Epitome
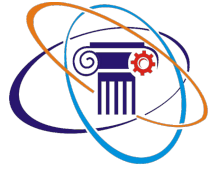
Let's make it easy...

Sarthak Parakh 0827CS171196
Shivam Goyan 0827CS171204
Somya Jain 0827CS171215
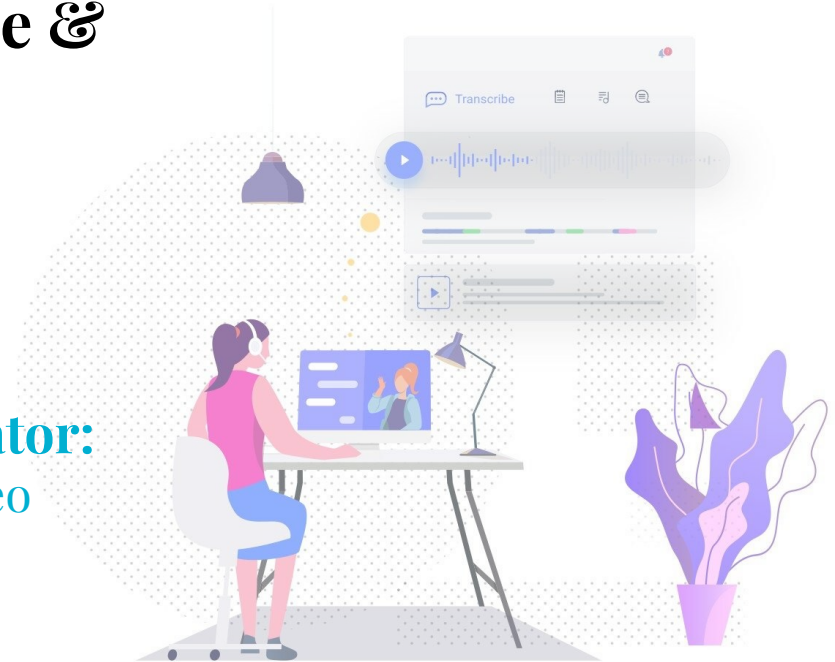
Department of Computer Science & Engineering

Major Project
Project Name: Epitome

Project Guide:
Prof. Kavita Namdeo
(Asst. Professor)

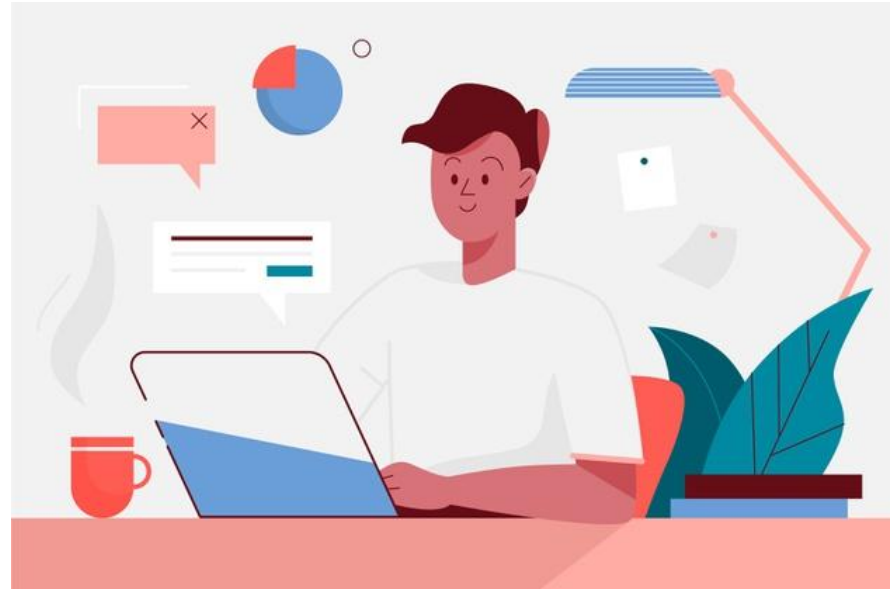**Project Coordinator:**
Prof. Kavita Namdeo
(Asst. Professor)

Aim & Utility

- ❏ Linking professionals and students to the new age technology.
- ❏ To provide instant notes of a digital meetings/classrooms.
- ❏ Bringing more attention in meetings/classrooms.
- ❏ We aim to provide a basic interface which can be used by anyone during online/offline meetings by supplying them the summaries of the text from which they want to gain information.

Introduction

- ❖ The goal of Epitome is to produce a shorter version of a source speech by preserving the meaning and the key contents of the original speech.
- ❖ A well written summary can significantly reduce the amount of work needed to digest large amount of text.

Epitome

Epitome is one of those applications of Natural Language Processing (NLP) which is bound to have a huge impact on our lives. With growing digital media and ever growing publishing – who has the time to go through entire articles / documents / books to decide whether they are useful or not?

Problem Statement

Naturally, there is so much critical information being said via voice that gets lost once we leave the meeting room. No matter how hard we try to multitask, it's impossible to remember everything, type everything, and also stay present while the meeting is happening.

Case Study

It was a hectic day when my sister was having multiple classes in a single day, she was not in the mood to attend any further classes, either if she did she wasn't able to retain it for long. So I asked her if I can build an application that can help her create notes for her class.

Types of Text Summarization

There are broadly two different approaches that are used for text summarization:

- Extractive Summarization
- Abstractive Summarization
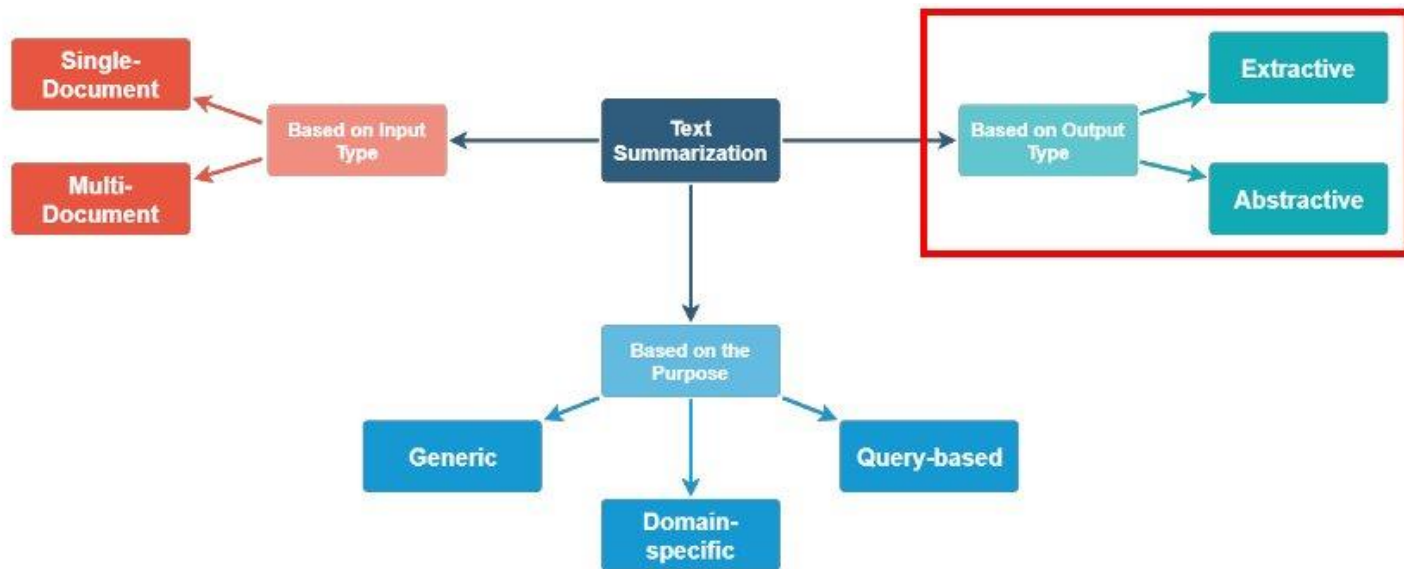
Extractive Summarization

The name gives away what this approach does. We identify the important sentences or phrases from the original text and extract only those from the text. Those extracted sentences would be our summary. The below diagram illustrates extractive summarization:
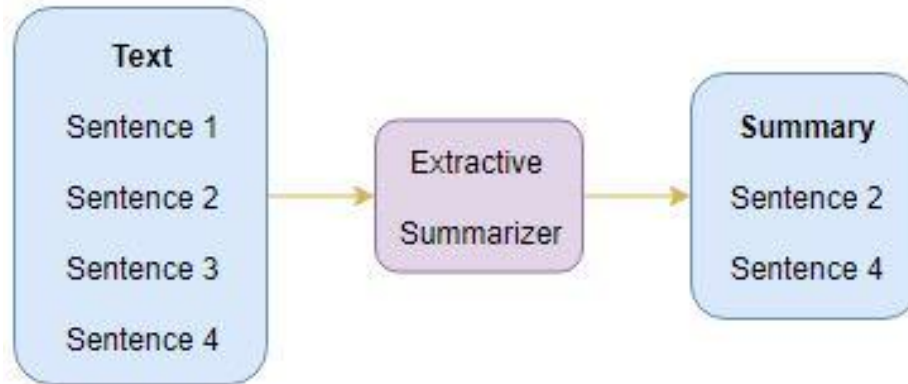
Abstractive Summarization

This is a very interesting approach. Here, we generate new sentences from the original text. This is in contrast to the extractive approach we saw earlier where we used only the sentences that were present. The sentences generated through abstractive summarization might not be present in the original text.
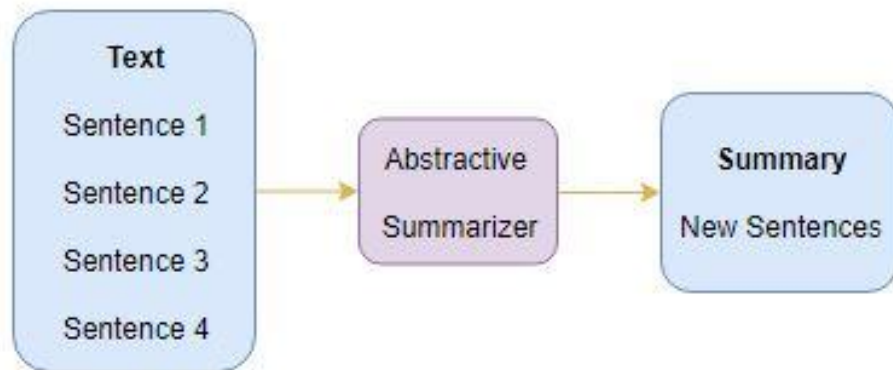
Our Proposed Solution

The project is to create a product which joins Google, Zoom, and other web conferencing meetings. While one is in the online/offline meeting speaking, Epitome is recording the audio of the conversation. Once the meeting concludes, the Epitome natural language processing (NLP) algorithms transcribe the audio into a copy of the text and skim it to provide the summary (key content) of the meeting rather than full transcription.

Existing Systems

❏ **<u>Fireflies.ai</u>** : Record, Transcribe and Search across your voice conversations to get the desired data.

❏ **<u>Chorus.ai</u>** : It is a Conversation Intelligence Platform that records, transcribes, and analyzes business conversations in real-time to coach reps on crushing their quota.
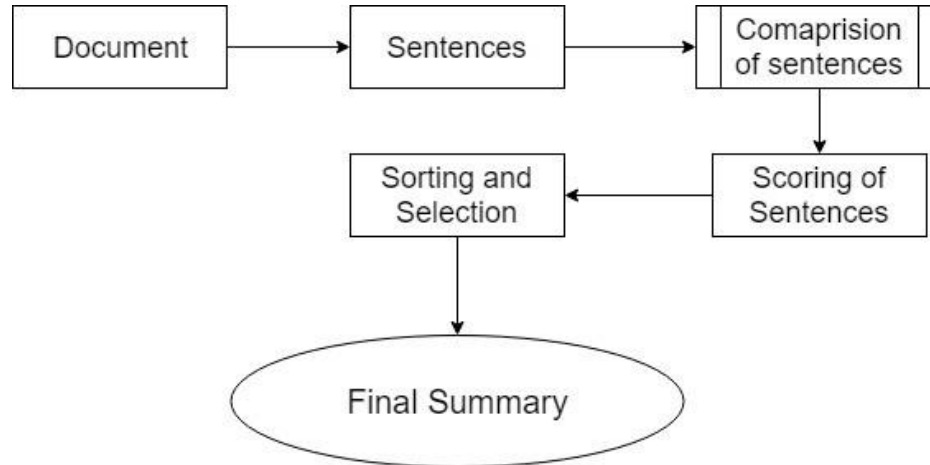
This application are very large level companies which is not available for every individual as their costs are too high(approx 1000$ per year).

Requirement Analysis

FUNCTIONAL REQUIREMENTS

To process the audio inserted by the user of the application into text and skim the text to provide a concise and fluent summary to the user. So the use of NLP is of extreme importance and for the application to work internet connection is required.

```
Document  →  Sentences  →  Comaprision of sentences
                                      ↓
Sorting and Selection  ←  Scoring of Sentences
        ↓
   Final Summary
```
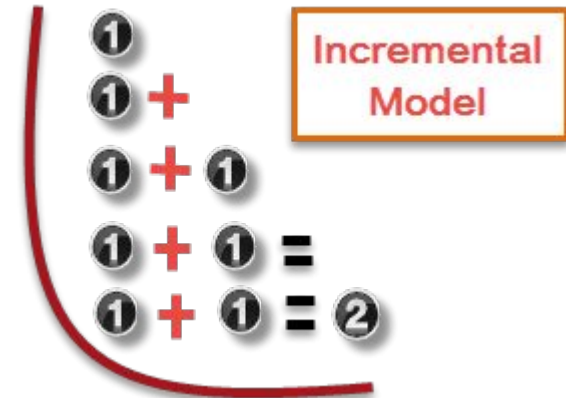
NON–FUNCTIONAL REQUIREMENTS

- **PERFORMANCE:** The application is lightweight and resulted in output available is concise and feasible.
- **RELIABILITY:** In case the user's devices crash, the user can run their text again to get the result.
- **AVAILABILITY:** The user should have Internet connection technology in the device and a well-charged device.
- **SECURITY:** Security systems need database storage just like many other applications. All the user data provided will be secured and protected in the application.
- **MAINTAINABILITY:** The main maintainability required is for the database and for processing algorithms which is done efficiently.
- **PORTABILITY:** The product is an android application so it can only be used on android platform operating systems.

SDLC Model

The Incremental Model is a process of software development where requirements are broken down into multiple standalone modules of software development cycle. Incremental development is done in steps from analysis design, implementation, testing/verification, maintenance
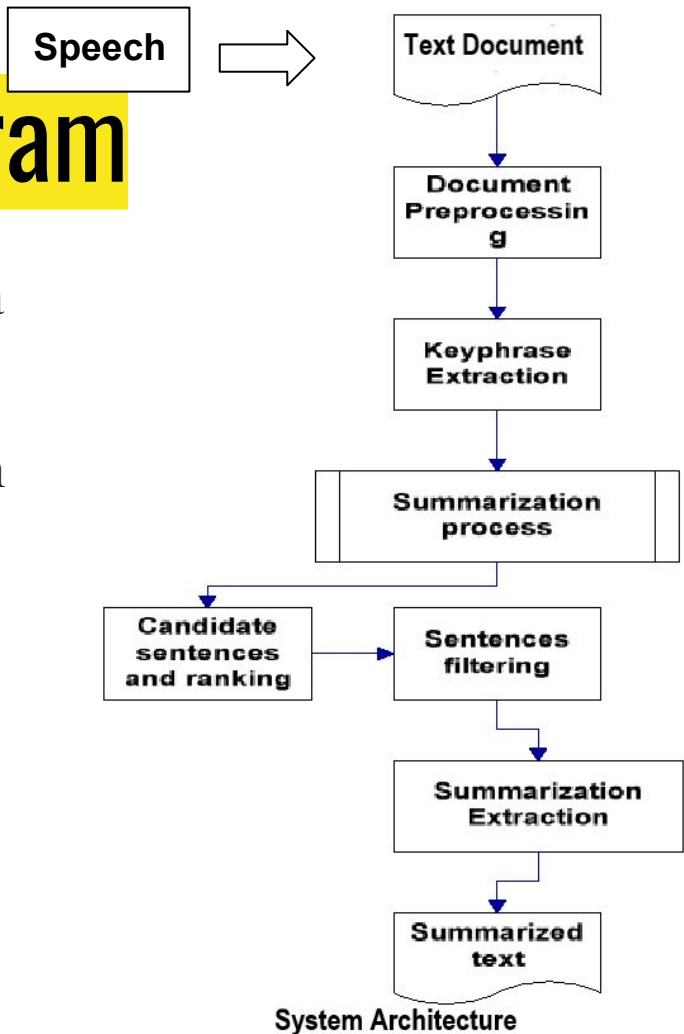
Analysis → Design → code → test → Increment-1

Analysis → Design → code → test → Increment-2

Analysis → Design → code → test → Increment-3
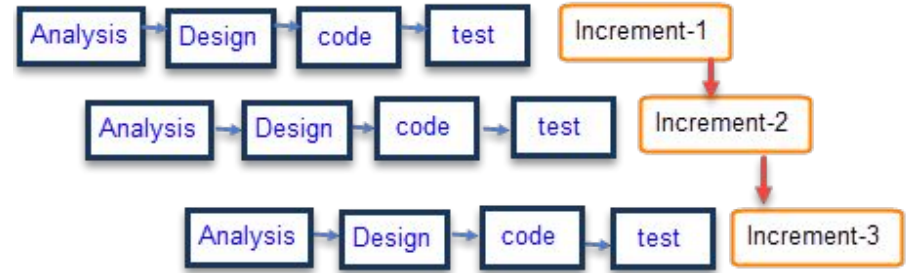
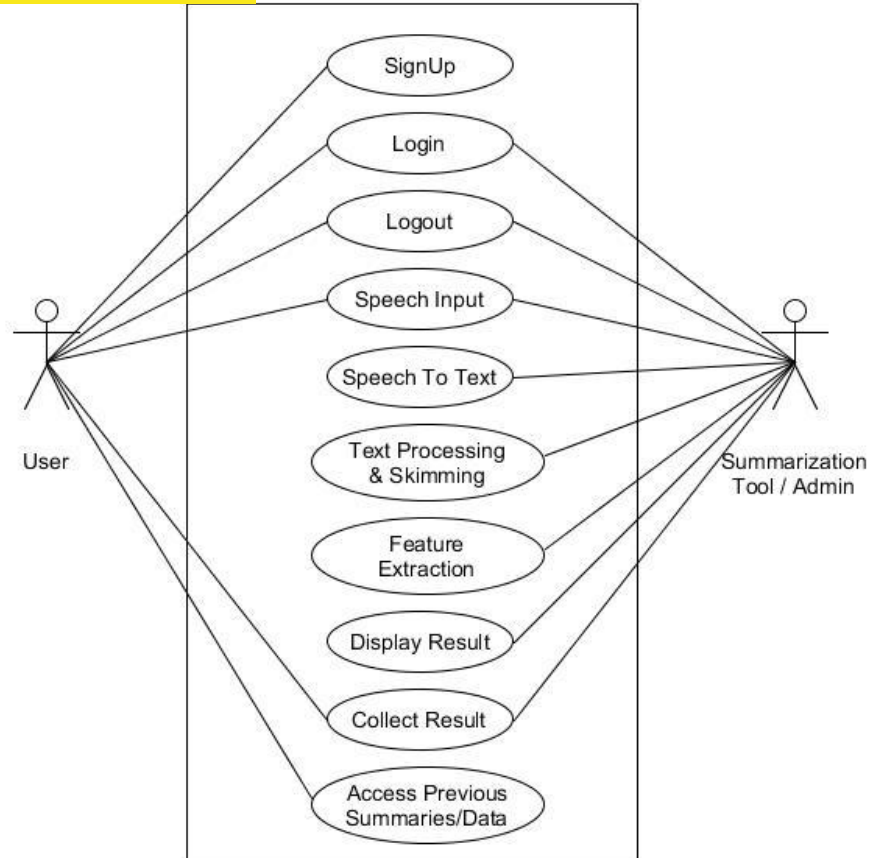**Incremental Model**

Incremental Model
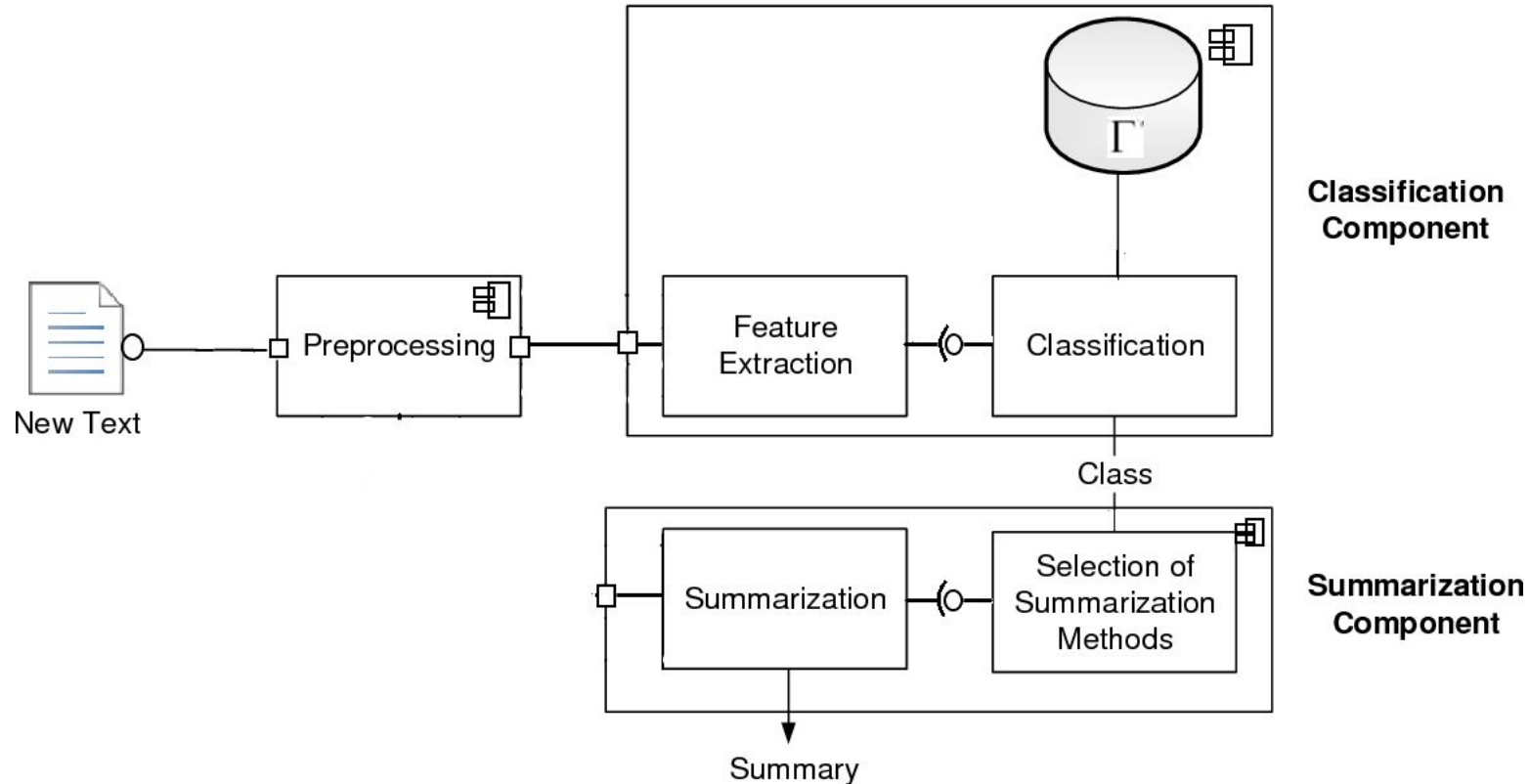
System Architecture Diagram

An architectural diagram is a diagram of a system that is used to abstract the overall outline of the software system and the relationships, constraints, and boundaries between components. It is an important tool as it provides an overall view of the physical deployment of the software system and its evolution roadmap

Speech ⟹ Text Document

```
Text Document
      │
      ▼
Document Preprocessing
      │
      ▼
Keyphrase Extraction
      │
      ▼
Summarization process
      │
      ▼
Candidate sentences and ranking  →  Sentences filtering
                                          │
                                          ▼
                                   Summarization Extraction
                                          │
                                          ▼
                                   Summarized text
```

System Architecture

Use Case Diagram

Component Diagram

New Text

Preprocessing

Feature Extraction

Classification

Classification Component

Class

Summarization

Selection of Summarization Methods

Summarization Component

Summary

E-R Diagram

Technology Used

MACHINE LEARNING

NLP Natural language processing

python django

DEEP LEARNING

DATA ANALYTICS
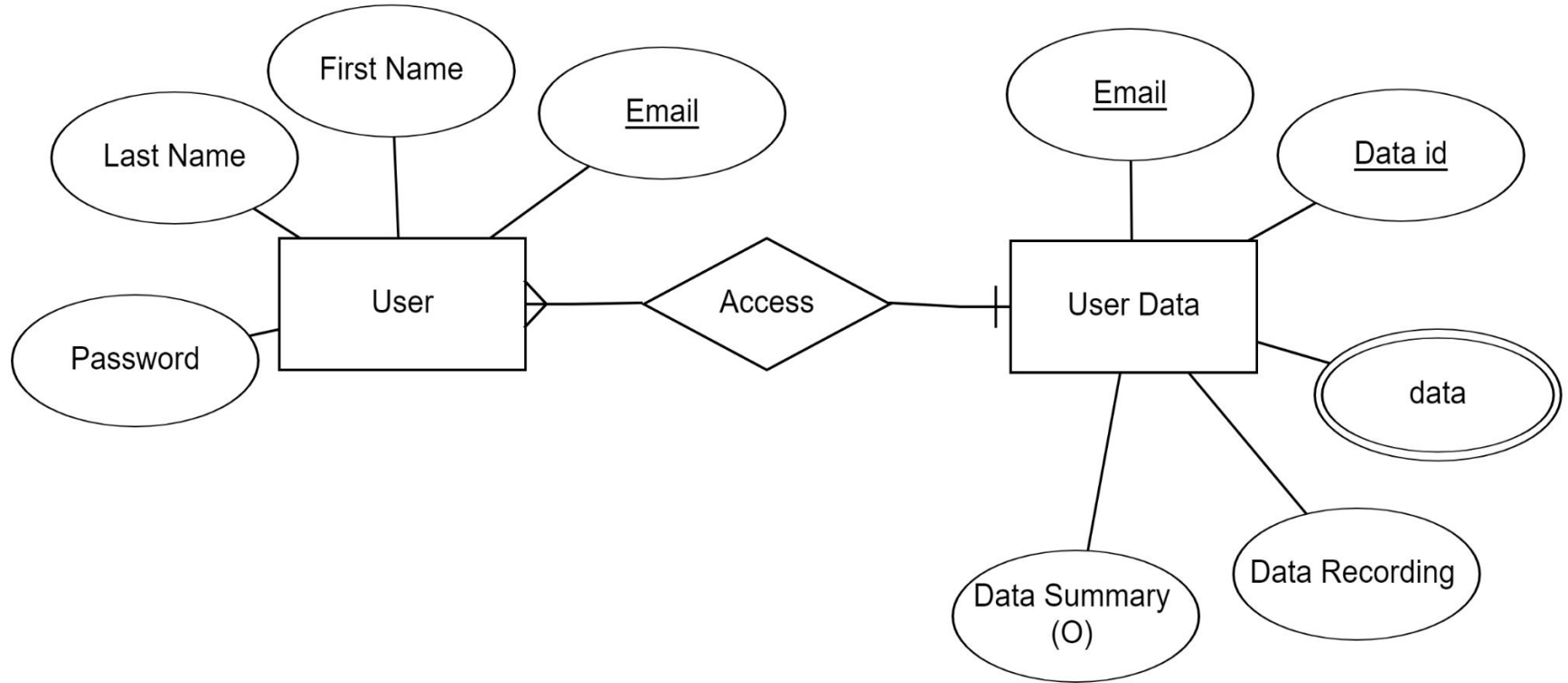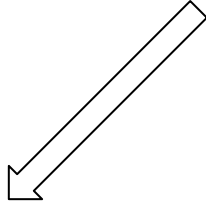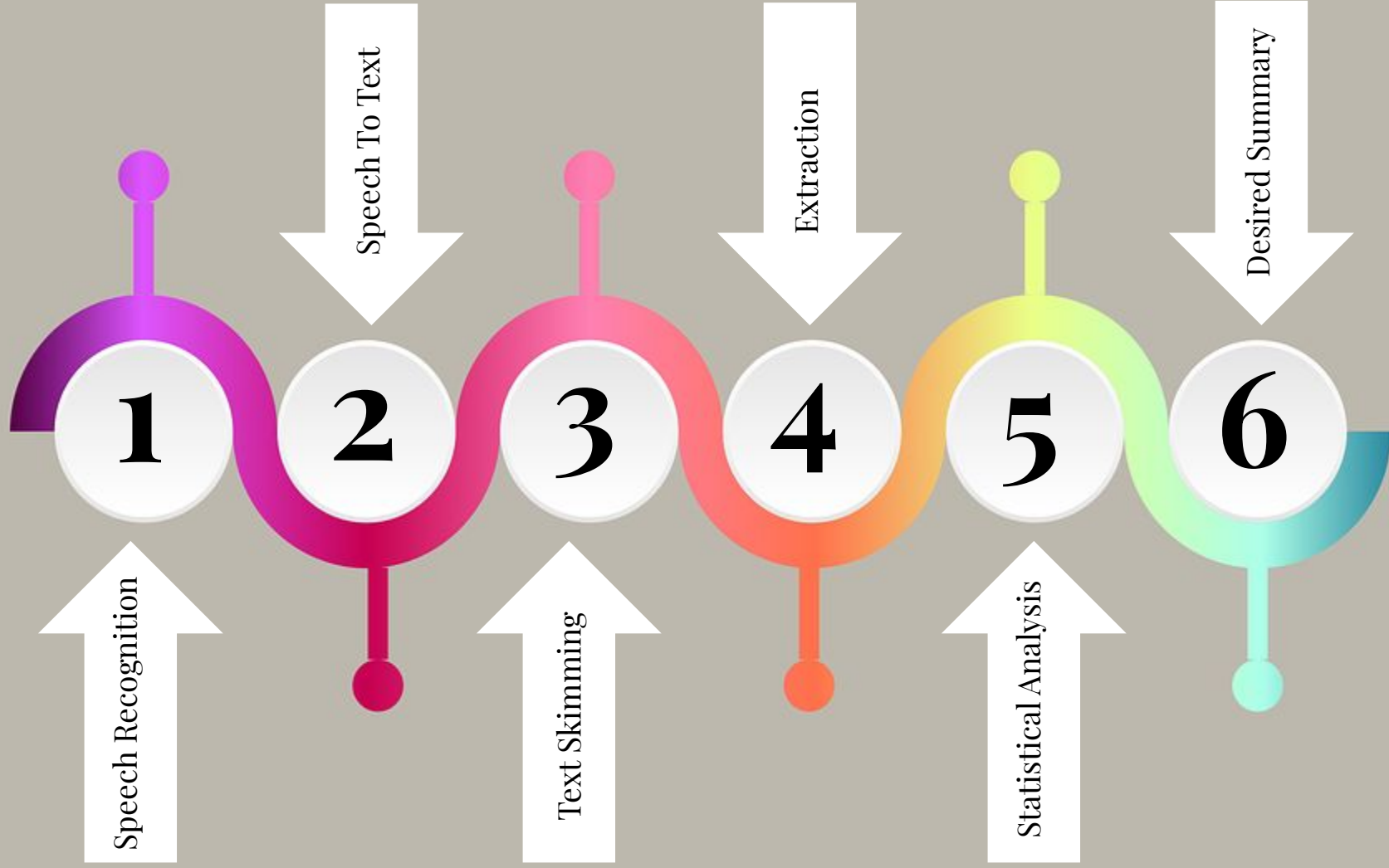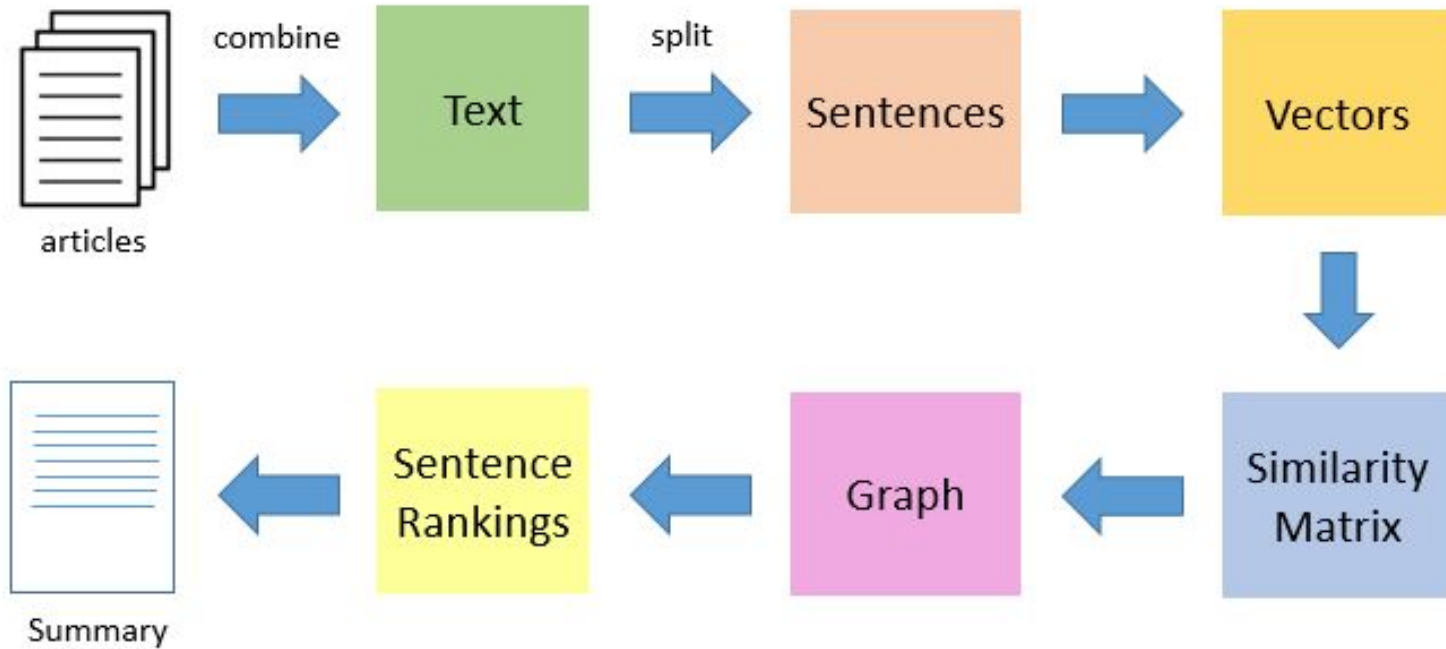
Process

Audio

Original Text

Text Summarization

Summarized output

Understanding the TextRank Algorithm
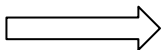
Implementation BackEnd

Speech
Input

Speech To
Text

```
In [2]: path = "C:/Users/fanta_000/Desktop/Sarthak/Major Project/Recording.wav"
        text=get_large_audio_transcription(path)
        print("\nFull text:",text)
```

audio-chunks\chunk1.wav : Machine learning is the study of computer algorithms that improve automatically through experience it is seen as a subset of artificial intelligence.
audio-chunks\chunk2.wav : Machine tools builder model based on sample data known as training data.
audio-chunks\chunk3.wav : How to make predictions or decisions without being explicitly programmed to do so.
audio-chunks\chunk4.wav : Machine codons and used in a wide variety of applications that is email filtering and computer vision.
audio-chunks\chunk5.wav : Where it is difficult for infeasible to develop conventional i got them to perform the needed task.
audio-chunks\chunk6.wav : Closing letter to computational statistics with focus on making predictions using computers.
audio-chunks\chunk7.wav : But not all machine learning statistical learning.
audio-chunks\chunk8.wav : Study of mathematical optimisation delivers methods theory and application domains to the field of machine learning.
audio-chunks\chunk9.wav : Data mining is related field of study focusing on exploratory data analysis to unsupervised learning.
audio-chunks\chunk10.wav : Application cause business problem machine learning is also is referred as predictive analysis.
audio-chunks\chunk11.wav : Computers discovering how they can perform task without being express entry program to do so.
audio-chunks\chunk12.wav : It was computer learning from data provided so that they can carry out certain task.
audio-chunks\chunk13.wav : Task assigned to computers it is possible to program algorithm check the machine how to execute all sets required to solve the problem at hand.
audio-chunks\chunk14.wav : Computers part no learning is needed.
audio-chunks\chunk15.wav : More advanced as it can be challenging for human to manually create the middle got them in practice it can turn out to be more effective to have the machine developed in algorithm.
audio-chunks\chunk16.wav : Human program must specify every needed step.
audio-chunks\chunk17.wav : Discipline of every machine learning employee various approaches to teach computer to accomplish tasks.
audio-chunks\chunk18.wav : Venu puri satisfactory record them is available.
audio-chunks\chunk19.wav : In cases where was numbers of potential answers exist 1 approaches to label some of the correct answer as valid.
audio-chunks\chunk20.wav : Can be used as training data for computer improves algorithms.
audio-chunks\chunk21.wav : It uses to determine the correct answers.
audio-chunks\chunk22.wav : A system for task of digital cat recognition the admin ist data set of handwritten digits has often been used.

Full text: Machine learning is the study of computer algorithms that improve automatically through experience it is seen as a subset of artificial intelligence. Machine tools builder model based on sample data known as training data. How to make predictions or decisions without being explicitly programmed to do so. Machine codons and used in a wide variety of applications that is email filtering and computer vision. Where it is difficult for infeasible to develop conventional i got them to perform the needed task. Closing letter to computational statistics with focus on making predictions using computers. But not all machine learning statistical learning. Study of mathematical optimisation delivers methods theory and application domains to the field of machine learning. Data mining is related field of study focusing on exploratory data analysis to unsupervised learning. Application cause business problem machine learning is also is referred as predictive analysis. Computers discovering how they can perform task without being express entry program to do so. It was computer learning from data provided so that they can carry out certain task. Task assigned to computers it is possible to program algorithm check the machine how to execute all sets required to solve the problem at hand. Computers part no learning is needed. More advanced as it can be challenging for human to manually create the middle got them in practice it can turn out to be more effective to have the machine developed in algorithm. Human program must specify every needed step. Discipline of every machine learning employee various approaches to teach computer to accomplish tasks. Venu puri satisfactory record them is available. In cases where was numbers of potential answers exist 1 approaches to label some of the correct answer as valid. Can be used as training data for computer improves algorithms. It uses to determine the correct answers. A system for task of digital cat recognition the admin ist data set of handwritten digits has often been used.

```
In [3]: f = open('C:/Users/fanta_000/Desktop/text1.txt','w')
        f.write(text)
        f.close()
```

Full Input Text Stores in a file

```
[4]  import spacy
     from spacy.lang.en.stop_words import STOP_WORDS
     from string import punctuation
```

```
stopwords = list(STOP_WORDS)
stopwords
```

```
'top',
'a',
'thus',
'wherever',
'side',
'make',
'becoming',
'per',
'we',
'always',
```

Text Preprocessing – removing stop words, punctuation...

```
[11] word_freq

    'improves': 1,
    'infeasible': 1,
    'intelligence': 1,
    'ist': 1,
    'known': 1,
    'label': 1,
    'learning': 9,
    'letter': 1,
    'machine': 6,
    'making': 1,
    'manually': 1,
    'mathematical': 1,
    'methods': 1,
    'middle': 1,
    'mining': 1,
    'model': 1,
    'needed': 3,
    'numbers': 1,
    'optimisation': 1,
    'perform': 2,
    'possible': 1,
    'potential': 1,
    'practice': 1,
    'predictions': 2,
    'predictive': 1,
    'problem': 2,
    'program': 3,
    'programmed': 1,
    'provided': 1,
    'puri': 1,
    'recognition': 1,
    'record': 1,
    'referred': 1,
```

Normalization & Sentence Tokenization

```
max_freq=max(word_freq.values())
max_freq
```

```
9
```

```
[14] for word in word_freq.keys():
         word_freq[word]=word_freq[word]/max_freq
```

```
[15] print(word_freq)
```

```
{'Machine': 0.3333333333333333, 'learning': 1.0, 'study': 0.2222222222222222, 'computer': 0.5555555555555556, 'algorithms': 0.2222222222222222, 'improve': 0.
```

```
[16] sentence_tokens=[sent for sent in doc.sents]
     print(sentence_tokens)
```

```
[Machine learning is the study of computer algorithms that improve automatically through experience it is seen as a subset of artificial intelligence., Machi
```

Sentence Token Scoring

```
sentence_score={}
for sent in sentence_tokens:
  for word in sent:
     if word.text.lower() in word_freq.keys():
        if sent not in sentence_score.keys():
           sentence_score[sent]=word_freq[word.text.lower()]
        else:
           sentence_score[sent]+=word_freq[word.text.lower()]
```

```
[18] sentence_score
```

```
{Machine learning is the study of computer algorithms that improve automatically through experience it is seen as a subset of artificial intelligence.: 3.444
Machine tools builder model based on sample data known as training data.: 2.888888888888889,
How to make predictions or decisions without being explicitly programmed to do so.: 0.5555555555555556,
Machine codons and used in a wide variety of applications that is email filtering and computer vision.: 2.0,
Where it is difficult for infeasible to develop conventional i got them to perform the needed task.: 1.6666666666666665,
Closing letter to computational statistics with focus on making predictions using computers.: 1.0,
But not all machine learning statistical learning.: 2.7777777777777777,
Study of mathematical optimisation delivers methods theory and application domains to the field of machine learning.: 2.888888888888889,
Data mining is related field of study focusing on exploratory data analysis to unsupervised learning.: 3.5555555555555556,
Application cause business problem machine learning is also is referred as predictive analysis.: 2.666666666666667,
Computers discovering how they can perform task without being express entry program to do so.: 1.5555555555555556,
It was computer learning from data provided so that they can carry on certain task.: 3.0,
Task assigned to computers it is possible to program algorithm check the machine how to execute all sets required to solve the problem at hand.: 3.0,
Computers part no learning is needed.: 1.5555555555555556,
More advanced as it can be challenging for human to manually create the middle got them in practice it can turn out to be more effective to have the machine
Human program must specify every needed step.: 1.0,
Discipline of every machine learning employee various approaches to teach computer to accomplish tasks.: 2.8888888888888893,
```

```
[19] from heapq import nlargest
```

```
[20] select_length=int(len(sentence_tokens)*0.3)
     select_length
```

```
     6
```

```
[21] summary=nlargest(select_length,sentence_score,key=sentence_score.get)
```

```
[22] summary
```

```
     [Data mining is related field of study focusing on exploratory data analysis to unsupervised learning.,
      Machine learning is the study of computer algorithms that improve automatically through experience it is seen as a subset of artificial intelligence.,
      It was computer learning from data provided so that they can carry on certain task.,
      Task assigned to computers it is possible to program algorithm check the machine how to execute all sets required to solve the problem at hand.,
      Discipline of every machine learning employee various approaches to teach computer to accomplish tasks.,
      Machine tools builder model based on sample data known as training data.]
```

```
[27] len(text)

     2037
```

```
[28] len(summary)

     656
```

Input Text Length
**Vs**
Summary Length

Implementation FrontEnd

Epitome   Home   About   Team   Contact Us   Login   Register

Welcome to Epitome! Let's make it easy...

Epitome is one of those applications of Natural Language Processing (NLP) which is bound to have a huge impact on our lives. With growing digital media and ever growing publishing – who has the time to go through entire articles / documents / books to decide whether they are useful or not?
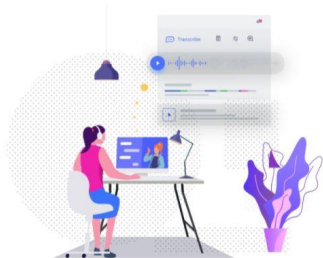
Epitome   Home   About   Team   Contact Us   Login   Register

Contact Us
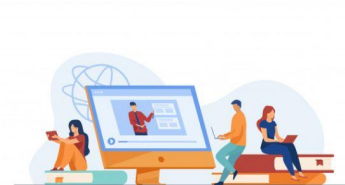### Contact us directly if you have any questions

Please write your name, email address and a message below if you have any questions. One of our staff members will be happy to contact you directly and answer your questions as soon as possible.

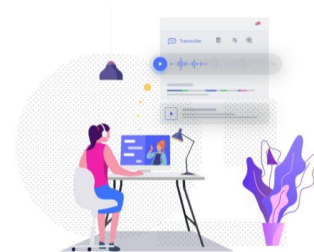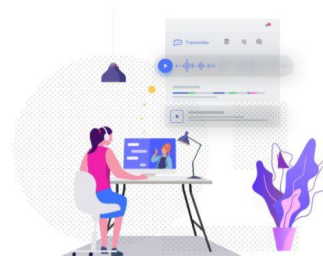First name*

Last name*

Email address*

Message*

Epitome   Home   About   Team   Contact Us   Login   Register

About

The project is to create a product which joins Google, Zoom, and other web conferencing meetings. While one is in the online/offline meeting speaking, Epitome is recording the audio of the conversation. Once the meeting concludes, the Epitome natural language processing (NLP) algorithms transcribe the audio into a copy of the text and skim it to provide the summary (key content) of the meeting rather than full transcription.
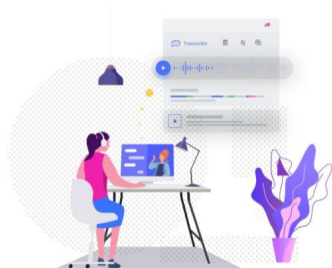

Epitome   Home   About   Team   Contact Us   Login   Register

Log In

Username*

Password*

Login

Need An Account? Sign Up Now

Sarthak May 18, 2021

Update   Audio To Text   Text To Summary   Record Live   Delete

Cloud Computing

Recording File Name: recordings/Cloud_VnrZ5F2.wav

Content: Cloud computing is the on-demand availability of computer system resources. Specific data storage and computing power without direct active management by the user. Generally used to describe data centers available to many users over the internet. Large clouds predominant today often have functions distributed over multiple locations from central service. The connection to the user is relative to the close it may be presented.

Summary: Specific data storage and computing power without direct active management by the user.

Join Today

Username*

Required. 150 characters or fewer. Letters, digits and @/./+/-/_ only.

Email*

Password*

- Your password can't be too similar to your other personal information.
- Your password must contain at least 8 characters.
- Your password can't be a commonly used password.

Summary Post

Title*

Title

Record*

Choose File   No file chosen

Post

Posts

Sarthak May 18, 2021

Cloud Computing

Content: Cloud computing is the on-demand availability of computer system resources. Specific data storage and computing power without direct active management by the user. Generally used to describe data centers available to many users over the internet. Large clouds predominant today often have functions distributed over multiple locations from central service. The connection to the user is relative to the close it may be presented.

Summary: Specific data storage and computing power without direct active management by the user.

Sarthak May 17, 2021

Machine Learning

Content: Machine learning is the study of computer algorithms that improve automatically through experience it is seen as a subset of artificial intelligence. Machine tools builder model based on sample data known as training data. How to make predictions or decisions without being explicitly programmed to do so. Machine codons and used in a wide variety of applications that is email filtering and

Sarthak

sarthak@gmail.com

Profile Info

Username*

Sarthak

Required. 150 characters or fewer. Letters, digits and @/./+/-/_ only.

Email*

sarthak@gmail.com

Image*

Currently: profile_pics/sarthakp_2_1BzXTwy.jpg
Change:

Testing/Test Cases

- ❏ Proper speech file (.wav, .mp4)
- ❏ Different languages are not permissible, only valid for English Language
- ❏ Speech To Text Conversion ⇻ Proper conversion without missing words, and preserving its original meaning
- ❏ Removal of stop words & all the punctuations
- ❏ Word frequence
- ❏ Sentence tokenization, normalization and scoring
- ❏ Sentences score heap
- ❏ Top 30% of sentences

Input Speech

Machine learning is the study of computer algorithms that improve automatically through experience it is seen as a subset of artificial intelligence. Machine tools builder model based on sample data known as training data. How to make predictions or decisions without being explicitly programmed to do so. Machine codons and used in a wide variety of applications that is email filtering and computer vision. Where it is difficult for infeasible to develop conventional i got them to perform the needed task. Closing letter to computational statistics with focus on making predictions using computers. But not all machine learning statistical learning. Study of mathematical optimisation delivers methods theory and application domains to the field of machine learning. Data mining is related field of study focusing on exploratory data analysis to unsupervised learning. Application cause business problem machine learning is also is referred as predictive analysis. Computers discovering how they can perform task without being express entry program to do so. It was computer learning from data provided so that they can carry on certain task. Task assigned to computers it is possible to program algorithm check the machine how to execute all sets required to solve the problem at hand. Computers part no learning is needed. More advanced as it can be challenging for human to manually create the middle got them in practice it can turn out to be more effective to have the machine developed in algorithm. Human program must specify every needed step. Discipline of every machine learning employee various approaches to teach computer to accomplish tasks. Venu puri satisfactory record them is available. In cases where was numbers of potential answers exist 1 approaches to label some of the correct answer as valid. Can be used as training data for computer improves algorithms. It uses to determine the correct answers. A system for task of digital cat recognition the admin ist data set of handwritten digits has often been used.

Output Summary

Data mining is related field of study focusing on exploratory data analysis to unsupervised learning. Machine learning is the study of computer algorithms that improve automatically through experience it is seen as a subset of artificial intelligence. It was computer learning from data provided so that they can carry on certain task. Task assigned to computers it is possible to program algorithm check the machine how to execute all sets required to solve the problem at hand. Discipline of every machine learning employee various approaches to teach computer to accomplish tasks. Machine tools builder model based on sample data known as training data.

Advantages

- ❖ Stakeholders is not being required to create notes during meetings.
- ❖ Participants can focus on the meeting.
- ❖ Reducing human effort to create notes.
- ❖ Capture, Clarify and consolidate key ideas quickly and easily.

Shortcomings

❖ Language Barrier- We have made our speech recognizer for english only.

❖ Usage of only single language in speech i.e only English.

The Outcome Discussion

We were successfully able to summarize and extract top 30% of sentences tokenized from the provided input speech with preserving the original context of the speech. We were able to skim the given data with the help and use of NLP and its respective libraries.

Future Work

❖ From our summarization result we have found that by reducing all sentences that do not contain any geographic information may lead to a loss of information, since there may exist link between that reduced sentences. Therefore, we will analyse this issue in detail, by studying graph based algorithm that capture the relationship between sentences.

❖ Implementation through RNN

❖ Visual Inclusions

❖ Other Research Work.

References

- ❖ Towards Data Science
- ❖ Analytics Vidhya
- ❖ Medium Blogs
- ❖ Geeks for Geeks
- ❖ Machine Learning Mastery
- ❖ Stack Overflows
- ❖ Some Research Papers
- ❖ Other blogging websites

Conclusion

The facts mentioned in the project reveal how Epitome can be used to fill major loopholes in online and offline meetings conducted globally.

Despite the drawbacks, this system can be a major break through in the digital, education and professional sector. Even in the worst conditions, with a few minor modifications, it will enable the optimization to reduce the obstacles that come their way.

In short, Epitome is a step forward towards technology, innovation and hence overall growth.

Thank You