**TEXTUAL, URL AND FILE SUMMARIZATION**

**USING BASIAN ALGORITHAM**

**A PROJECT REPORT**

***Submitted by***

**PRITHVIRAJ R (420418205302)**

**SRIKANDHAN K (420418205303)**

**TAMIZHSELVAN R (420418205304)**

***in partial fulfillment for the award of the degree of***

# BACHELOR OF TECHNOLOGY

***IN***

# INFORMATION TECHNOLOGY

**ADHIPARASAKTHI ENGINEERING COLLEGE**

**MELMARUVATHUR**



**ANNA UNIVERSITY :: CHENNAI 600 025**

**JUNE 2022**

# BONAFIDE CERTIFICATE

Certified that this report titled “**TEXTUAL, URL AND FILE SUMMARIZATION USING BASIAN ALGORITHAM**”, is a bonafide work of **Mr. PRITHIVRAJ R (420418205302), Mr. SRIKANDHAN K (420418205303) and Mr. TAMIZHSELVAN R (420418205304),** who carried out the work under my supervision.

## SIGNATURE SIGNATURE

### Dr. A. BHUVANESWARI M.E., Ph.D., Mr. M. EZHILVENDAN B.Tech., M.E.,

**PROFESSOR, ASSISTANT PROFESSOR,**

**HEAD OF THE DEPARTMENT SUPERVISOR**

Department of Information Technology, Department of Information Technology, Adhiparasakthi Engineering College, Adhiparasakthi Engineering College, Melmaruvathur-603 319. Melmaruvathur-603 319.

**(ii)**

## CERTIFICATE OF EVALUATION

College Code/Name : 4204/ Adhiparasakthi Engineering College Branch/semester : Information Technology / VIII

|  |  |  |  |
| --- | --- | --- | --- |
| **S.**  **No** | **Names of the Students**  **& Register Number** | **Title of the Project** | **Name of the Supervisor**  **with Designation** |
| 1. | **Mr. PRITHIVRAJ R**  (**420418205302)** | **TEXTUAL, URL**  **AND FILE SUMMARIZATION**  **USING BASIAN ALGORITHAM** | **Mr. M. EZHILVENDAN**  B.Tech., M.E.,  Assistant Professor  Department of Information Technology,  Adhiparasakthi Engineering College,  Melmaruvathur. |
| 2. | **Mr. SRIKANDHAN K**  **(420418205303)** |
| 3. | **Mr. TAMIZHSELVAN R**  (**420418205304)** |
|  |

The reports of the project work submitted by the above students in partial fulfillment for the award of **Bachelor of Technology** degree in **Information Technology** of Anna University were evaluated and confirmed to be the work done by the above students and then evaluated.

Submitted for the Project Work and Viva-Voce Examination held on…………

## INTERNAL EXAMINER EXTERNAL EXAMINER

**(iii)**

**ACKNOWLEDGEMENT**

It is indeed a great pleasure and proud privilege to acknowledge the help and support we received from the positive minds around us in making this endeavor a successful one. The spiritual blessings of His Holiness **ARULTHIRU AMMA** and the divine guidance **THIRUMATHI AMMA** have undoubtedly taken us to the path of victory in completing this project.

The infrastructural support with all kinds of lab facilities have been a motivating factor in this completion of project work, all because of our **CORRESPONDENT, SAKTHI THIRU Dr. G.B. SENTHIL KUMAR B.E.,** with great pleasure we take this opportunity to thank him.

From the academic side the constant support from our **PRINCIPAL Dr. J. RAJA., M.E., Ph.D.,** has encouraged us to work hard and attain this goal of completing the project. We sincerely thank our motivating and respected to **HEAD OF THE DEPARTMENT Dr. A. BHUVANESHWARI M.E., Ph.D.,** and **PROJECT COORDINATOR Mrs. N. JAINTHI., M.E., who** has given us both moral and technical support adding experience to the job we have undertaken. We thank our **SUPERVISOR Mr. M. EZHILVENDAN B.Tech., M.E.,** who helped us in crossing in the path to our glory and we also thank other staff members who have given their constant support and motivation in all our endeavors.

I takes immense pleasure in thanking my parents who gave me moral supports to complete this project

**(iv)**

**ABSTRACT**

The project entitled **TEXTUAL, URL and File summarization** aims at creating a summary of a certain document that contains the most important information of the original one. Summarization of multi-documents aims to generate a concentrated version of the document while keeping the main information. Due to the massive amount of data these days, the importance of summarization arises.

The **TEXTUAL, URL and File summarization** has the potential to reduce the time, required to familiarize oneself with the subject of studies by condensing long-form detailed descriptions to concise, meaning-preserving synopses.

**(v)**

### 

|  |  |  |
| --- | --- | --- |
| **CONTENTS** | | |
| **CHAPTER** | **TITLE** | **PAGE NO** |
|  | **BONAFIDE CERTIFICATE** | ii |
|  | **ACKNOWLEDGEMENT** | iv |
|  | **ABSTRACT** | v |
|  | **CONTENTS** | vi |
|  | **LIST OF FIGURES** | viii |
| 1 | **INTRODUCTION** | 1 |
|  | 1.1 OVERVIEW | 1 |
|  | 1.2 MOTIVATION | 2 |
|  | 1.3 PROBLEM STATEMENT | 2 |
|  | 1.4 PROPOSED SYSTEM | 3 |
| 2 | **LITERATURE SURVEY** | 4 |
| 3 | **PROPOSED METHOD** | 9 |
|  | 3.1 DEEP REPRESENTATION OF A DOCUMENT | 10 |
|  | 3.1.1 PRE-PROCESSING | 10 |
|  | 3.1.2 PRE-TRAINED SENTENCE MODEL | 10 |
|  | 3.1.3 POSITIONAL ENCODING | 11 |
|  | 3.1.4 SCALED DOT-PRODUCT ATTENTION | 12 |
|  | 3.2 PRINCIPAL COMPONENT ANALYSIS | 13 |
|  | 3.3 INTEGER LINEAR PROGRAMMING | 15 |
|  | 3.3.1 OPTIMIZATION | 15 |
|  | 3.3.2 SENTENCE PRUNING | 16 |
|  | 3.3.3 EVALUATING THE  SUMMARIZATION SYSTEMS | 17 |
|  | **(vi)** |  |
|  | 3.4 TEXT SUMMARIZATION TECHNIQUES | 17 |
|  | 3.5 TEXT SUMMARIZATION IN NLP | 18 |
|  | 3.5.1 EXTRACTIVE SUMMARIZATION | 19 |
|  | 3.5.2 ABSTRACTIVE SUMMARIZATION | 20 |
|  | 3.6 INTRODUCTION TO SEQUENCE-TO-SEQUENCE (SEQ2SEQ) MODELING | 21 |
|  | 3.7 UNDERSTANDING THE ENCODER-DECODER ARCHITECTURE | 22 |
|  | 3.7.1 TRAINING PHASE | 23 |
|  | 3.7.2 ENCODER | 24 |
|  | 3.7.3 DECODER | 24 |
|  | 3.7.4 INFERENCE PHASE | 25 |
| **4** | **IMPLEMENTATION** | 27 |
|  | 4.1 GUI INTEGRATION | 27 |
|  | 4.2 IMPLEMENTATION | 28 |
|  | 4.2.1 PARAGRAPH TO SENTENCE AND  CLEANING | 28 |
|  | 4.2.2 POSITIONAL ENCODING | 29 |
|  | 4.2.3 ATTENTION | 30 |
|  | 4.2.4 PCA | 30 |
|  | 4.2.5 ILP | 31 |
|  | 4.3 SUMMARY | 31 |
| **5** | **CONCLUSION AND FUTURE WORK** | 37 |
|  | 5.1 CONCLUSION | 37 |
|  | 5.2 FUTURE WORK | 37 |
| **6** | **REFERENCES** | 38 |
|  | **(vii)** |  |

|  |
| --- |
| **LIST OF FIGURES** |

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Figure Name** | **Page No** |
| Figure 3.1 | Block diagram | 9 |
| Figure 3.2 | Attention Mechanism | 12 |
| Figure 3.4 | Text Summarization Techniques | 17 |
| Figure 3.5 | Text Summarization in NLP | 18 |
| Figure 3.5.1 | Extractive Summarization | 19 |
| Figure 3.5.2 | Abstractive Summarization | 20 |
| Figure 3.6 | Introduction to sequence-to-sequence (seq2seq) modeling | 22 |
| Figure 3.7 | Understanding the encoder-decoder Architecture | 23 |
| Figure 3.7.2 | Encoder | 24 |
| Figure 3.7.3 | Decoder | 25 |
| Figure 3.7.4 | Inference Phase | 26 |
| Figure 4.0 | Implementation in GUI | 27 |
| Figure 4.1 | GUI Homepage | 28 |
| Figure 4.2.1 | Cleaned Text | 29 |
| Figure 4.2.2 | Positional encoding | 29 |
| Figure 4.2.3 | Attention Scores | 30 |
| Figure 4.2.4 | Principal components | 30 |
| Figure 4.2.5 | Sentence importance | 31 |
| Figure 4.3 | Input Before Textbox summarization | 31 |
|  | **(viii)** |  |

|  |  |  |
| --- | --- | --- |
| Figure 4.3.1 | Output After Text Summarization | 32 |
| Figure 4.3.2 | Input Before URL Summarization | 33 |
| Figure 4.3.3 | Output After URL Summarization | 34 |
| Figure 4.3.4 | Input Before File Summarization | 35 |
| Figure 4.3.5 | Output After File Summarization | 36 |

**(ix)**