TEXT PROCESSING IN PYTHON

A PROJECT REPORT

Submitted by

NAME OF THE CANDIDATES
B. V. Karthik Sai Ram (20BAI10018)
Christopher Sojan (20BAI10246)
Arindam Goswami (20BAI10291)
Mahi (20BAI10342)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
in
PROGRAM OF STUDY

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

KOTHRIKALAN, SEHORE
MADHYA PRADESH - 466114

APRIL 2022

VIT BHOPAL UNIVERSITY, KOTHRIKALAN, SEHORE
MADHYA PRADESH – 466114

BONAFIDE CERTIFICATE

Certified that this project report titled **“TEXT PROCESSING IN PYTHON.”** is the bonafide work of “**B. V. Karthik Sai Ram (20BAI10018), Christopher Sojan(20BAI10246), Arindam Goswami(20BAI10291), Mahi(20BAI10342)**” who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROGRAM CHAIR
Dr. S. Sountharrajan,Programme Chair
B.Tech CSE(AI & ML)
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

PROJECT GUIDE
Dr. S. Suthir
Senior Assistant Professor
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on 29th April 2022

ACKNOWLEDGEMENT

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to Dr. S. Sountharrajan, Programme Chair B.Tech CSE (AI & ML), School of Computer Science and Engineering for much of his valuable support and encouragement in carrying out this work.

I would like to thank my internal guide Dr. S. Suthir, Senior Assistant Professor, for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of theSchool of Computer Science and Engineering, who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

LIST OF FIGURES AND GRAPHS

ABSTRACT

Using a practical learning environment, we wanted to get a better understanding of how a simple sentence is constructed, framed and structured and put into practice the concepts of Text and Natural Language Processing. What we are hoping to accomplish with this study is a deeper understanding of sentiment analysis, information extraction, machine translation and question answering. We worked on a project that will have a profound impact on the way students approach their education. It is possible to conduct three separate procedures on the same phrase or paragraph. Summarize, locate the keywords, and rephrase the content of a sentence or paragraph. Paraphrasing involves coming up with other ways to present a text. Use the Summarize tool to shorten a long and redundant text. Using the keyword finder, you may learn about the most significant and relevant keywords in a paragraph and utilize that information to build your own paragraph around those key phrases.

TABLE OF CONTENTS

PROJECT DESCRIPTION AND OUTLINE

1.1 Introduction

Python Programming can be used to process text data for the requirements in various textual data analysis. A very important area of application of such text processing ability of python is for NLP (Natural Language Processing). NLP is used in search engines, newspaper feed analysis and more recently for voice -based applications like Siri and Alexa. Python's Natural Language Toolkit (NLTK) is a group of libraries that can be used for creating such Text Processing systems.

Text processing has a direct application to Natural Language Processing, also known as NLP. NLP is aimed at processing the languages spoken or written by humans when they communicate with one another. This is different from the communication between a computer and a human where the communication is either a computer program written by humans or some gesture by humans like clicking the mouse at some position. NLP tries to understand the natural language spoken by humans and classify it, analyze it as well if required to respond to it. Python has a rich set of libraries which cater to the needs of NLP. The Natural Language ToolKit (NLTK) is a suite of such libraries which provides the functionalities required for NLP.

1.2 Motivation for the work

We collaborated on a project that will have a significant impact on how students approach their education. Three distinct procedures can be applied to the same phrase or paragraph. Summarize, locate the keywords, and rewrite a sentence or paragraph's content. Paraphrasing is the process of coming up with alternative ways to present a text. To shorten a long and redundant text, use the Summarize tool. You can use the keyword finder to learn about the most important and relevant keywords in a paragraph and then use that information to create your own paragraph based on those key phrases.

1.3 Problem Statement

Many students all over the world struggle with their studies while attempting to prepare for their exams. At the last minute, individuals may be unable to cope with the stress of exams. So, in order to assist many students, teachers, and other working professionals, we have devised a solution that will make their lives easier while also assisting them with their day-to-day activities.

1.4 Objective of the work

The objective of our project is to help students with their studies. It helps the students use three key features with their study material. It helps summarize, paraphrase as well as extract all the keywords used in the sentence. With the help of these tools students can shorten, rephrase or find out the important and key terms being used in a given sentence.

RELATED WORK INVESTIGATION

2.1 Introduction

Many different approaches are there for this but we have chosen to use modules and libraries which make use of Natural Language Processing as it helps us give an accurate and an accurate answer both grammatically and structurally

2.2 Existing Approaches/Methods

There are many existing approaches and methods, but they primarily focus on the paraphrasing and summarizing part of our idea, and few of them are used to find the keywords of a given paragraph or sentence. We take a sentence and make appropriate changes to it using Natural Language Processing modules in Python, and then present the output. In addition, when compared to the rest of the existing methods and approaches on the market, the accuracy of the paraphrase and summarize tool is extremely high.

2.3 Issues / Observations from Investigations

We have observed that our project gives the most accurate answer when we use a compound or a complex sentence. It fails to give an ideal answer when a simple sentence is given in as an input. The answer tends to be grammatically incorrect and also conveys a different meaning than it is supposed to give.

REQUIREMENT ARTIFACTS

3.1 Hardware and Software requirements

1. Windows 7 or 10
2. Mac OS X 10.11 or higher, 64-bit
3. Linux: RHEL 6/7, 64-bit (almost all libraries also work in Ubuntu)
4. x86 64-bit CPU (Intel / AMD architecture)
5. 8 GB RAM (16 GB RAM Recommended)
6. 20 GB free disk space
7. Python
8. IDE (PyCharm, VSCode, etc.)
9. Required Libraries for implementation
10. Google Colab to be used while Implementation through Website

DESIGN METHODOLOGY AND ITS NOVELTY

4.1 Methodology and goal

The increment in the usage of Social Media has grown the size of text data, and boosted the studies or research in Natural Language Processing (NLP), for example, Information Retrieval and Sentiment Analysis. Most of the time, the documents or the text files to be analyzed are gigantic and contain a lot of noise, directly used raw texts for analysis is inapplicable. Hence, text processing is essential to provide clean input for modeling and analysis.

Text processing contains two main phases, which are tokenization and normalization. Tokenization is the process of splitting a longer string of text into smaller pieces, or tokens. Normalization refers to converting numbers to their word equivalent, removing punctuation, converting all text to the same case, removing stopwords, removing noise, lemmatization and stemming.

We intend to learn more and implement the concepts of Text and Natural Language Processing through a practical medium.

Through this project we intend to have a better understanding of information retrieval, sentiment analysis, information extraction, machine translation and question answering.

4.2 Software Architecture Diagram

Using tokenizer to separate the sentences into a list of single words (tokens).

Stopwords referring to the word which does not carry much insight, such as preposition.

We remove the tokens which are not words, for example, the symbols and numbers, also tokens that only contain less than two letters or contain only consonants. This might not be useful in a lot of cases, but it’s pretty useful when you are dealing with a large set of text data, it helps to clean up much noise.

This process refers to tagging the word with their part-of-speech position, for example, verbs, adjectives and nouns. The POS Tag is the task performed straight after tokenization, which is a smart move when you know you only need a particular part of speech like adjectives and nouns.

Lemmatization and stemming both help to reduce the dimension of the vocabulary by returning the words to their root form (lemmatizing) or remove all the suffix, affix, prefix and so on (stemming).

Stemming is nice for reducing the dimension of vocabulary, but most of the time the word becomes meaningless as stemming only chopped off the suffix but not returning the words to their base form. For example, *houses* will become *hous* after stemming, which completely loses its meaning. Hence, lemmatizing is more preferable for text analytics. It is used to obtain the root form for the nouns, adjectives and verbs.

After tokenizing and normalizing the text, now we obtained a list of clean tokens, which are ready to be plug-in into WordCloud or other text analytics models.

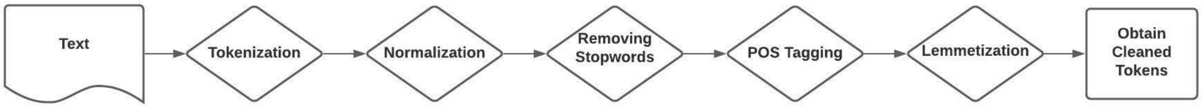


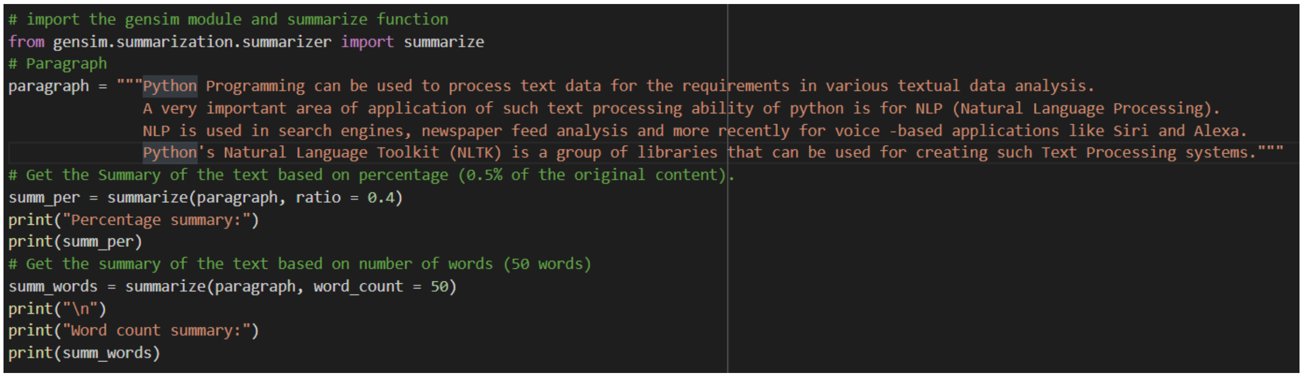
Figure 1 - Software Architecture Diagram

TECHNICAL IMPLEMENTATION & ANALYSIS

5.1 Outline

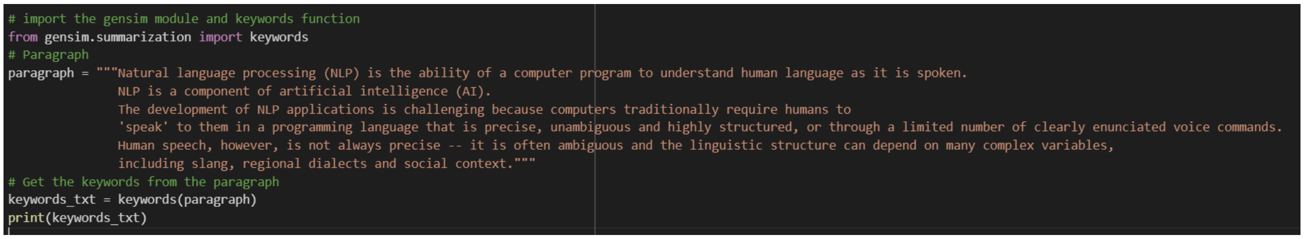
We have designed a basic website using wix which gives a brief about our project and talks about how it works. It also has the team information in the about section in the heading. In the website there is a button under the “How it Works?” section called Implementation when we click on this button it redirects us to Google Colab where we can either implement the code or directly download it.

5.2 Technical coding and code solutions



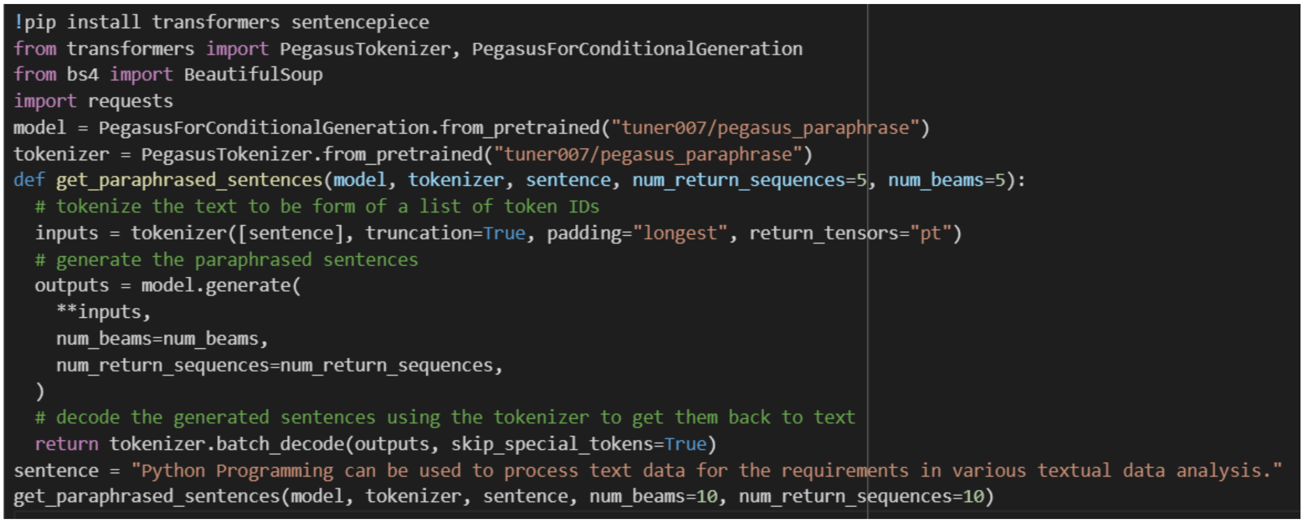
```
# import the gensim module and summarize function
from gensim.summarization.summarizer import summarize
# Paragraph
paragraph = """Python Programming can be used to process text data for the requirements in various textual data analysis.
               A very important area of application of such text processing ability of python is for NLP (Natural Language Processing).
               NLP is used in search engines, newspaper feed analysis and more recently for voice -based applications like Siri and Alexa.
               Python's Natural Language Toolkit (NLTK) is a group of libraries that can be used for creating such Text Processing systems."""
# Get the Summary of the text based on percentage (0.5% of the original content).
summ_per = summarize(paragraph, ratio = 0.4)
print("Percentage summary:")
print(summ_per)
# Get the summary of the text based on number of words (50 words)
summ_words = summarize(paragraph, word_count = 50)
print("\n")
print("Word count summary:")
print(summ_words)
```

Figure 2 - Summarize Tool



```
# import the gensim module and keywords function
from gensim.summarization import keywords
# Paragraph
paragraph = """Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken.
             NLP is a component of artificial intelligence (AI).
             The development of NLP applications is challenging because computers traditionally require humans to
             'speak' to them in a programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands.
             Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables,
             including slang, regional dialects and social context."""
# Get the keywords from the paragraph
keywords_txt = keywords(paragraph)
print(keywords_txt)
```

Figure 3 - Key Word Checker

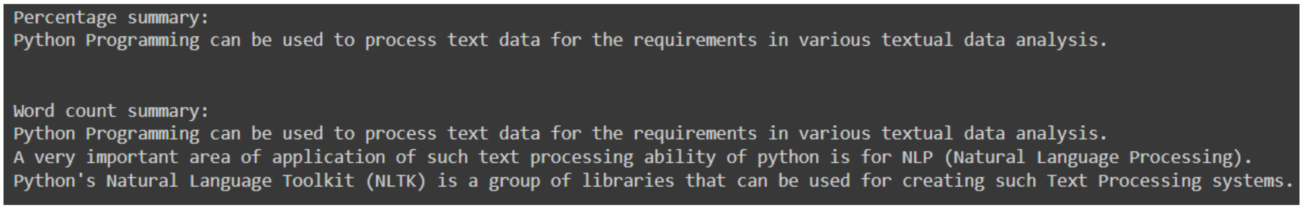


```
!pip install transformers sentencepiece
from transformers import PegasusTokenizer, PegasusForConditionalGeneration
from bs4 import BeautifulSoup
import requests
model = PegasusForConditionalGeneration.from_pretrained("tuner007/pegasus_paraphrase")
tokenizer = PegasusTokenizer.from_pretrained("tuner007/pegasus_paraphrase")
def get_paraphrased_sentences(model, tokenizer, sentence, num_return_sequences=5, num_beams=5):
  # tokenize the text to be form of a list of token IDs
  inputs = tokenizer([sentence], truncation=True, padding="longest", return_tensors="pt")
  # generate the paraphrased sentences
  outputs = model.generate(
    **inputs,
    num_beams=num_beams,
    num_return_sequences=num_return_sequences,
  )
  # decode the generated sentences using the tokenizer to get them back to text
  return tokenizer.batch_decode(outputs, skip_special_tokens=True)
sentence = "Python Programming can be used to process text data for the requirements in various textual data analysis."
get_paraphrased_sentences(model, tokenizer, sentence, num_beams=10, num_return_sequences=10)
```

Figure 4 - Paraphrase Tool

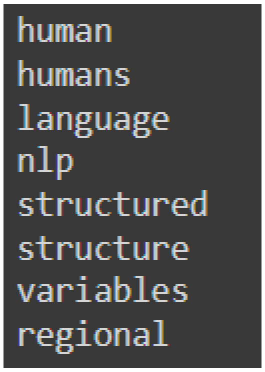
5.3 Test and validation

Output:



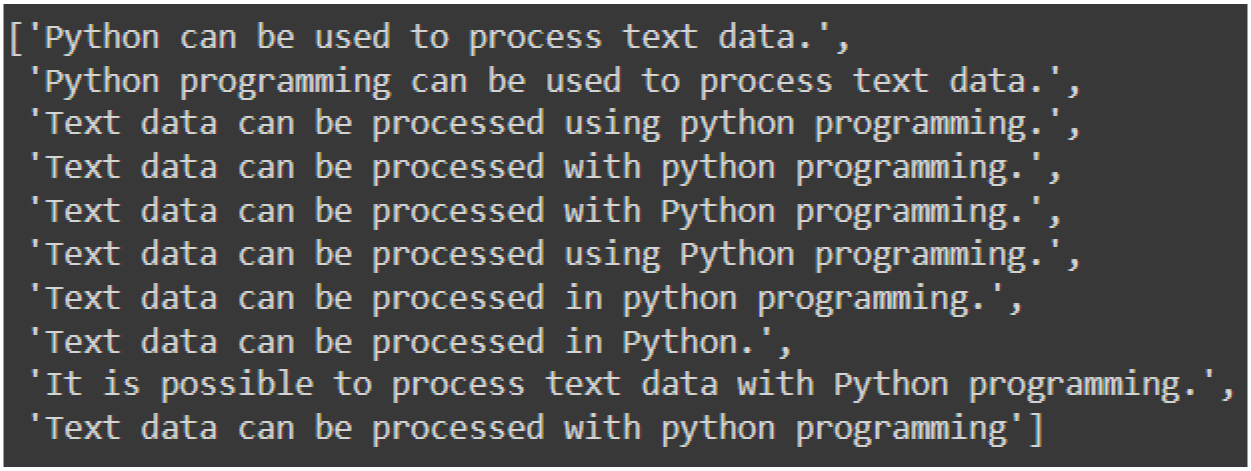
```
Percentage summary:
Python Programming can be used to process text data for the requirements in various textual data analysis.


Word count summary:
Python Programming can be used to process text data for the requirements in various textual data analysis.
A very important area of application of such text processing ability of python is for NLP (Natural Language Processing).
Python's Natural Language Toolkit (NLTK) is a group of libraries that can be used for creating such Text Processing systems.
```

Figure 5 - Output for Summarize Tool



```
human
humans
language
nlp
structured
structure
variables
regional
```

Figure 6 - Output for Keyword Checker



```
['Python can be used to process text data.',
 'Python programming can be used to process text data.',
 'Text data can be processed using python programming.',
 'Text data can be processed with python programming.',
 'Text data can be processed with Python programming.',
 'Text data can be processed using Python programming.',
 'Text data can be processed in python programming.',
 'Text data can be processed in Python.',
 'It is possible to process text data with Python programming.',
 'Text data can be processed with python programming']
```

Figure 7 - Output for Paraphrase Tool

PROJECT OUTCOME AND APPLICABILITY

6.1 Key implementations outlines of the System

TextProcessor is a firm believer in the educational value of its products. For example, it changes the way students study for examinations and produce assignments. There is still a long way to go, but we've already learned that everything we've made has needed a lot of hard effort and a fearless mentality. For our project, we have summarizer, keyword finder and summarizer at its core. Students throughout the globe will benefit greatly from the usage of these three instruments. There will be an increase in the number of pupils who can more easily summarize, paraphrase, and locate keywords in a sentence or paragraph. If you've written a lengthy post or response and would want to reduce it down, Summarize is the solution for you. In order to identify better versions of the same phrase or to increase the quality of the sentence, the programme simply searches for other methods to compose the same paragraph.

6.2 Significant project outcomes

TextProcessor's purpose is to make text simple and powerful. We use Natural Language Processing (NLP) to improve writing. Our algorithms evaluate material in real time and show writers how viewers react. We know that employing words or phrases that connect with particular audiences or industry peers.Our solutions will help pupils, we believe. It changes how students study for examinations, present, and compose assignments. We're only just getting started, but we already know that everything we've produced took devotion and risk.

6.3 Project applicability on Real-world applications

We worked on a project that will have a profound impact on the way students approach their education. It is possible to conduct three separate procedures on the same phrase or paragraph. Summarize, locate the keywords, and rephrase the content of a sentence or paragraph. Paraphrasing involves coming up with other ways to present a text. Use the Summarize tool to shorten a long and redundant text. Using the keyword finder, you may learn about the most significant and relevant keywords in a paragraph and utilize that information to build your own paragraph around those key phrases.

CONCLUSIONS AND RECOMMENDATION

7.1 Limitation/Constraints of the System

We've discovered that when we use a compound or complex sentence, our project provides the most accurate answer. When given a simple sentence as input, it fails to provide an ideal answer. The answer is frequently grammatically incorrect and conveys a different meaning than intended.

7.2 Future Enhancements

With the help of the Keyword Generator we can implement the Search option Optimizer for many websites which help in giving the most accurate results when we use it to search for a movie or a book or it can be a normal article in a newspaper.

7.3 Inference

The goal of TextProcessor is to make text simple and powerful. Natural Language Processing (NLP) is used to improve writing. Our algorithms assess content in real time and show writers how viewers react. We understand the importance of using words or phrases that resonate with specific audiences or industry peers. We believe that our solutions will benefit students. It alters how students prepare for exams, present information, and write assignments. We're only getting started, but we already know that everything we've created requires dedication and risk.

REFERENCES

1. https://towardsdatascience.com/text-processing-in-python-29e86ea4114c

2. https://www.tutorialspoint.com/python_text_processing/index.htm#:~:text=Python%20Programming%20can%20be%20used,NLP%20(Natural%20Language%20Processing).

3. https://realpython.com/nltk-nlp-python/