SUMMER INSTITUTIONAL TRAINING REPORT

ON

[Text Summarization (Basics of Machine Learning & Deep Learning using Tensorflow)]

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD

OF THE DEGREE OF

BACHELOR OF ENGINEERING

(Computer Science & Engineering)

JUNE-JULY,2022

SUBMITTED BY:

NAME: PARTH AWASTHI

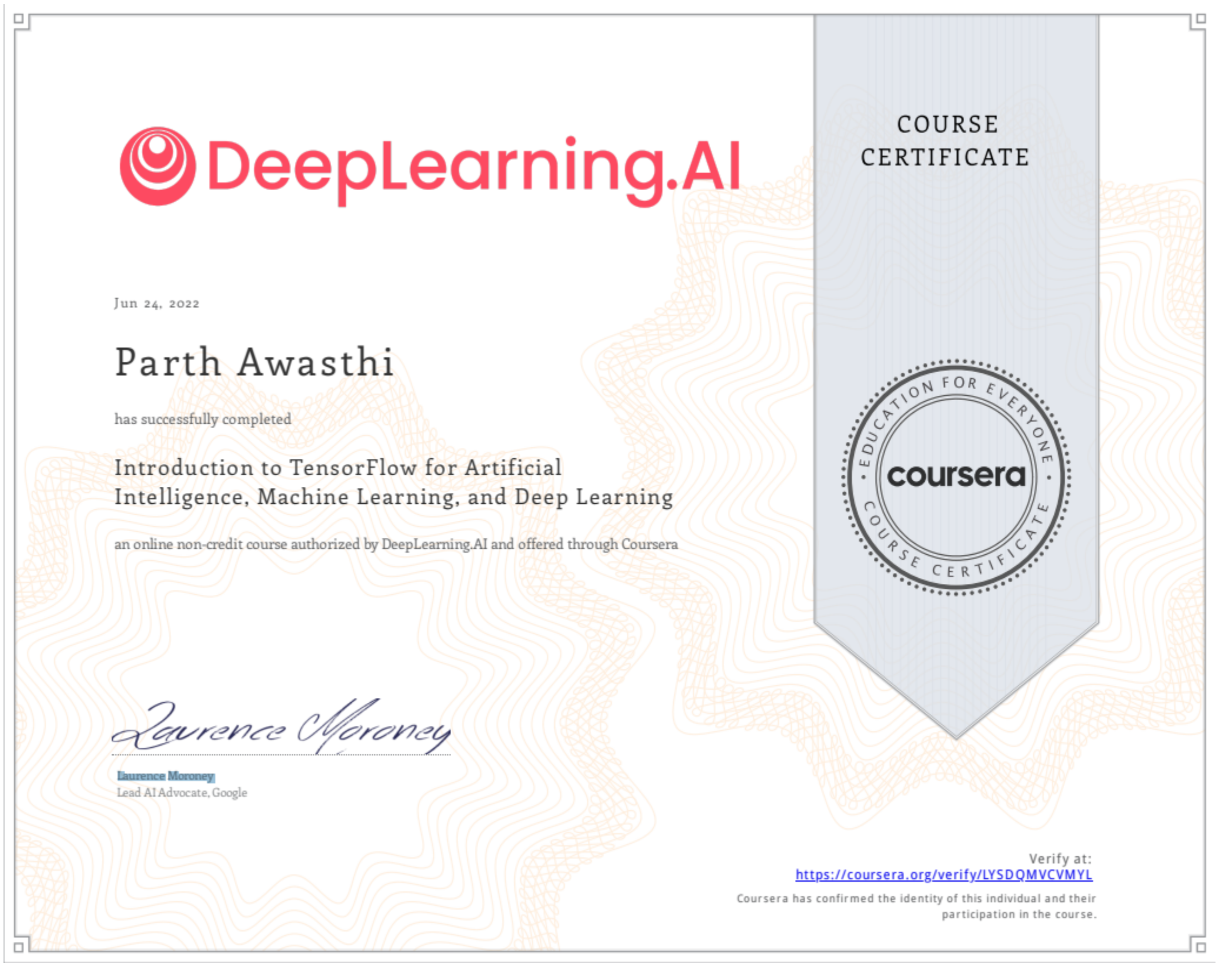
UNIVERSITY UID: 21BCS1447

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CHANDIGARH UNIVERSITY GHARUAN, MOHALI

CERTIFICATES BY COURSERA

1. Introduction to TensorFlow for Artificial



DeepLearning.AI

COURSE CERTIFICATE

Jun 24, 2022

Parth Awasthi

has successfully completed

Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning

an online non-credit course authorized by DeepLearning.AI and offered through Coursera

EDUCATION FOR EVERYONE · coursera · COURSE CERTIFICATE

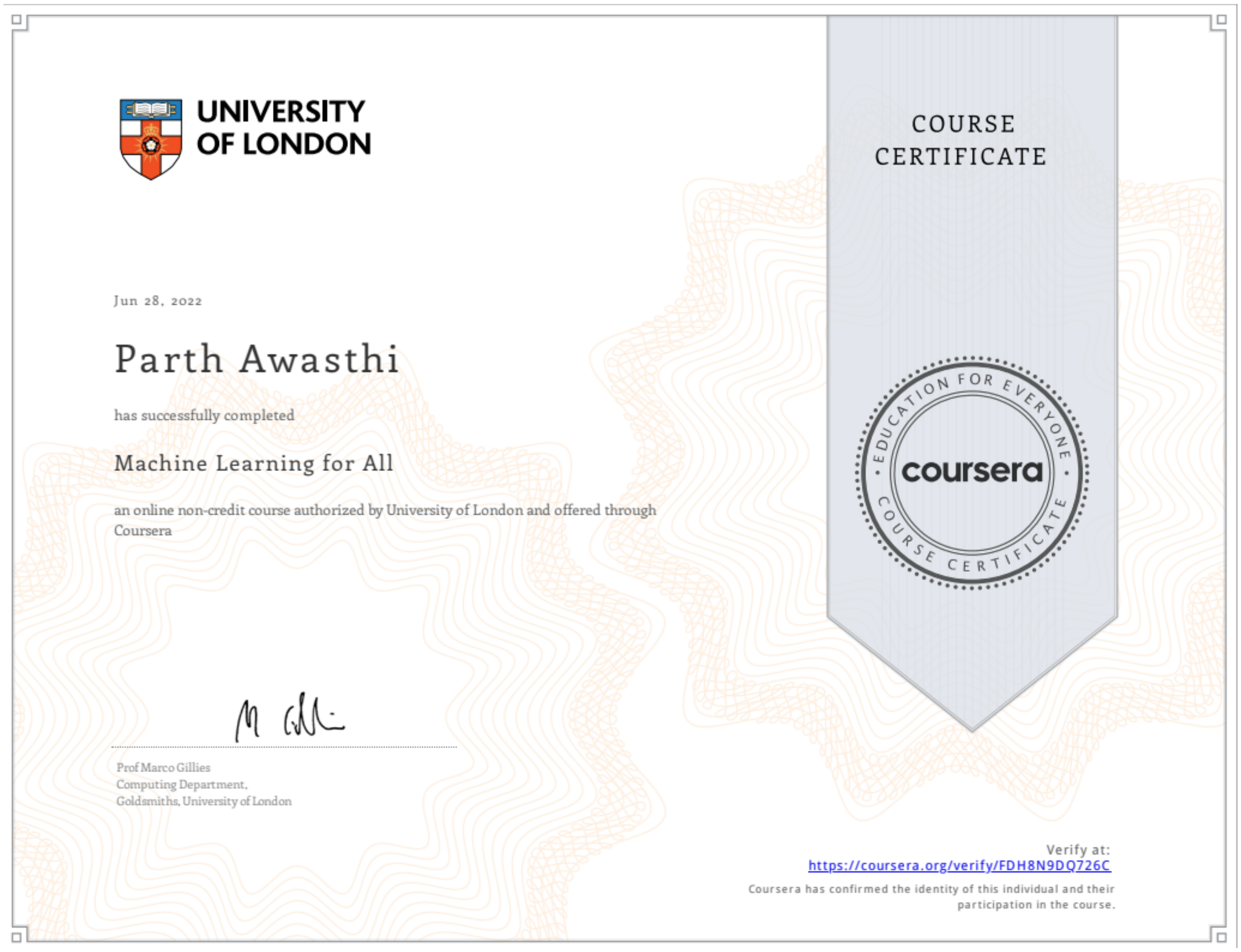
Laurence Moroney

Laurence Moroney
Lead AI Advocate, Google

Verify at:
https://coursera.org/verify/LYSDQMVCVMYL
Coursera has confirmed the identity of this individual and their participation in the course.

Verification link- https://coursera.org/verify/LYSDQMVCVMYL

2. Machine Learning For All



UNIVERSITY OF LONDON

COURSE CERTIFICATE

Jun 28, 2022

Parth Awasthi

has successfully completed

Machine Learning for All

an online non-credit course authorized by University of London and offered through Coursera

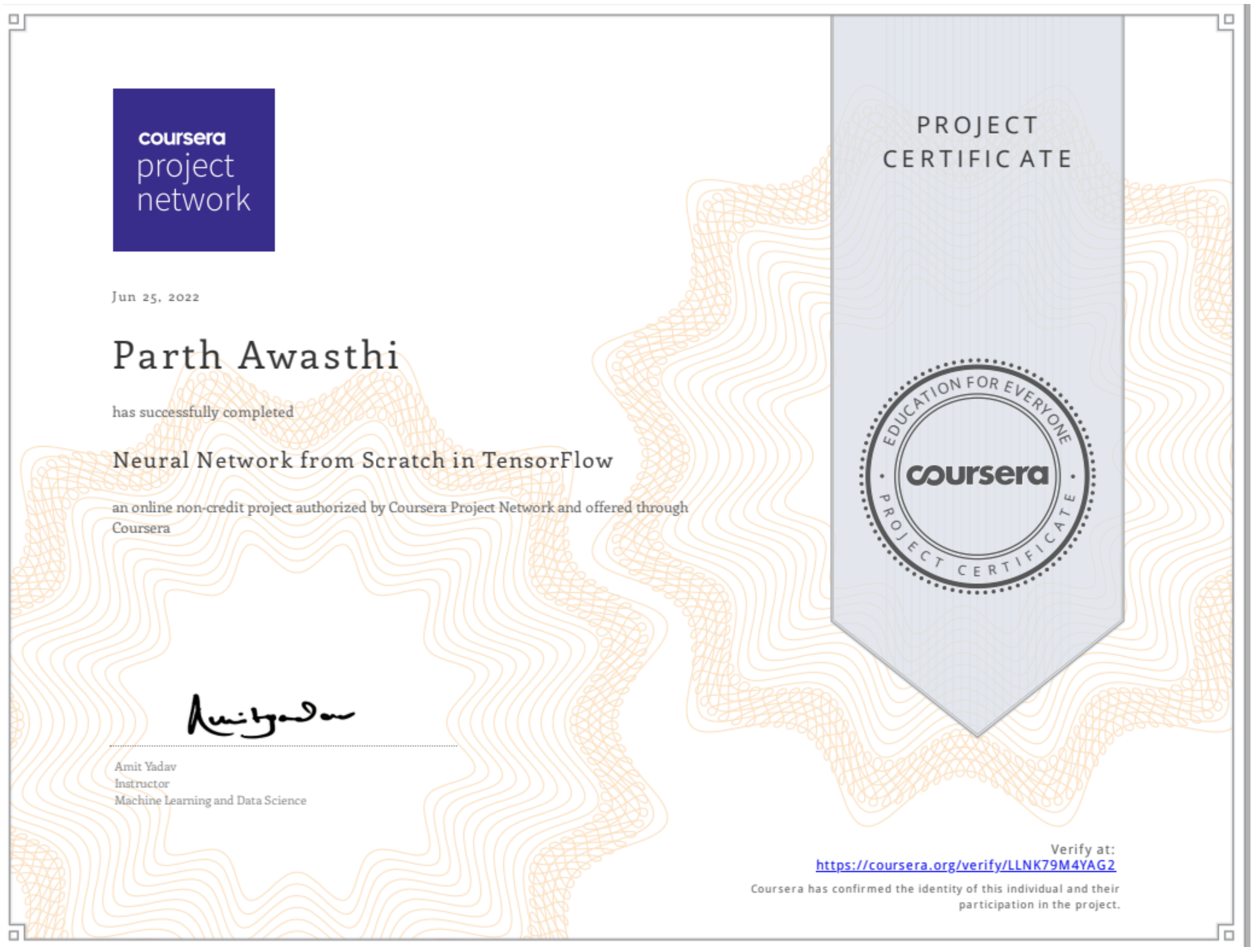
EDUCATION FOR EVERYONE
coursera
COURSE CERTIFICATE

Prof Marco Gillies
Computing Department,
Goldsmiths, University of London

Verify at:
https://coursera.org/verify/FDH8N9DQ726C
Coursera has confirmed the identity of this individual and their participation in the course.

Verification Link- https://coursera.org/verify/FDH8N9DQ726C

3. Neural Network from Scratch in TensorFlow (Guided Project)



coursera
project
network

PROJECT CERTIFICATE

Jun 25, 2022

Parth Awasthi

has successfully completed

Neural Network from Scratch in TensorFlow

an online non-credit project authorized by Coursera Project Network and offered through Coursera

EDUCATION FOR EVERYONE
coursera
PROJECT CERTIFICATE

Amit Yadav
Instructor
Machine Learning and Data Science

Verify at:
https://coursera.org/verify/LLNK79M4YAG2
Coursera has confirmed the identity of this individual and their participation in the project.

Verification Link- https://coursera.org/verify/LLNK79M4YAG2

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI

CANDIDATE'S DECLARATION

I “**Parth Awasthi** ” hereby declare that I have undertaken Summer Training and developed a project titled **Text Summarization_using machine learning_** during a period from 15th JUNE’22 to 02nd JULY’22 in partial fulfillment of requirements for the award of the degree of B.E (COMPUTER SCIENCE & ENGINEERING) at CHANDIGARH UNIVERSITY GHARUAN, MOHALI. The work which is being presented in the training report submitted to the Department of Computer Science & Engineering at CHANDIGARH UNIVERSITY GHARUAN, MOHALI is an authentic record of training work.

Signature of the Student

The training Viva–Voce Examination of__________________ has been held on ____________ and accepted.

Signature of Internal Examiner Signature of External Examiner

ABSTRACT

The Internet is an excellent source of information, where you can get information on all the topics. But due to the large availability of content, it becomes a challenging job to extract exact information. A text summarization system's main objective is to define and present the most relevant information from the given text to the end-users. Nowadays, the data is available in a considerable quantity. It becomes difficult for the user to deal with exact information. It's not possible to read all the information and make a conclusion from specific data. With text summarization, a significant content of data is converted into a short set of information. If we talk specifically about Wikipedia's information, almost all the areas are open on this website. If we search for a specific keyword, it will provide huge details. Text summarization will convey the same extensive information by converting it into small pieces without losing its message. We have taken data from Wikipedia and applied summarization techniques to reduce the content without changing its meaning for demonstrating the concept during this research. Abstractive methods generate an internal semantic representation of the original content. For several years, text summarization has been an active area of study. By summarizing parts of the source document, abstraction can transform the removed content to condense more strongly than extraction.

ACKNOWLEDGMENT

Gratitude cannot be seen or expressed. It can only be felt in the heart and is beyond description. Often words are inadequate to serve as a model of the expression of one’s feelings, especially the sense of indebtedness and gratitude to all those who helped us do our duty.

I would like to express my deep and sincere gratitude to my course instructor, Laurence Moroney, Amit Yadav, Prof MarcoGillies the Batch Head (CSE) Prof (Dr.) Puneet Kumar and my University for giving me the golden opportunity to do this Summer/Institutional Training in the form of an online course and for providing invaluable guidance throughout the course on the topic “Machine Learning for all ” and the guided project on the topic” Neural Network from Scratch in TensorFlow”. It was a privilege and honor to work and study under such instructors. I am extremely grateful for what they have offered to me. It helped me in doing a lot of research and I came to know I learned a lot of things related to this topic.

Finally, I would also like to thank my parents and friends who helped me a lot in finalizing this project within time.

ABOUT THE COURSE

Machine Learning, often called Artificial Intelligence or AI is one of the most exciting areas of technology at the moment. We see daily news stories that herald new breakthroughs in facial recognition technology, self-driving cars, or computers that can have a conversation just like a real person. Machine Learning technology is set to revolutionize almost any area of human life and work, and so will affect all our lives, so you are likely to want to find out more about it. Machine Learning has a reputation for being one of the most complex areas of computer science, requiring advanced mathematics and engineering skills to understand it. While it is true that working as a Machine Learning engineer does involve a lot of mathematics and programming, we believe that anyone can understand the basic concepts of Machine Learning, and given the importance of this technology, everyone should. The big AI breakthroughs sound like science fiction, but they come down to a simple idea: the use of data to train statistical algorithms. In this course, you will learn to understand the basic idea of machine learning, even if you don't have any background in math or programming.

LIST OF FIGURES

LIST OF TABLES

DEFINITION, ACRONYMS, AND ABBREVIATIONS

Standard Abbreviation and Definitions used:

(i) NLTK: Natural Language Toolkit.

(ii) Punkt: Pre-trained Punkt tokenizer

(iii) Matplotlib: cross-platform, data visualization and graphical plotting library

(iv) Word Cloud: represents text data in which the size of each word indicates its frequency or importance

(v) Rake: Rapid Automatic Keyword Extraction algorithm

CONTENTS

Topic **Page No.**

CHAPTER 1 INTRODUCTION

CHAPTER 2 TRAINING WORK UNDERTAKEN

CHAPTER 3 RESULTS AND DISCUSSION

CHAPTER 4 CONCLUSION AND FUTURE SCOPE

CHAPTER 1

1.0 INTRODUCTION

Python is a programming language that is preferred for programming due to its vast features, applicability, and simplicity. The Python programming language best fits machine learning due to its independent platform and its popularity in the programming community.

Machine learning is a section of Artificial Intelligence (AI) that aims at making a machine learn from experience and automatically do the work without necessarily being programmed on a task. On the other hand, Artificial Intelligence (AI) is the broader meaning of machine learning, where computers are made to be receptive to the human level by recognizing visually, by speech, language translation, and consequently making critical decisions.

1.1 BACKGROUND INFORMATION

1.1.1 PYTHON

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. It's high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy-to-learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed.

1.1.2 MACHINE LEARNING

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google, and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

1.1.3 GOOGLE COLLAB

Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members - just the way you edit documents in Google Docs. Colab supports many popular machine learning libraries which can be easily loaded into your notebook.

1.2 FEATURES OF PYTHON

1. **Free and Open-Source:** Python is developed under an OSI-approved open source license. Hence, it is completely free to use, even for commercial purposes. It doesn't cost anything to download Python or to include it in your application. It can also be freely modified and re-distributed. Python can be downloaded from the official Python website.
2. **Robust Standard Library:** Python has an extensive standard library available for anyone to use. This means that programmers don’t have to write their code for every single thing, unlike other programming languages. There are libraries for image manipulation, databases, unit-testing, expressions, and a lot of other functionalities. In addition to the standard library, there is also a growing collection of thousands of components, which are all available in the Python Package Index.

3. **Interpreted:** When a programming language is interpreted, it means that the source code is executed line by line, and not all at once. Programming languages such as C++ or Java are not interpreted, and hence need to be compiled first to run them. There is no need to compile Python because it is processed at runtime by the interpreter.
4. **Portable:** Python is portable in the sense that the same code can be used on different machines. Suppose you write Python code on a Mac. If you want to run it on Windows or Linux later, you don't have to make any changes to it. As such, there is no need to write a program multiple times for several platforms.
5. **Object-Oriented and Procedure**: Oriented A programming language is object-oriented if it focuses on design around data and objects, rather than functions and logic. On the contrary, a programming language is procedure-oriented if it focuses more on functions (code that can be reused). One of the critical Python features is that it supports both object-oriented and procedure-oriented programming.

1.1.3 FEATURES OF MACHINE LEARNING

1. **Preprocessing of Data:** A data mining technique that involves transforming raw data into an understandable format.
2. **Feature Engineering:** Feature engineering is the process of altering the data to help machine learning algorithms work better, which is often time-consuming and expensive
3. **Diverse Algorithms:** Every dataset contains unique information that reflects the individual events and characteristics of a business. Due to the variety of situations and conditions, one algorithm cannot successfully solve every possible business problem or dataset. Because of this, we need access to a diverse repository of algorithms to test against our data, in order to find the best one for our particular data.
4. **Algorithm Selection:** Look for an automated machine learning platform that knows which algorithms make sense for your data and runs only those. That way you will get better algorithms, faster.

1.1.4 FEATURES OF GOOGLE COLAB

1. Interactive tutorials to learn machine learning and neural networks.
2. Write and execute Python 3 code without having a local setup.
3. Execute terminal commands from the Notebook.
4. Import datasets from external sources such as Kaggle.
5. Save your Notebooks to Google Drive.
6. Import Notebooks from Google Drive.
7. Free cloud service, GPUs, and TPUs.
8. Integrate with PyTorch, Tensor Flow, and Open CV.
9. Import or publish directly from/to GitHub.

CHAPTER 2

2.0 TRAINING WORK UNDERTAKEN

In this Summer/Institutional Training, I have learned different uses and features of machine learning and python programming language. through online learning platform Coursera. Also, I made a mini project using all the

concepts of Machine Learning. I have made a "Text Summarizer using ML" which aims to convert big paragraphs into a summarized text using machine learning.

2.1 LEARNING PROCEDURE

1.0 WEEK-1-

1. describe and illustrate different artificial bits of intelligence and machine learning problems and techniques
2. explain the process of training and using a machine learning model
3. train and debug a machine learning system based on a notional machine model of that system

.

2.0 WEEK-2

1. describe and illustrate the role of data features in machine learning

2. Describe a number of data features used in machine learning
3. reason about the effect of data features used on the function of machine learning systems.

3.0 WEEK-3

1. Explain a number of typical errors and problems that can occur in machine learning
2. Test a machine learning model in order to evaluate its effectiveness
3. Critique uses of machine learning in terms of their benefits and dangers

4.0 WEEK-4

1. describe and illustrate how data is collected, analyzed and used
2. collect a dataset suitable for use in training
3. reason about the effect of data sets on the function of machine learning systems.

2.2 PROJECT UNDERTAKEN

2.2.1 OBJECTIVE OF THE PROJECT

The objective of this project is to create a Text Summarization_using machine learning.

2.2.2 SCOPE OF THE PROJECT

Now that information in digital spaces is largely non-structured, requiring tools that allow people to extract key points easily is more crucial than ever. Text segmentation is the challenge of condensing a long document into a concise, accurate, fluent summary. We enjoy tremendous quantities of data at our fingertips. But the majority of this information is unnecessary, nonessential, and will only serve to confuse. When summarising a text, extract the important ideas from it. With the exception of having the summarization turned on, however, however, that is not the case. The software can now reduce a text of 5,000 words to an engaging sound bite in less than five seconds. This is true regardless of the language. It is also possible that you must condense a text that you can't understand into a few simple words. Text extraction is entirely language-independent. Since most people's ability is limited to speaking only one language, it exceeds that of human beings. The summarization software covers this gap and makes the task simple for those who don't need to have to work hard. Both text-to-text and text-to-English conversion are in a single program.

2.3 PROJECT REQUIREMENTS

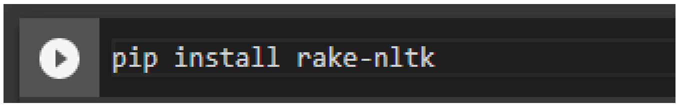
Packages- nltk,rake,matplot.lib, stopwords, punkt, string, heapq and nlargest and sample text file for summarizing.

S.No	Description
1.	NTLK
2.	PUNKT
3.	STRING
4.	NLARGEST
5.	RAKE
6.	MATPLOT.LIB

Table 1.1 Project Requirements

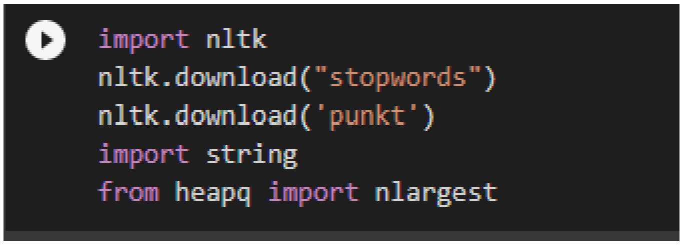
2.4 METHODOLOGY

Step 1: Install nltk Packages by using the following command



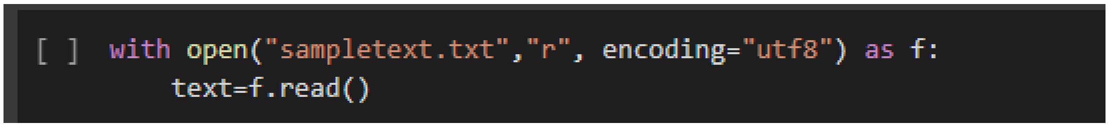
```
pip install rake-nltk
```

Step 2: Download stopwords by using the following command and **:** Import packages, namely nltk, stopwords, punkt, string, heapq and nlargest.



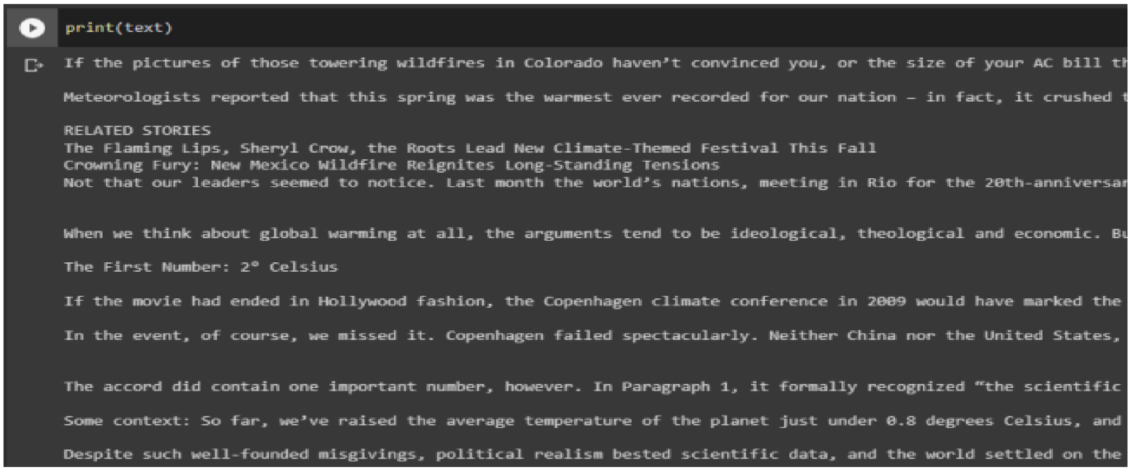
```
import nltk
nltk.download("stopwords")
nltk.download('punkt')
import string
from heapq import nlargest
```

Step 3: To open a sample paragraph and read the text we use text=f.read



```
[ ]  with open("sampletext.txt","r", encoding="utf8") as f:
         text=f.read()
```

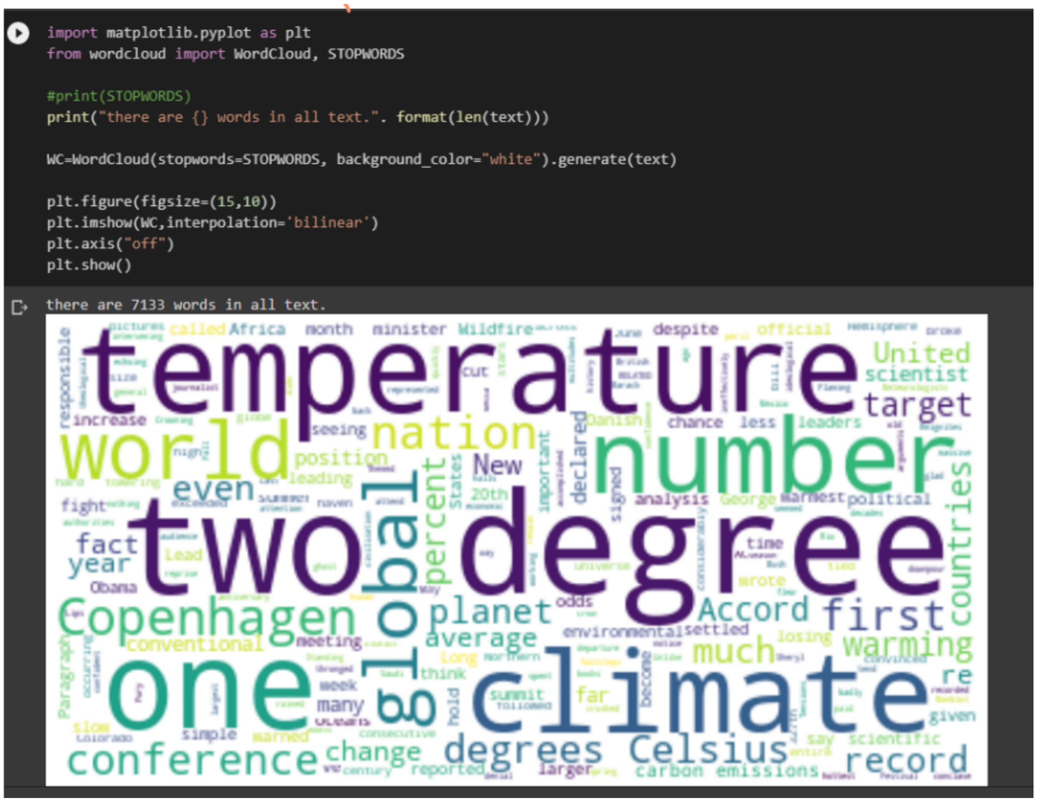
Step 4: To print the text from sample paragraph we use print(text)



```
print(text)
```

```
If the pictures of those towering wildfires in Colorado haven’t convinced you, or the size of your AC bill th

Meteorologists reported that this spring was the warmest ever recorded for our nation – in fact, it crushed t

RELATED STORIES
The Flaming Lips, Sheryl Crow, the Roots Lead New Climate-Themed Festival This Fall
Crowning Fury: New Mexico Wildfire Reignites Long-Standing Tensions
Not that our leaders seemed to notice. Last month the world’s nations, meeting in Rio for the 20th-anniversar


When we think about global warming at all, the arguments tend to be ideological, theological and economic. Bu

The First Number: 2° Celsius

If the movie had ended in Hollywood fashion, the Copenhagen climate conference in 2009 would have marked the

In the event, of course, we missed it. Copenhagen failed spectacularly. Neither China nor the United States,


The accord did contain one important number, however. In Paragraph 1, it formally recognized “the scientific

Some context: So far, we’ve raised the average temperature of the planet just under 0.8 degrees Celsius, and

Despite such well-founded misgivings, political realism bested scientific data, and the world settled on the
```

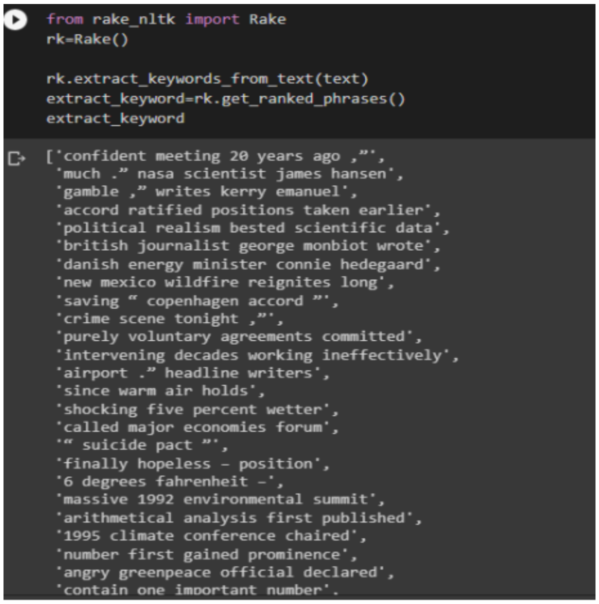
Step 5: Import matplotlib.pyplot package, wordcloud package and then print the wordcloud.



```
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS

#print(STOPWORDS)
print("there are {} words in all text.". format(len(text)))

WC=WordCloud(stopwords=STOPWORDS, background_color="white").generate(text)

plt.figure(figsize=(15,10))
plt.imshow(WC,interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
there are 7133 words in all text.
```

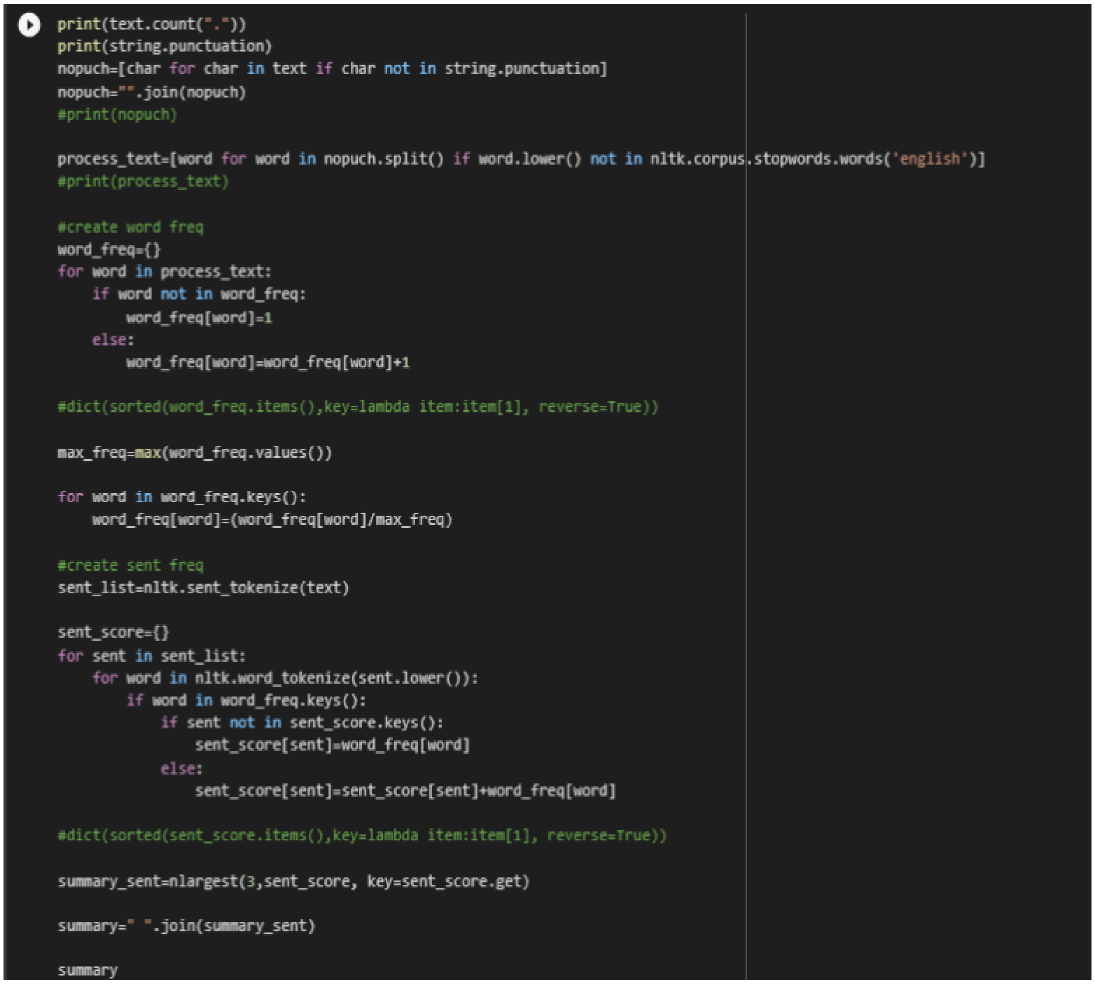
Step 6: Extract keywords from the para



```
from rake_nltk import Rake
rk=Rake()

rk.extract_keywords_from_text(text)
extract_keyword=rk.get_ranked_phrases()
extract_keyword
```

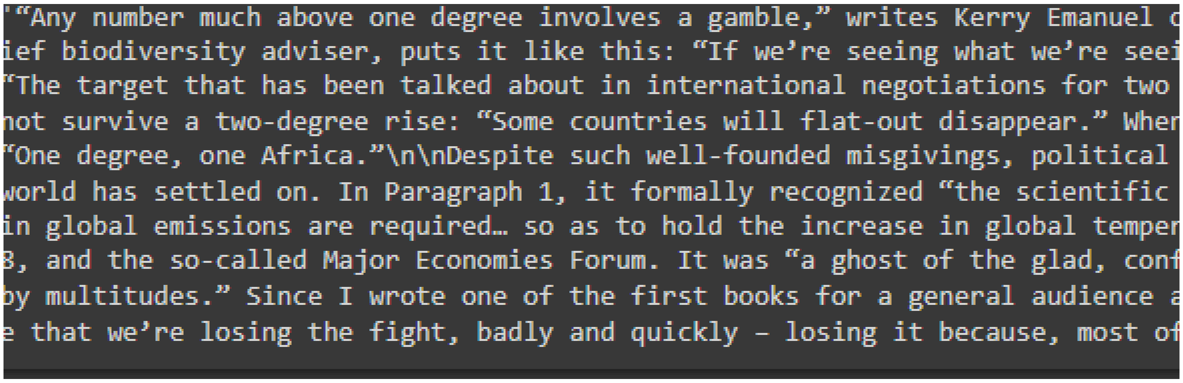
```
['confident meeting 20 years ago ,”',
 'much .” nasa scientist james hansen',
 'gamble ,” writes kerry emanuel',
 'accord ratified positions taken earlier',
 'political realism bested scientific data',
 'british journalist george monbiot wrote',
 'danish energy minister connie hedegaard',
 'new mexico wildfire reignites long',
 'saving “ copenhagen accord ”',
 'crime scene tonight ,”',
 'purely voluntary agreements committed',
 'intervening decades working ineffectively',
 'airport .” headline writers',
 'since warm air holds',
 'shocking five percent wetter',
 'called major economies forum',
 '“ suicide pact ”',
 'finally hopeless – position',
 '6 degrees fahrenheit –',
 'massive 1992 environmental summit',
 'arithmetical analysis first published',
 '1995 climate conference chaired',
 'number first gained prominence',
 'angry greenpeace official declared',
 'contain one important number',
```

Step 8: Process the summary



```
print(text.count("."))
print(string.punctuation)
nopuch=[char for char in text if char not in string.punctuation]
nopuch="".join(nopuch)
#print(nopuch)

process_text=[word for word in nopuch.split() if word.lower() not in nltk.corpus.stopwords.words('english')]
#print(process_text)

#create word freq
word_freq={}
for word in process_text:
    if word not in word_freq:
        word_freq[word]=1
    else:
        word_freq[word]=word_freq[word]+1

#dict(sorted(word_freq.items(),key=lambda item:item[1], reverse=True))

max_freq=max(word_freq.values())

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

#create sent freq
sent_list=nltk.sent_tokenize(text)

sent_score={}
for sent in sent_list:
    for word in nltk.word_tokenize(sent.lower()):
        if word in word_freq.keys():
            if sent not in sent_score.keys():
                sent_score[sent]=word_freq[word]
            else:
                sent_score[sent]=sent_score[sent]+word_freq[word]

#dict(sorted(sent_score.items(),key=lambda item:item[1], reverse=True))

summary_sent=nlargest(3,sent_score, key=sent_score.get)

summary=" ".join(summary_sent)

summary
```

Step 9: Output Summary



```
'“Any number much above one degree involves a gamble,” writes Kerry Emanuel c
ief biodiversity adviser, puts it like this: “If we’re seeing what we’re seei
“The target that has been talked about in international negotiations for two 
not survive a two-degree rise: “Some countries will flat-out disappear.” Wher
“One degree, one Africa.”\n\nDespite such well-founded misgivings, political 
world has settled on. In Paragraph 1, it formally recognized “the scientific 
in global emissions are required… so as to hold the increase in global temper
8, and the so-called Major Economies Forum. It was “a ghost of the glad, conf
by multitudes.” Since I wrote one of the first books for a general audience a
e that we’re losing the fight, badly and quickly – losing it because, most of
```

CHAPTER 3

3.0 RESULTS AND DISCUSSION

3.1 INTRODUCTION

This chapter discusses the various implementation of the codes that have been written to Create this project. The different components working and their responsiveness is shown here And also a brief discussion about the guiding principle and process of the entire project.

3.2 RESULT

We have searched the word INDIA on Wikipedia, which returns 12389. If you search the same keyword simultaneously, get a different number of words produced by the search. It doesn’t matter how many stories have been returned at a specific moment. Our system will work dynamically and It will summarize the result. Every time it will give you summarized data of updated information on Wikipedia. To perform this task, we have integrated python and the Natural Language Toolkit (NLTK) library. In Python, with the integration of the NLTK library, text summarization has been achieved. As shown in Table 1, most occurred words have been calculated, minimizing the overall data. After text summarization, the same message has been transferred to the end-user (Fig. 3).

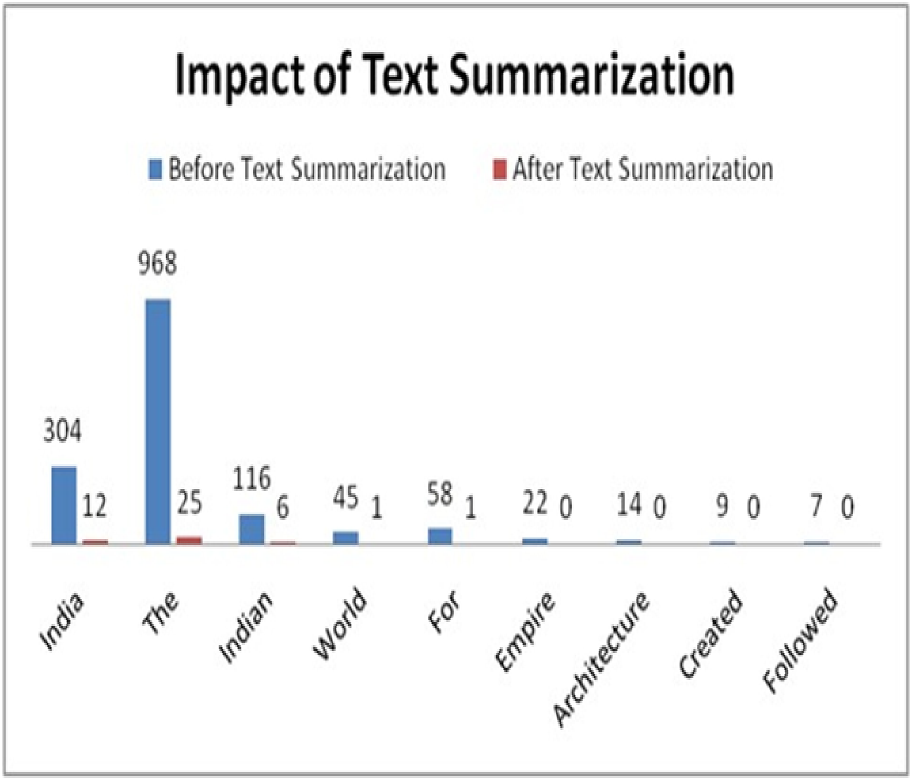


Fig. 3. Impact of text summarization

3.3 DISCUSSION

As you can see, it does a pretty good job. You can further customize it to reduce to number to characters instead of lines.

It is important to understand that we have used text rank as an approach to rank the sentences. TextRank does not rely on any previous training data and can work with any arbitrary piece of text. TextRank is a general-purpose graph-based ranking algorithm for NLP.

CHAPTER 4

4.0 CONCLUSION AND FUTURE SCOPE

4.1 CONCLUSION

Since a lot of information is available on the web server, it is impossible to read each available document to determine the record and know whether it is a required document. Hence, a description of these documents will help the reader decide whether they are available for misappropriation. Some diverse models can find the essential material from the database to produce an extended text description. In this research paper, we have integrated the NLTK library for text summarization. The outcome of this process is the summary of effective content without changing its meaning. Ultimately users can extract enormous information in the form of an outline. An extended range is essential to business analysts, marketing executives, governments, students, researchers, and teachers. The executive is seen as having to be summarized to allow a maximum amount of information to be processed within a limited time frame. This paper delves into the details of both the extractive and the abstractive approaches and their overall performance. The methods and techniques used, with their strengths and shortcomings.

4.2 FUTURE SCOPE

In order to enhance and add more features to our project, we can add much-advanced techniques available for text summarization and we can use python scripts to automate this task giving us more efficiency and reducing manual working dependency.

REFERENCES

[1] Gaikwad, D., Mehender, C.: A review paper on text summarization. Int. J. Adv. Res. Comput. Commun. Eng. 154–160 (2016)

[2] Goncalves, L.: Automatic Text Summarization with Machine Learning — An overview: 12 April 2020 (22 March 2021)

[3] Allahyari, M., et al.: Text summarization techniques: a brief survey. Int.J.Adv.ComputerSci.Appl.8(10),397–405(2017) https://doi.org/10.14569/IJACSA.2017.081052

[4] Moratanch, N., Chitrakala, S.: A survey on extractive text summarization. In: EEE International Conference on Computer, Communication, and Signal Processing, pp. 1–6. IEEE, Chennai, India (2017)

[5] Aggarwal, R., Gupta, L.: Automatic text summarization. Int. J. Comput. Sci. Mobile Comput. 158–167 (2017)

[6] Asawa, Y., Balaji, V., Dey, I.I.: Modern multi-document text summarization techniques. Int. J. Recent Technol. Eng. 9(1), 654–670 (2020). https://doi.org/10.35940/ijrte.A1945.059120

[7]Widyassari, A.P., et al.: Review of automatic text summarization techniques and methods. J. King Saud Univ. Comput. Inform. Sci., p. S1319157820303712 (2020). https://doi.org/10. 1016/j.jksuci.2020.05.006

[8] Arpita Sahoo, A.K.: Review paper on extractive text summarization. Int. J. Eng. Res. Comput. Sci. Eng. 46–52 (2018)

[9]Mehdi Allahyari, S.P.: Text summarization techniques: a brief survey. arXiv, pp. 1–9 (2017)

[10] Neelima, G.V.M.: Extractive text summarization using deep natural language fuzzy process- ing. Int. J. Innov. Tech. Explor. Eng. 990–993 (2019)

[11]Shai Erera, M.S.-S.: A Summarization System for Scientific Document

[12]EMNLP and the 9th IJCNLP, pp. 211–216. Association for Computational Linguistics, Hong Kong (2019)