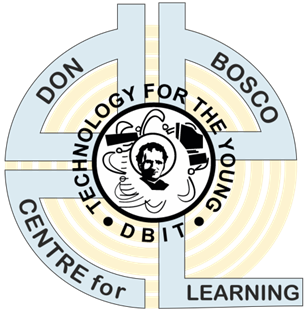
**Mini Project Report**

Topic : Summarizer



Course/Lab: ITL (CODE 404)

Semester: IV

Academic Year: 2021-2022

Subject In-charge: **Mr. Shiv Negi**

Project Group Members

1. **Rhea Dsouza** (Roll No. 16)

2. **Aniket Parate** (Roll No. 38)

3. **Shubham Talawadeka**r (Roll No. 58)

Department of Information Technology

Don Bosco Institute of Technology, Mumbai 400070

Table of contents

| Sr.No. | Content | Page No |
| --- | --- | --- |
| 1. | Problem Statement | 03 |
| 2. | Project Scope | 03 |
| 3. | Relevant Literature on Existing Scenario | 04 |
| 4. | Proposed Solution | 05 |
| 5. | System Design and Implementation | 05 |
| 6. | Conclusion | 07 |
| 7. | References | 07 |
| 8. | Contributions | 07 |
| 9. | Timesheet | 09 |

**Problem Statement:**

Create a Summarizer in order to get a short summary of long paragraphs and with various features. Should be able to handle PDFs as well.

**Project Scope:**

A Summarizer will make it very easy to summarise PDFs or paragraphs in a few seconds. In today’s busy world we cannot spend much time on simple things like writing a summary, so we use technology to make tasks easier and more efficient. Summarising teaches us how to discern the most important ideas in a given text and how to ignore irrelevant information. It also teaches us to integrate the central idea in a meaningful way.

**Relevant Literature on Existing Scenario:**

Some earlier work on text summarization and its methods are listed below :

| **Methods** | **Advantages** | **Disadvantages** |
| --- | --- | --- |
| Simple surface  level features  of sentences. | The sentences which  include most frequent  words are selected as  summary sentences. | Duplication in  summary sentences. |
| Artificial  Intelligence | Frames or templates are used to identify the  conceptual relation of  entities and extract the  relation between entities by an assumption. | Only limited frames  or templates may  lead to incomplete  analysis of  conceptual entities. |
| Cognitive  science  theories | The system can  overcome the  redundancy in some  extent. Extract the  representative sentences from the source text. | Complex task and  limited to specified  area. |
| Information  retrieval  techniques | Generate significant  sentence from source  text same as information retrieval  techniques. | Doesn’t consider the  semantic aspects  such as synonym  and polysemy |
| Machine learning  techniques | Different machine learning algorithms are  used and provides a more generalised summary. | Computationally complex and lack of  semantic analysis of  source text. |
| Statistical and  Algebraic  methods | Depended on some  heuristics, linguistics  and mathematical  techniques. Easy to  implement | Without any  syntactic analysis of  the source text. |

**Proposed Solution:**

There are mainly two approaches that have been developed over time for summarising a long text into a shorter one.

**Extraction Summarization:**

This approach entails the method to extract keywords and phrases from sentences and then joining them to produce a compact meaningful summary.

### **Abstractive Summarization:**

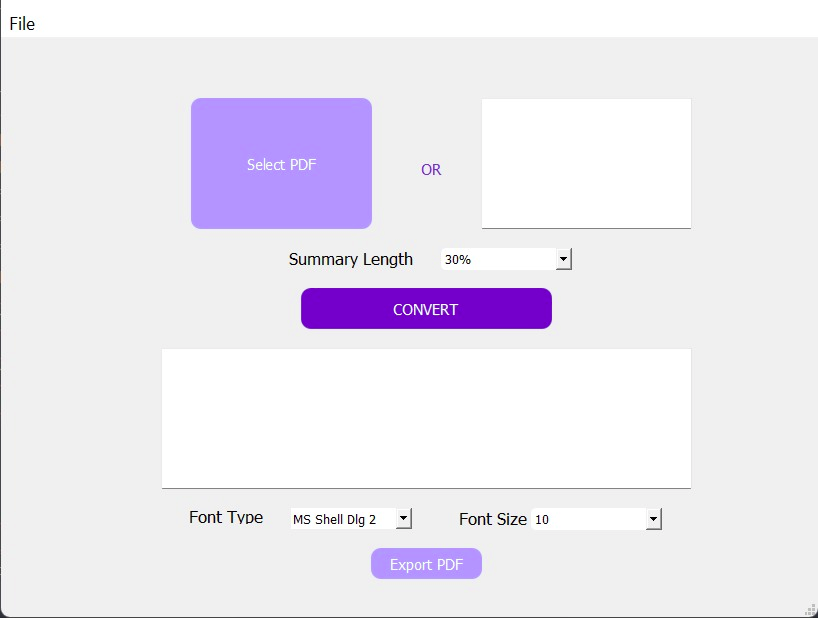
### In this method, algorithms are developed in such a way to reproduce a long text into a shorter one by NLP. It retains its meaning but changes the structure of sentences.

We have opted for the Extraction Summarization method.

**System Design and Implementation:**

First, we have two options, we can either select a PDF file from our computer or we can directly paste the text into the given area. Then we have a dropdown menu containing the different percentages to which the text should be summarised. The percentages are 30%, 50%, and 75%. Later we have the Convert button, when the user clicks on the convert button, the text starts getting summarised. In a few seconds, the result of the conversion appears in the large output box given below. Now, the user can select different fonts and font sizes from the drop-down menu provided below. There is an export PDF button at the bottom, once the user clicks on it the text result will be converted into a PDF. This PDF can be saved where the user wants. The font and the font size in this PDF will be according to what the user has chosen.

**UI Design:**



**Technologies used:**

* **PyQt5** - UI Development
* **NLTK** - tokenization and remove stopwords
* **fPDF** - create pdf from text
* **pdfPlumber** - extract text from pdf

**Conclusion:**

We have created a python based application to summarise text with various features.

**References:**

<http://ischolar.info/index.php/IJCSIT/article/download/120562/109066>

[Text Summarization with NLTK in Python (stackabuse.com)](https://stackabuse.com/text-summarization-with-nltk-in-python/)

<https://pyfpdf.github.io/fpdf2>

[NLTK :: Natural Language Toolkit](https://www.nltk.org/)

**Contribution :**

Shubham Talawadekar - Report, Code Research, Connecting different

codes

Aniket Parate - Report, Code Logic, Integration, Research

Rhea Dsouza - UI Design, Report, Connection of UI to the

code

**Learnings :**

**Aniket Vinayak Parate**

Python is a programming language that lets you work more quickly and integrate your systems more effectively. It was a very exciting and informative experience while working on this group project. I got to learn about summarization in detail and come across different types of summarization. We also learned how to automate the summarization process as well as we used python for this project and got a better understanding of python coding as well. This project is a purely python based project from coded logic to UI design everything was done with python and its libraries. I also learned UI designing and code integration with the UI and error management.

**Shubham Ramesh Talawadekar** :

Python is one of the most popular and in-demand programming languages in the world. I got to learn it from basic fundamentals to expert level coding. Python makes it easy to automate processes. As explained above, in summarisation, there are two basic types: extraction and abstraction:- which requires a lot of knowledge, and machine learning, so we went with extraction. By reading a few papers, we realised how easy it is to get important sentences from an article. It’s not going to work for stories, but it can get the gist of research papers. It is essential to know what is possible and what part of the project will take more time than other things. As we progressed, we realised which parts of the project we should keep and which parts we should remove. I think the biggest learning from this is understanding the difficulty of the project and planning ahead for the flow of making and completing the project.

**Rhea Donald Dsouza:**

It was a wholesome experience in which I got to improve my UI Designing skills, python coding and team work. We faced various challenges while creating the functional application and adding different features to it. We all had our own ideas and have implemented it in the project. I learned how to integrate all the required fields in the UI, and also overcame the challenges faced while coding it and connecting the UI to the Backend. Also, through this project I have got a better understanding of summarization and how to implement it in the code. While working on the project I got a clear idea of the direction in which I have to think in order to fulfil the needs of the user and how to make it more user friendly.

**Timesheet:**

| **Sr. No.** | **Date** | **Hours** |
| --- | --- | --- |
| 1 | 3/2/2022 | 2 |
| 2 | 4/2/2022 | 3 |
| 3 | 5/2/2022 | 2.5 |
| 4 | 18/2/2022 | 1 |
| 5 | 19/2/2022 | 1.5 |
| 6 | 20/2/2022 | 1 |
| 7 | 24/2/2022 | 2 |
| 8 | 25/2/2022 | 1 |
| 9 | 8/3/2022 | 2 |
| 10 | 15/3/2022 | 3 |
| 11 | 18/3/2022 | 1 |
| 12 | 29/3/2022 | 2 |
| 13 | 2/4/2022 | 3 |
| 14 | 13/4/2022 | 1 |
| 15 | 24/4/2022 | 2.5 |
| 16 | 29/4/2022 | 3 |
| 17 | 1/5/2022 | 2 |
| 18 | 2/5/2022 | 2.5 |
| 19 | 3/5/2022 | 4 |
| 20 | 5/5/2022 | 1.5 |
| 21 | 7/5/2022 | 3.5 |
| 22 | 8/5/2022 | 4 |