NOTES SUMMARIZATION

Submitted by

S.MANISHMA (113218205018)
P.RASHMI (113218205035)

DEPARTMENT OF INFORMATION TECHNOLOGY

VELAMMAL ENGINEERING COLLEGE, AMBATTUR,

CHENNAI 600-066

MAY - JUNE 2021

VELAMMAL ENGINEERING COLLEGE

DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

It is certified that this mini project report titled **"NOTES SUMMARIZATION"** is the bonafide work of **Ms.Manishma.S (113218205018)** and **Ms.Rashmi.P (113218205035)** who carried out the mini-project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report.

SIGNATURE

_____-

SATHYABAMA A R, M.Tech

ASSISTANT PROFESSOR-I

INTERNAL GUIDE

VELAMMAL ENGINEERING COLLEGE
CHENNAI 600-066

SIGNATURE

_____-

Dr.Jeeva Katiravan, M.Tech.,Ph.D

HEAD OF THE DEPARTMENT

DEPARTMENT OF INFORMATION TECHNOLOGY

VELAMMAL ENGINEERING COLLEGE
CHENNAI 600-066

ACKNOWLEDGEMENT

Behind every achievement lies an unfathomable sea of gratitude to those who achieved it, without whom it would ever have come into existence. To them we express our words of gratitude.

We give all the glory and thanks to our almighty GOD for showering upon, the necessary wisdom and grace for accomplishing this project. We express our gratitude and thanks to our parents first for giving health and sound mind for completing this project.

First of all, we would like to express our deep gratitude to our beloved and respectable Chairman, Thiru M.V Muthuramalingam and Chief Executive Officer, Thiru M.V.M Velmurugan for their kind encouragement.We express our deep gratitude to Dr.N.Duraipandian, principal of Velammal Engineering College for his helpful attitude towards this project.We wish to express our sincere thanks and gratitude to Dr.Jeeva Katiravan Head of the Department of Information Technology for motivating and encouraging every part of our project.

We express our sincere gratitude to The Project Guide Sathyabama A R Assistant Professor-I, Department of Information Technology for her valuable guidance in shaping the project.

Our thanks to all Teaching and Non-teaching staff members of our department for their support and peers for having stood by us and helped us to complete this project.

ABSTRACT

We spent hours of time reading journals and research papers to get a clear idea about a domain. Reading long lessons before exams at constrained time is always an issue for a student. We always prefer the content to be small and informative. Summarising gives you a brief view of a particular topic in a crisp sense. Taking this into consideration we propose our Idea "Notes Summarisation". An effective NLP and OCR Models are used to gain the objective. This intakes long paras of text or images containing text and summarises it without missing important points. This ensures a reduction of time as well as effective learning.

.

INDEX

1. INTRODUCTION

1.1 Artificial Intelligence

Artificial intelligence (AI) is wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. AI is an interdisciplinary science with multiple approaches, but advancements in machine learning and deep learning are creating a paradigm shift in virtually every sector of the tech industry.

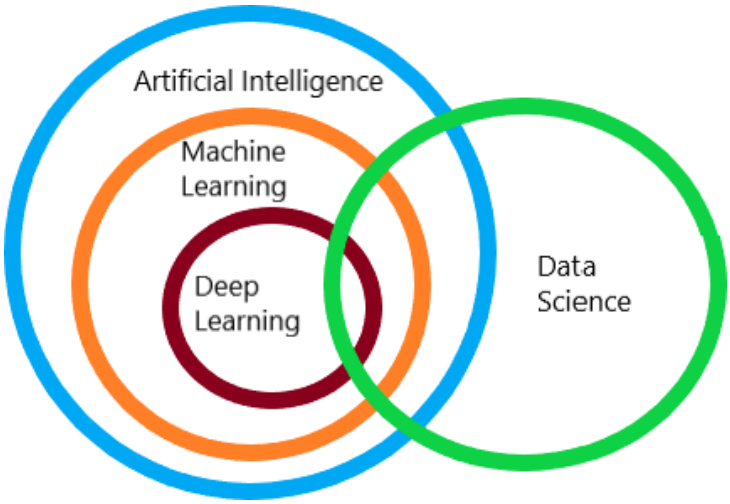
Artificial intelligence can be divided into two different categories: weak and strong.

Weak artificial intelligence embodies a system designed to carry out one particular job. Weak AI systems include video games such as the chess example from above and personal assistants such as Amazon's Alexa and Apple's Siri. You ask the assistant a question, it answers it for you.

Strong artificial intelligence systems are systems that carry on the tasks considered to be human-like. These tend to be more complex and complicated systems. They are programmed to handle situations in which they may be required to problem solve without having a person intervene. These kinds of systems can be found in applications like self-driving cars or in hospital operating rooms.

Although artificial intelligence evokes thoughts of science fiction, artificial intelligence already has many uses today, for example:

- **Email filtering:** Email services use artificial intelligence to filter incoming emails. Users can train their spam filters by marking emails as "spam".
- **Personalization:** Online services use artificial intelligence to personalize your experience. Services, like Amazon or Netflix, "learn" from your previous purchases and the purchases of other users in order to recommend relevant content for you.
- **Fraud detection:** Banks use artificial intelligence to determine if there is strange activity on your account. Unexpected activity, such as foreign transactions, could be flagged by the algorithm.
- **Speech recognition:** Applications use artificial intelligence to optimize speech recognition functions. Examples include intelligent personal assistants, e.g. Amazon's "Alexa" or Apple's "Siri".



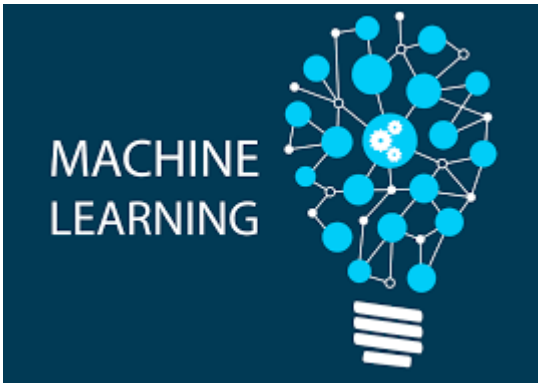
1.2 Machine Learning

Algorithms are a sequence of instructions used to solve a problem. Algorithms, developed by programmers to instruct computers in new tasks, are the building blocks of the advanced digital world we see today. Computer algorithms organize enormous amounts of data into information and services, based on certain instructions and rules.

Instead of programming the computer every step of the way, this approach gives the computer instructions that allow it to learn from data without new step-by-step instructions by the programmer. This means computers can be used for new, complicated tasks that could not be manually programmed. Things like

photo recognition applications for the visually impaired , or translating pictures into speech.

The basic process of machine learning is to give training data to a learning algorithm. The learning algorithm then generates a new set of rules, based on inferences from the data. This is in essence generating a new algorithm, formally referred to as the machine learning model. By using different training data, the same learning algorithm could be used to generate different models. For example, the same type of learning algorithm could be used to teach the computer how to translate languages or predict the stock market.



Inferring new instructions from data is the core strength of machine learning. It also highlights the critical role of data: the more data available to train the algorithm, the more it learns. In fact, many recent advances in AI have not been due to radical innovations in learning algorithms, but rather by the enormous amount of data enabled by the Internet.

1.3 How machines learn :

Although a machine learning model may apply a mix of different techniques, the methods for learning can typically be categorized as three general types:

- Supervised learning: The learning algorithm is given labeled data and the desired output. For example, pictures of dogs labeled “dog” will help the algorithm identify the rules to classify pictures of dogs.
- Unsupervised learning: The data given to the learning algorithm is un labeled, and the algorithm is asked to identify patterns in the input data. For example, the recommendation system of an e-commerce website where the learning algorithm discovers similar items often bought together.
- Reinforcement learning: The algorithm interacts with a dynamic environment that provides feedback in terms of rewards and punishments. For example, self-driving cars being rewarded to stay on the road.

1.4 Growth of AI and ML :

Machine learning is not new. Many of the learning algorithms that spurred new interest in the field, such as neural networks, are based on decades old research. The current growth in AI and machine learning is tied to developments in three important areas:

- **Data availability**: Just over 3 billion people are online with an estimated 17 billion connected devices or sensors. That generates a large amount of data which, combined with decreasing costs of data storage, is easily available for use. Machine learning can use this as training data for learning algorithms, developing new rules to perform increasingly complex tasks.
- **Computing power**: Powerful computers and the ability to connect remote processing power through the Internet make it possible for machine-learning techniques that process enormous amounts of data.
- **Algorithmic innovation**: New machine learning techniques, specifically in layered neural networks – also known as "deep learning" – have inspired new services, but is also spurring investments and research in other parts of the field.

1.5 Deep Learning

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It's achieving results that were not possible before.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.

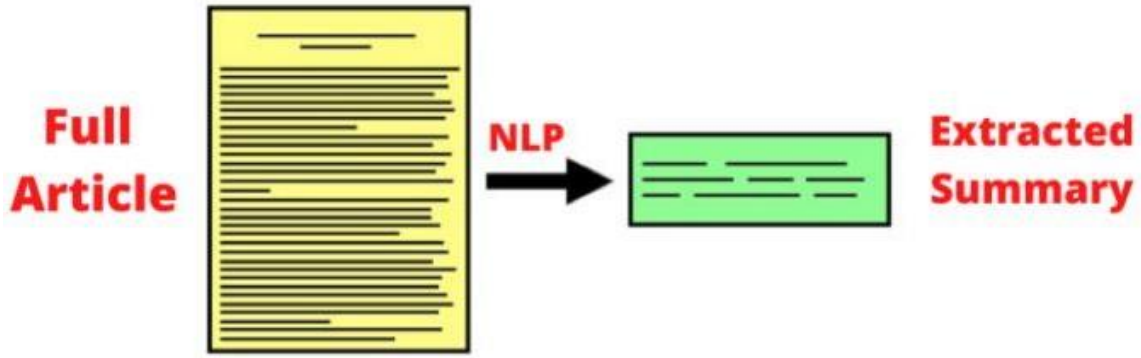
Deep learning achieves recognition accuracy at higher levels than ever before. This helps consumer electronics meet user expectations, and it is crucial for safety-critical applications like driverless cars. Recent advances in deep learning have improved to the point where deep learning outperforms humans in some tasks like classifying objects in images.

While deep learning was first theorized in the 1980s, there are two main reasons it has only recently become useful:

1. Deep learning requires large amounts of **labeled data**. For example, driverless car development requires millions of images and thousands of hours of video.
2. Deep learning requires substantial **computing power**. High-performance GPUs have a parallel architecture that is efficient for deep learning. When combined with clusters or cloud computing, this enables development teams to reduce training time for a deep learning network from weeks to hours or less.

1.6 Summarization

Summarization can be defined as a task of producing a concise and fluent summary while preserving key information and overall meaning.



Convert Article to Summary

Impact

Summarization systems often have additional evidence they can utilize in order to specify the most important topics of document(s). For example, when summarizing blogs, there are discussions or comments coming after the blog post that are good sources of information to determine which parts of the blog are critical and interesting.

In scientific paper summarization, there is a considerable amount of information such as cited papers and conference information which can be leveraged to identify important sentences in the original paper.

How text summarization works

In general there are two types of summarization, **abstractive** and **extractive** summarization.

1. **Abstractive Summarization:** Abstractive methods select words based on semantic understanding, even those words did not appear in the source documents. It aims at producing important material in a new way. They interpret and examine the text using advanced natural language techniques in order to generate a new shorter text that conveys the most critical information from the original text.

It can be correlated to the way human reads a text article or blog post and then summarizes in their own word.

Input document → understand context → semantics → create own summary.

2. **Extractive Summarization:** Extractive methods attempt to summarize articles by selecting a subset of words that retain the most important points.

This approach weights the important part of sentences and uses the same to form the summary. Different algorithm and techniques are used to define weights for the sentences and further rank them based on importance and similarity among each other.

Input document → sentences similarity → weight sentences → select sentences with higher rank.

The limited study is available for abstractive summarization as it requires a deeper understanding of the text as compared to the extractive approach.

Purely extractive summaries often times give better results compared to automatic abstractive summaries. This is because of the fact that abstractive summarization methods cope with problems such as semantic representation, inference and natural language generation which is relatively harder than data-driven approaches such as sentence extraction.

What is TextRank algorithm?

TextRank is an extractive summarization technique. It is based on the concept that words which occur more frequently are significant. Hence, the sentences containing highly frequent words are important.

Based on this, the algorithm assigns scores to each sentence in the text. The top-ranked sentences make it to the summary.

How does LexRank work?

A sentence which is similar to many other sentences of the text has a high probability of being important. The approach of LexRank is that a particular sentence is recommended by other similar sentences and hence is ranked higher. Higher the rank, higher is the priority of being included in the summarized text.

LSA (Latent semantic analysis)

Latent Semantic Analysis is an unsupervised learning algorithm that can be used for extractive text summarization. It extracts semantically significant sentences by applying singular value decomposition (SVD) to the matrix of term-document frequency.

1.7 OCR

OCR, or Optical Character Recognition, is a process of recognizing text inside images and converting it into an electronic form. These images could be of handwritten text, printed text like documents, receipts, name cards, etc., or even a natural scene photograph. OCR has two parts to it. The first part is **text detection** where the textual part within the image is determined. This localization of text within the image is important for the second part of OCR, **text recognition**, where the text is extracted from the image. Using these techniques together is how you can extract text from any image.

Text Recognition with Tesseract OCR

Tesseract is an open-source OCR engine originally developed as proprietary software by HP (Hewlett-Packard) but was later made open source in 2005. Google has since then adopted the project and sponsored its development.

2. SOTWARE INVOLVED

1. HARDWARE REQUIREMNTS

Processor	:	Intel Core-i3
Speed	:	1.7 GHz
RAM	:	256 MB(min)
Hard Disk	:	500 GB
Keyboard	:	Standard Windows Keyboard
Mouse	:	Two or three button Mouse
Monitor	:	SVGA

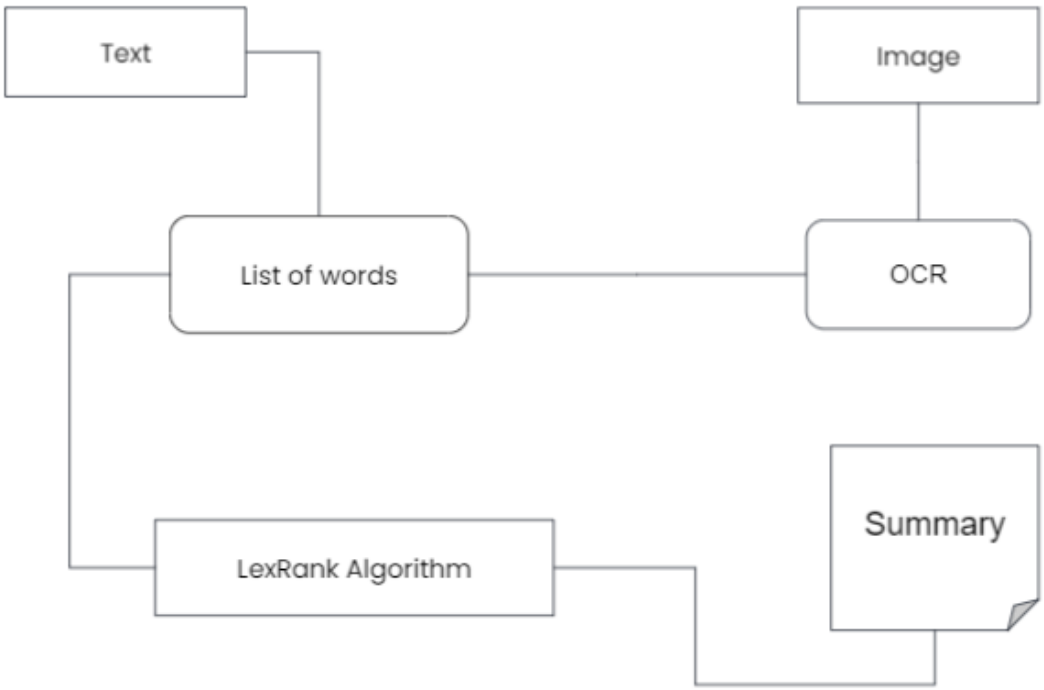
2. SOFTWARE REQUIREMNTS

Platform : Google colab or Python IDLE

Libraries :

1. sumy
2. opencv-Python
3. pytesseract
4. tesseract-ocr
5. nltk

3. BLOCK DIAGRAM



4. SOURCE

```
#Installing sumy package
!pip install sumy

#Installing Opencv
!pip install opencv-python
#Installing Pytesseract
!pip install pytesseract
#Installing tesseract
!sudo apt install tesseract-ocr

#Installing nltk package
!pip install nltk


#Import the Reqquired Libraries
import  sumy  from  sumy.parsers.plaintext
import PlaintextParser from sumy.nlp.tokenizers
import Tokenizer from sumy.summarizers.lex_rank
import LexRankSummarizer import nltk,re,pprint
from nltk import word_tokenize
from urllib import request import cv2
import pytesseract from google.colab.patches
import cv2_imshow
from PIL import Image

#Download punkt for Tokenization
nltk.download('punkt')


print("Choices:")
print("1. Textual Input")
print("2. Image Input")
ch=int(input("Enter the choice number : "))

if(ch==1):

  #Getting Textual Input from the user
  userInput = input("Enter the Text :\n").splitlines()
  print(userInput)
  data=''.join(userInput)
  #Initialsing the parser
  parser = PlaintextParser.from_string(data,Tokenizer('english'))

elif(ch==2):

  #Getting Image input from user
  from google.colab import files
  #Getting the name of image file
```

```
  image = list(files.upload().keys())[0]
  image = cv2.imread(image)
  print(" Input Image : ")
  cv2_imshow(image)
  #Converting into Grayscale
  gray_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
  print("Gray Scale Image : ")
  cv2_imshow(gray_image)
  #Threshold image
  threshold_img = cv2.threshold(gray_image,0,255, cv2.THRESH_BINARY | c
v2.THRESH_OTSU)[1]
  print("Threshold Image : ")
  cv2_imshow(threshold_img)
  pytesseract.pytesseract.tesseract_cmd = (r'/usr/bin/tesseract')
  #Text extraction from image using pytesseract
  data = pytesseract.image_to_string(image, lang='eng', config='-psm 1'
)
  #Initialising the parser
  parser = PlaintextParser.from_string(data,Tokenizer('english'))

else:

  print("Invalid Data!!")
```

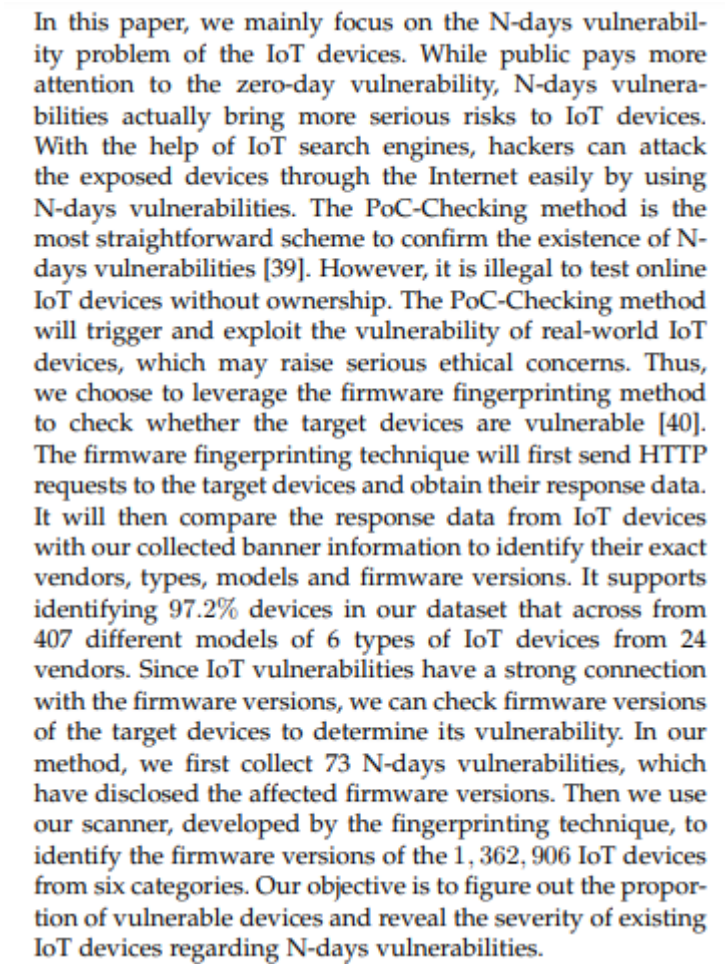
```
#Getting summary using Lex Rank
lex_rank_summarizer = LexRankSummarizer()
summary = lex rank summarizer(parser.document,sentences count=10)

#Printing the Sentence in the summary

for sentence in summary: print(sentence)
```

5. SCREENSHOTS

5. 1 INPUT : IMAGE



In this paper, we mainly focus on the N-days vulnerability problem of the IoT devices. While public pays more attention to the zero-day vulnerability, N-days vulnerabilities actually bring more serious risks to IoT devices. With the help of IoT search engines, hackers can attack the exposed devices through the Internet easily by using N-days vulnerabilities. The PoC-Checking method is the most straightforward scheme to confirm the existence of N-days vulnerabilities [39]. However, it is illegal to test online IoT devices without ownership. The PoC-Checking method will trigger and exploit the vulnerability of real-world IoT devices, which may raise serious ethical concerns. Thus, we choose to leverage the firmware fingerprinting method to check whether the target devices are vulnerable [40]. The firmware fingerprinting technique will first send HTTP requests to the target devices and obtain their response data. It will then compare the response data from IoT devices with our collected banner information to identify their exact vendors, types, models and firmware versions. It supports identifying 97.2% devices in our dataset that across from 407 different models of 6 types of IoT devices from 24 vendors. Since IoT vulnerabilities have a strong connection with the firmware versions, we can check firmware versions of the target devices to determine its vulnerability. In our method, we first collect 73 N-days vulnerabilities, which have disclosed the affected firmware versions. Then we use our scanner, developed by the fingerprinting technique, to identify the firmware versions of the 1,362,906 IoT devices from six categories. Our objective is to figure out the proportion of vulnerable devices and reveal the severity of existing IoT devices regarding N-days vulnerabilities.

OUTPUT

```
In this paper, we mainly focus on the N-days vulnerabil- ity problem of the loT devices.
The PoC-Checking method is the ‘most straightforward scheme to confirm the existence of N- days vulnerabilities [39].
However, it is illegal to test online oT devices without ownership.
The PoC-Checking method will trigger and exploit the vulnerability of real-world IoT devices, which may raise serious ethical concems.
Thus, ‘we choose to leverage the firmware fingerprinting method to check whether the target devices are vulnerable [40] The firmware fing
erprinting technique will first send HTTP requests to the target devices and obtain their response data.
It will then compare the response data from IoT devices with our collected banner information to identify their exact vendors, types, mod
els and firmware versions.
It supports, identifying 97.2% devices in our dataset that across from 407 different models of 6 types of IoT devices from 24 vendors.
Since loT vulnerabilities have a strong connection with the firmware versions, we can check firmware versions of the target devices to de
termine its vulnerability.
In our method, we first collect 73 N-days vulnerabilities, which have disclosed the affected firmware versions.
Our objective is to figure out the propor- tion of vulnerable devices and reveal the severity of existing IoT devices regarding N-days vu
lnerabilities.
```

5.2 INPUT : TEXT

In an attempt to build an AI-ready workforce, Microsoft announced Intelligent Cloud Hub which has been launched to empower the next generation of students with AI-ready skills. Envisioned as a three-year collaborative program, Intelligent Cloud Hub will support around 100 institutions with AI infrastructure, course content and curriculum, developer support, development tools and give students access to cloud and AI services. As part of the program, the Redmond giant which wants to expand its reach and is planning to build a strong developer ecosystem in India with the program will set up the core AI infrastructure and IoT Hub for the selected campuses. The company will provide AI development tools and Azure AI services such as Microsoft Cognitive Services, Bot Services and Azure Machine Learning.According to Manish Prakash, Country General Manager-PS, Health and Education, Microsoft India, said, "With AI being the defining technology of our time, it is transforming lives and industry and the jobs of tomorrow will require a different skillset. This will require more collaborations and training and working with AI. That’s why it has become more critical than ever for educational institutions to integrate new cloud and AI technologies. The program is an attempt to ramp up the institutional set-up and build capabilities among the educators to educate the workforce of tomorrow." The program aims to build up the cognitive skills and in-depth understanding of developing intelligent cloud connected solutions for applications across industry. Earlier in April this year, the company announced Microsoft Professional Program In AI as a learning track open to the public. The program was developed to provide job ready skills to programmers who wanted to hone their skills in AI and data science with a series of online courses which featured hands-on labs and expert instructors as well. This program also included developer-focused AI school that provided a bunch of assets to help build AI skills.

OUTPUT: (If sentence count ==5)

In an attempt to build an AI-ready workforce, Microsoft announced Intelligent Cloud Hub which has been launched to empower the next generation of students with AI-ready skills.

Envisioned as a three-year collaborative program, Intelligent Cloud Hub will support around 100 institutions with AI infrastructure, course content and curriculum, developer support, development tools and give students access to cloud and AI services.

This will require more collaborations and training and working with AI.

That’s why it has become more critical than ever for educational institutions to integrate new cloud and AI technologies.

The program aims to build up the cognitive skills and in-depth understanding of developing intelligent cloud connected solutions for applications across industry.

6. CONCLUSION

Notes Summarization is done using Sumy and OCR libraries. Sumy is used to summarize the given text. OCR is used to extract the text from an image.

Achieved,

- Made learning easier.
- Got a summary of long time taking paragraphs.
- Effective Model constructed
- Reduced time.

7. REFERENCES

1. **https://www.machinelearningplus.com/nlp/text-summarization-approaches-nlp-example/**
2. **https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70**
3. **https://www.analyticsvidhya.com/blog/2020/05/build-your-own-ocr-google-tesseract-opencv/**
4. **https://www.kaggle.com/imkrkannan/text-summarization-with-nltk-in-python**
5. **https://www.geeksforgeeks.org/python-text-summarizer/**
6. **https://www.geeksforgeeks.org/text-detection-and-extraction-using-opencv-and-ocr/#:~:text=Python%2Dtesseract%20is%20a%20wrapper,the%20imread%20function%20of%20opencv.**
7. **Stack Overflow**