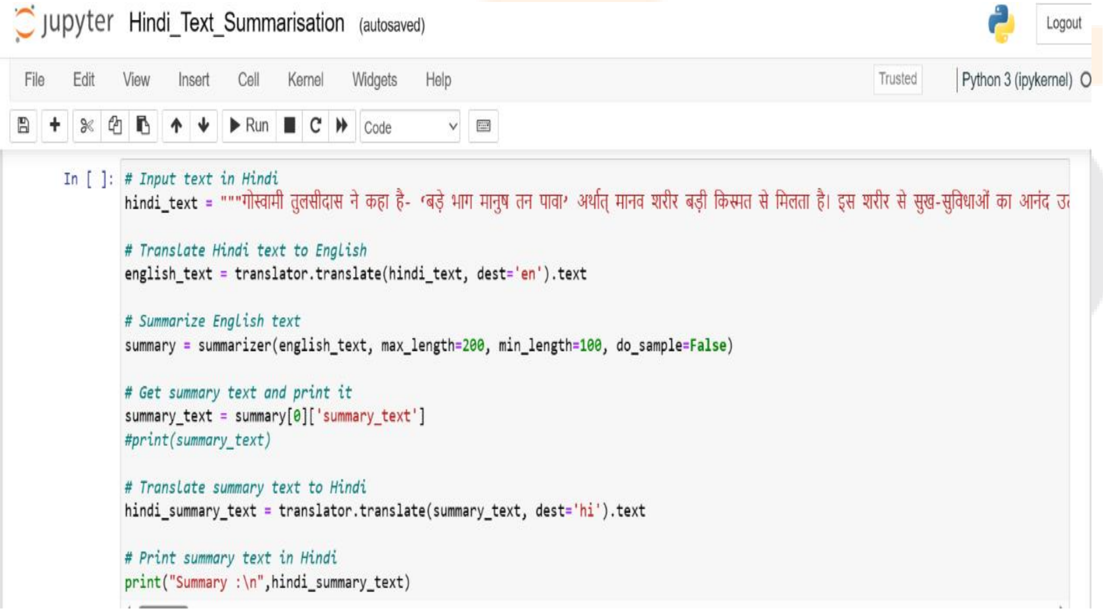


```
# Generate the position-based summary
pos_summary_legal_document = pos_summarize(legal_document)
print("\nPosition-based summary:\n", pos_summary_legal_document)


#Test Case 3
#Scientific paper
scientific_paper = """In this study, we investigate the effects of a new drug on patients with a rare genetic disorder. Our resul

# Summarize the scientific paper
# Generate the frequency-based summary
freq_summary_scientific_paper = freq_summarize(scientific_paper)
print("\n\n\nTest Case 3 - Scientific paper\nFrequency-based summary:\n", freq_summary_scientific_paper)

# Generate the position-based summary
pos_summary_scientific_paper = pos_summarize(scientific_paper)
print("\nPosition-based summary:\n", pos_summary_scientific_paper)



article = "Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and WAS

freq_summary_article = freq_summarize(article)
print("\n\nSample Case - \nFrequency-based summary:\n", freq_summary_article)
```



```
article = "Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and WAS

freq_summary_article = freq_summarize(article)
print("\n\nSample Case - \nFrequency-based summary:\n", freq_summary_article)

# Generate the position-based summary
pos_summary_article = pos_summarize(article)
print("\nPosition-based summary:\n", pos_summary_article)
```

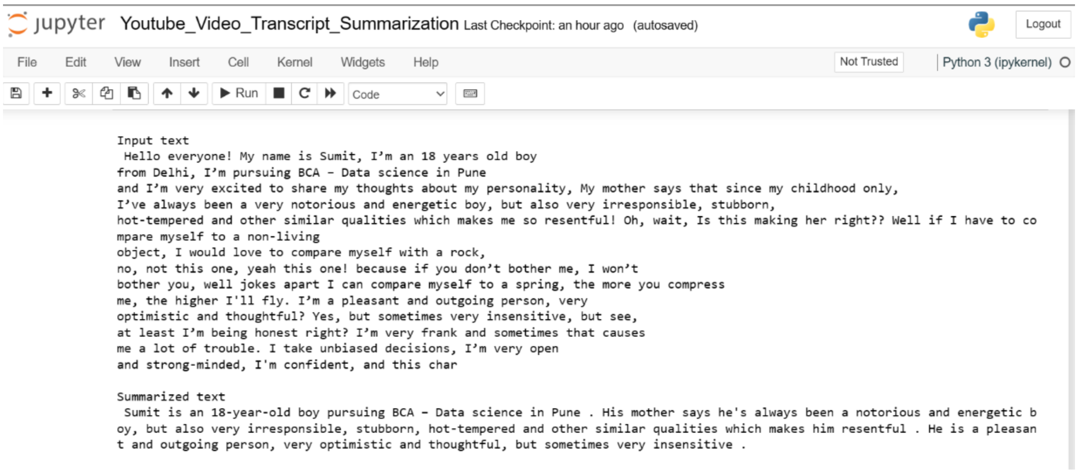
6.2 Hindi Text Summarization:



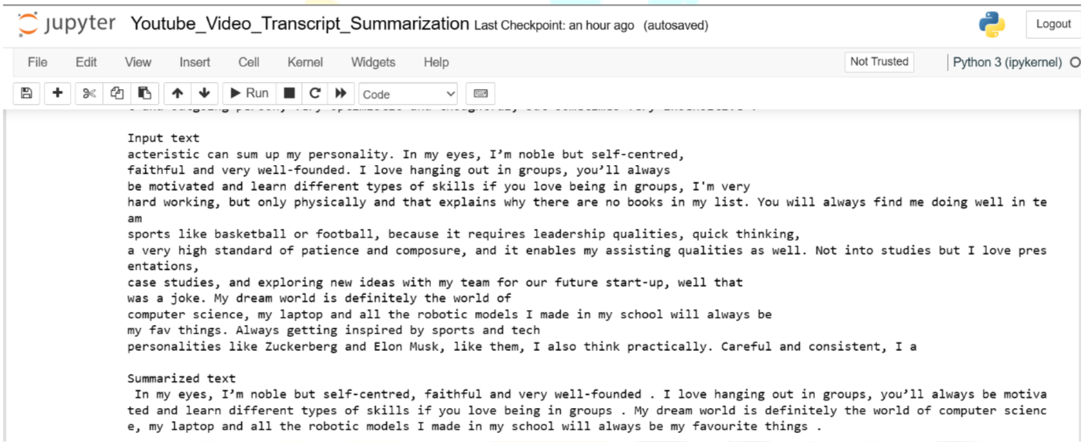
jupyter Hindi_Text_Summarisation (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [ ]: # Input text in Hindi
        hindi_text = """गोस्वामी तुलसीदास ने कहा है- ‘बड़े भाग मानुष तन पावा’ अर्थात् मानव शरीर बड़ी किस्मत से मिलता है। इस शरीर से सुख-सुविधाओं का आनंद उ

        # Translate Hindi text to English
        english_text = translator.translate(hindi_text, dest='en').text

        # Summarize English text
        summary = summarizer(english_text, max_length=200, min_length=100, do_sample=False)

        # Get summary text and print it
        summary_text = summary[0]['summary_text']
        #print(summary_text)

        # Translate summary text to Hindi
        hindi_summary_text = translator.translate(summary_text, dest='hi').text

        # Print summary text in Hindi
        print("Summary :\n",hindi_summary_text)
```

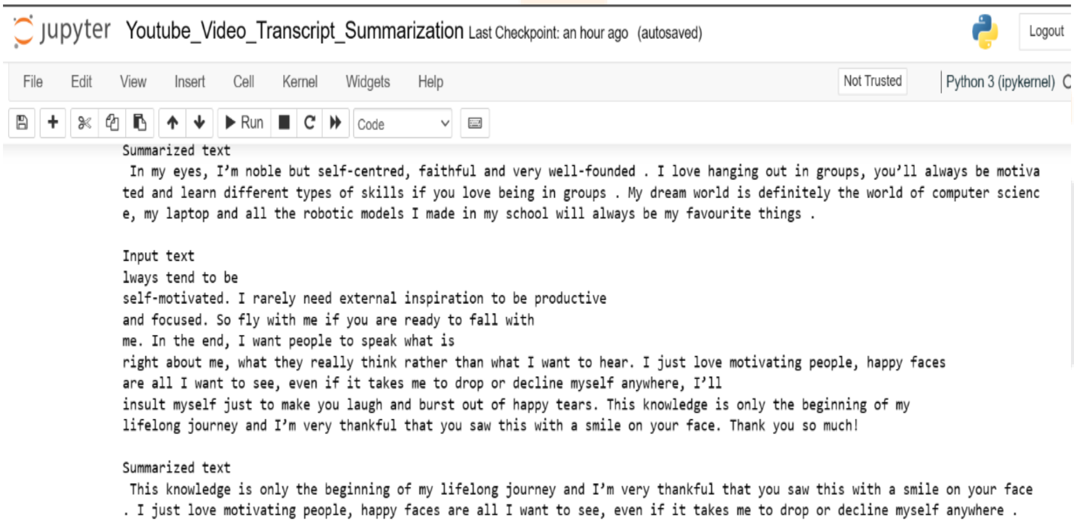
6.3 You-Tube Video Summarization:


jupyter
Youtube_Video_Transcript_Summarization Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
Input text
 Hello everyone! My name is Sumit, I’m an 18 years old boy
from Delhi, I’m pursuing BCA – Data science in Pune
and I’m very excited to share my thoughts about my personality, My mother says that since my childhood only,
I’ve always been a very notorious and energetic boy, but also very irresponsible, stubborn,
hot-tempered and other similar qualities which makes me so resentful! Oh, wait, Is this making her right?? Well if I have to co
mpare myself to a non-living
object, I would love to compare myself with a rock,
no, not this one, yeah this one! because if you don’t bother me, I won’t
bother you, well jokes apart I can compare myself to a spring, the more you compress
me, the higher I'll fly. I’m a pleasant and outgoing person, very
optimistic and thoughtful? Yes, but sometimes very insensitive, but see,
at least I’m being honest right? I’m very frank and sometimes that causes
me a lot of trouble. I take unbiased decisions, I’m very open
and strong-minded, I'm confident, and this char

Summarized text
 Sumit is an 18-year-old boy pursuing BCA – Data science in Pune . His mother says he's always been a notorious and energetic b
oy, but also very irresponsible, stubborn, hot-tempered and other similar qualities which makes him resentful . He is a pleasan
t and outgoing person, very optimistic and thoughtful, but sometimes very insensitive .
```


jupyter
Youtube_Video_Transcript_Summarization Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

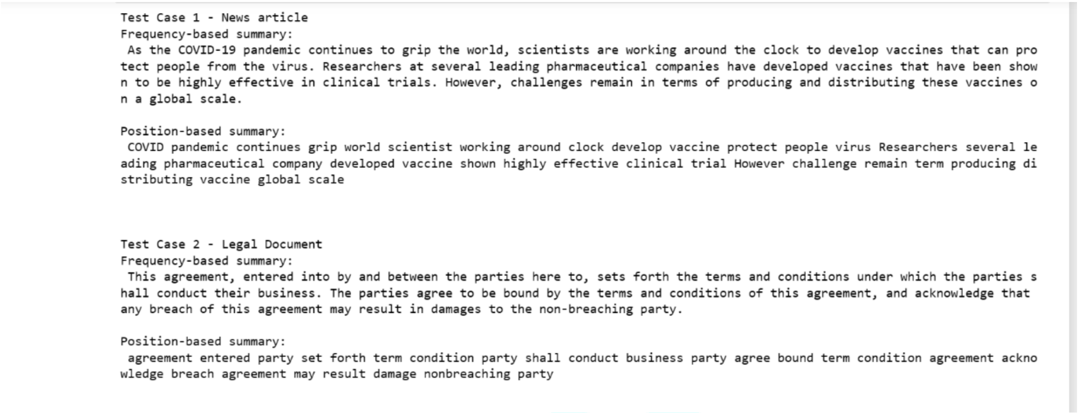
```
Input text
acteristic can sum up my personality. In my eyes, I’m noble but self-centred,
faithful and very well-founded. I love hanging out in groups, you’ll always
be motivated and learn different types of skills if you love being in groups, I'm very
hard working, but only physically and that explains why there are no books in my list. You will always find me doing well in te
am
sports like basketball or football, because it requires leadership qualities, quick thinking,
a very high standard of patience and composure, and it enables my assisting qualities as well. Not into studies but I love pres
entations,
case studies, and exploring new ideas with my team for our future start-up, well that
was a joke. My dream world is definitely the world of
computer science, my laptop and all the robotic models I made in my school will always be
my fav things. Always getting inspired by sports and tech
personalities like Zuckerberg and Elon Musk, like them, I also think practically. Careful and consistent, I a

Summarized text
 In my eyes, I’m noble but self-centred, faithful and very well-founded . I love hanging out in groups, you’ll always be motiva
ted and learn different types of skills if you love being in groups . My dream world is definitely the world of computer scienc
e, my laptop and all the robotic models I made in my school will always be my favourite things .
```


jupyter
Youtube_Video_Transcript_Summarization Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

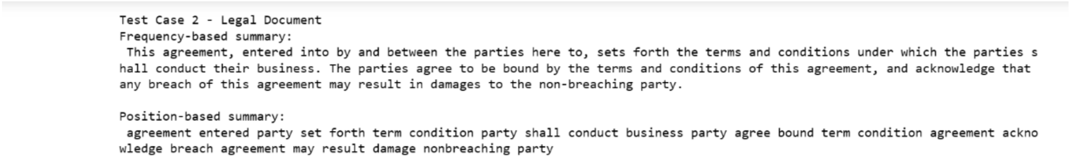
```
Summarized text
 In my eyes, I’m noble but self-centred, faithful and very well-founded . I love hanging out in groups, you’ll always be motiva
ted and learn different types of skills if you love being in groups . My dream world is definitely the world of computer scienc
e, my laptop and all the robotic models I made in my school will always be my favourite things .

Input text
lways tend to be
self-motivated. I rarely need external inspiration to be productive
and focused. So fly with me if you are ready to fall with
me. In the end, I want people to speak what is
right about me, what they really think rather than what I want to hear. I just love motivating people, happy faces
are all I want to see, even if it takes me to drop or decline myself anywhere, I’ll
insult myself just to make you laugh and burst out of happy tears. This knowledge is only the beginning of my
lifelong journey and I’m very thankful that you saw this with a smile on your face. Thank you so much!

Summarized text
 This knowledge is only the beginning of my lifelong journey and I’m very thankful that you saw this with a smile on your face
. I just love motivating people, happy faces are all I want to see, even if it takes me to drop or decline myself anywhere .
```

7. Experiments and Results:

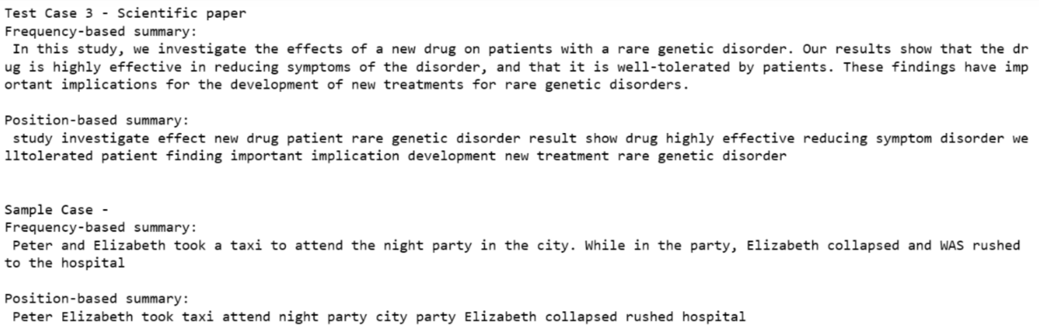
7.1 Frequency and Position Based Summary:



```
Test Case 1 - News article
Frequency-based summary:
 As the COVID-19 pandemic continues to grip the world, scientists are working around the clock to develop vaccines that can pro
tect people from the virus. Researchers at several leading pharmaceutical companies have developed vaccines that have been show
n to be highly effective in clinical trials. However, challenges remain in terms of producing and distributing these vaccines o
n a global scale.

Position-based summary:
 COVID pandemic continues grip world scientist working around clock develop vaccine protect people virus Researchers several le
ading pharmaceutical company developed vaccine shown highly effective clinical trial However challenge remain term producing di
stributing vaccine global scale


Test Case 2 - Legal Document
Frequency-based summary:
 This agreement, entered into by and between the parties here to, sets forth the terms and conditions under which the parties s
hall conduct their business. The parties agree to be bound by the terms and conditions of this agreement, and acknowledge that
any breach of this agreement may result in damages to the non-breaching party.

Position-based summary:
 agreement entered party set forth term condition party shall conduct business party agree bound term condition agreement ackno
wledge breach agreement may result damage nonbreaching party
```

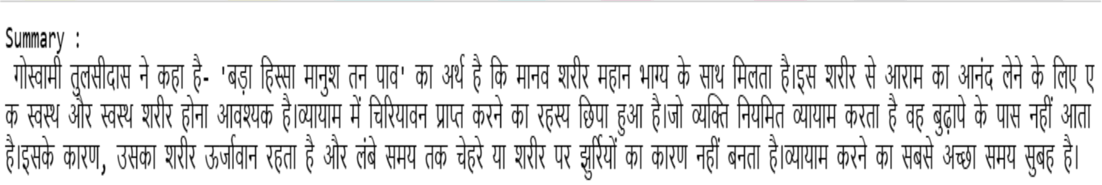


```
Test Case 2 - Legal Document
Frequency-based summary:
 This agreement, entered into by and between the parties here to, sets forth the terms and conditions under which the parties s
hall conduct their business. The parties agree to be bound by the terms and conditions of this agreement, and acknowledge that
any breach of this agreement may result in damages to the non-breaching party.

Position-based summary:
 agreement entered party set forth term condition party shall conduct business party agree bound term condition agreement ackno
wledge breach agreement may result damage nonbreaching party
```



```
Test Case 3 - Scientific paper
Frequency-based summary:
 In this study, we investigate the effects of a new drug on patients with a rare genetic disorder. Our results show that the dr
ug is highly effective in reducing symptoms of the disorder, and that it is well-tolerated by patients. These findings have imp
ortant implications for the development of new treatments for rare genetic disorders.

Position-based summary:
 study investigate effect new drug patient rare genetic disorder result show drug highly effective reducing symptom disorder we
lltolerated patient finding important implication development new treatment rare genetic disorder


Sample Case -
Frequency-based summary:
 Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and WAS rushed
to the hospital

Position-based summary:
 Peter Elizabeth took taxi attend night party city party Elizabeth collapsed rushed hospital
```

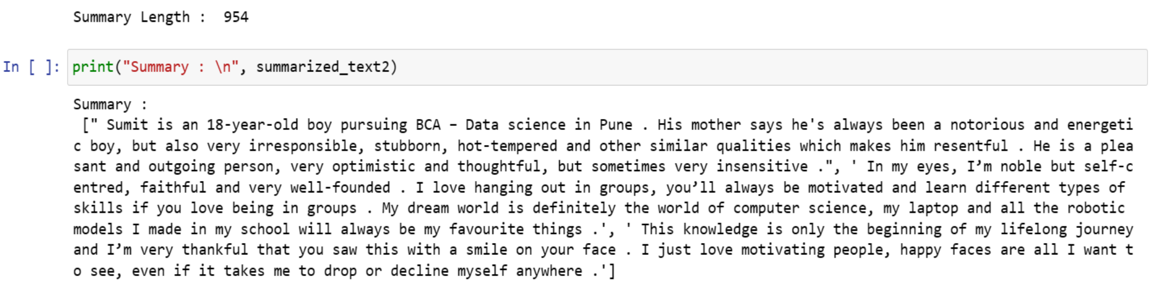
7.2 Hindi Text Summarization:



```
Summary :
 गोस्वामी तुलसीदास ने कहा है- 'बड़ा हिस्सा मानुश तन पाव' का अर्थ है कि मानव शरीर महान भाग्य के साथ मिलता है।इस शरीर से आराम का आनंद लेने के लिए ए
क स्वस्थ और स्वस्थ शरीर होना आवश्यक है।व्यायाम में चिरियावन प्राप्त करने का रहस्य छिपा हुआ है।जो व्यक्ति नियमित व्यायाम करता है वह बुढ़ापे के पास नहीं आता
है।इसके कारण, उसका शरीर ऊर्जावान रहता है और लंबे समय तक चेहरे या शरीर पर झुर्रियों का कारण नहीं बनता है।व्यायाम करने का सबसे अच्छा समय सुबह है।
```

Research Through Innovation

7.3 You-Tube Video Summarization:



```
Summary Length :  954

In [ ]: print("Summary : \n", summarized_text2)

Summary :
 [" Sumit is an 18-year-old boy pursuing BCA - Data science in Pune . His mother says he's always been a notorious and energeti
c boy, but also very irresponsible, stubborn, hot-tempered and other similar qualities which makes him resentful . He is a plea
sant and outgoing person, very optimistic and thoughtful, but sometimes very insensitive .", ' In my eyes, I’m noble but self-c
entred, faithful and very well-founded . I love hanging out in groups, you’ll always be motivated and learn different types of
skills if you love being in groups . My dream world is definitely the world of computer science, my laptop and all the robotic
models I made in my school will always be my favourite things .', ' This knowledge is only the beginning of my lifelong journey
and I’m very thankful that you saw this with a smile on your face . I just love motivating people, happy faces are all I want t
o see, even if it takes me to drop or decline myself anywhere .']
```

8. CONCLUSION:

Text summarization using NLP is a powerful tool that can help save time and effort for readers by providing a condensed version of a longer text while retaining its most important information. It can be applied to a wide range of applications, from news articles to research papers to legal documents, and can be performed using either extractive or abstractive techniques.

While extractive summarization is relatively straightforward and relies on statistical and linguistic techniques to identify the most important sentences or phrases from the original text, abstractive summarization is more challenging and requires a deep understanding of the content and context of the text. However, abstractive summarization can provide more concise and human-like summaries by synthesizing and paraphrasing the most important information from the original text.

Overall, text summarization using NLP has a wide range of applications and can help improve efficiency in various industries and domains. As NLP technology continues to advance, we can expect text summarization to become more accurate and effective, leading to further improvements in information retrieval and knowledge management.

9. REFERENCES:

- (PDF) Text Summarizing Using NLP (researchgate.net)
- Gunes Erkan and Dragomir R. Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In¨ Proceedings of EMNLP, pages 365–371
- https://data-flair.training/blogs/machine-learning-text-summarization/
- https://paperswithcode.com/task/text-summarization

International Research Journal
IJNRD
Research Through Innovation