***MAJOR PROJECT REPORT***

***TEXT SUMMARIZATION USING NLP***

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

**Bachelor of Technology Computer Science and Engineering**

Guide(s):

Anil Kumar Vats

Asst. Professor

Dept. of Computer Science Submitted by:

Pratyush Garg

35413302717

Suhit Garg

04513302717

Shivam Goel

03913302717

Yash Sharma

04813302717

**HMR INSTITUTE OF TECHNOLOGY & MANAGEMENT**

**HAMIDPUR, DELHI 110 036**

**Affiliated to**

**GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY**

**Sector - 16C Dwarka, Delhi - 110075, India**

**DECLARATION**

We hereby declare that the project report work titled “ TEXT SUMMARIZATION USING NLP “ submitted to the HMR Institute of Technology & Management, Hamidpur Delhi , is a record of an original work done by the us under the guidance of Mr. Anil Kumar Vats , Department of Computer Science and Engineering , HMRITM and with the aid of training provided during the summer training program. This project is submitted in the partial fulfilment of the requirements of the award of the degree of Bachelor of Technology in Computer Science & Engineering. The results embodied in this report work have not been submitted anywhere or in any other university for any award for any degree or diploma.

**ACKNOWLEDGEMENT**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them.

We are highly indebted to Mr Anil Kumar Vats for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

We would like to express our gratitude towards members of HMR for their kind co-operation and encouragement which help me in completion of this project.

We would like to express our special gratitude and thanks to industry persons for giving us such attention and time.

Suhit Garg (04513302717)

Pratyush Garg(35413302717)

Yash Sharma (04813302717)

Shivam Goel (03913302717)

**ABSTRACT**

Summarization has become a very helpful way of tackling the issue of data overburden in the 21st century. What is text summarization? It’s basically a task to generate an accurate summary while maintaining key information and not losing overall meaning. Due to an exponential growth in the generation of textual data, the need for tools and mechanisms for automatic summarization of documents has become very critical. Text documents are vital to any organization's day-to-day working and as such, long documents often hamper trivial work. Therefore, an summarizer is vital towards reducing human effort. Text summarization is an important activity in the analysis of a high-volume text documents and is currently a major research topic in Natural Language Processing. It is the process of generation of the summary of input text by extracting the representative sentences from it. In this project, we present a novel technique for generating the summarization of domain specific text by using Semantic Analysis for text summarization, which is a subset of Natural Language Processing.

**TABLE OF CONTENTS**

**Declaration i**

**Acknowledgement ii**

**Abstract iii**

**Chapter– 1 Introduction**

1.1 Introduction 1

1.2 Overview of the project 2

1.3 Problem Statement 3

1.4 Scope of Study 6

**Chapter – 2 System Analysis**

2.1 Introduction 8

2.2 SRS 17

2.3 Feasibility Study 23

**Chapter – 3 System Design**

3.1 Design Methods 27

3.2 Various development approaches 28

**Chapter –4 Testing**

4.1 Testing techniques and strategies 30

4.2 Debugging & Improvement 32

4.3 Outputs 35

**Chapter – 5 Implementation**

5.1 System Implementation 42

5.2 Software Implementation 44

5.2.1 Current Uses 46

5.2.2 Ongoing research and improvement 46

**Conclusion & Future Uses 47**

**References 48**

**Chapter 1**

**Introduction**

**1.1 Introduction**

Text summarization (or automatic summarization) is the creation of a shortened version of a text by a computer program. The product of this procedure still contains the most important points of the original text and is generally referred to as an abstract or a summary. Broadly, one distinguishes two approaches to text summarization: extraction and abstraction. Extraction techniques merely copy information deemed to be most important by the system to the summary, while abstraction involves paraphrasing sections of the source document. In general, abstraction can produce summaries that are more condensed than extraction, but these programs are considered much harder to develop. Both techniques exploit the use of natural language processing and/or statistical methods for generating summaries. Extractive summarization is basically creating a summary based on strictly what you get in the text. It can be compared to copying down the main points of a text without any modification to those points and rearranging the order of that points and the grammar to make more sense out of the summary. Abstractive summarization techniques tend to mimic the process of ‘paraphrasing’ from a text than just simply summarizing it. Texts summarized using this technique looks more human-like and produce more condensed summaries. These techniques are much harder to implement than extractive summarization techniques in general.

**1.2 Overview of the Project**

The proposed system as shown in figure 1 uses Latent Semantic Analysis [1] to summarize documents from the user. The user inputs a document to the summarizer which has classes derived from the NLP libraries implemented on it. These classes are a collection of semantic rules (which allows the system to group the content using world knowledge) and dictionaries, which aid in the semantic analysis. The input document is first parsed or pre-processed, wherein there is a removal of unneeded words such as ‘stop words’ which are simply small function words, like “the”, “and”, “a”, which do not contribute meaning to the text summary. After the removal of the stop words, sentences are tokenised for the purpose of words with the same root word is taken as single entity, for example, ‘understanding’ and ‘understand’ are not taken as different words when scoring is done in the later stages. A word count dictionary is created, more the frequency of the word, higher the relevance of the sentence in which it comes in it. Later, on the basis of the word count dictionary, the sentences are scored by adding individual scores of the words present in sentence. Finally, top N sentences are taken in summary, where N is the total number of sentences divided by 4 and is displayed and saved as a .txt file.

**1.3 Problem Statement**

A problem statement is a concise description of an issue to be addressed or a condition to be improved upon. It identifies the gap between the current (problem) state and desired (goal) state of a process or product. Focusing on the facts, the problem statement should be designed to address the Five W’s. The first condition of solving a problem is understanding the problem, which can be done by way of a problem statement.

Problem statements are widely used by businesses and organizations to execute process improvement projects. A simple and well-defined problem statement will be used by the project team to understand the problem and work toward developing a solution. It will also provide management with specific insights into the problem so that they can make appropriate project-approving decisions. As such, it is crucial for the problem statement to be clear and unambiguous.

The main purpose of the problem statement is to identify and explain the problem. This includes describing the existing environment, where the problem occurs, and what impacts it has on users, finances, and ancillary activities. Additionally, the problem statement is used to explain what the expected environment looks like. Defining the desired condition provides an overall vision for the process or product. It makes clear the purpose for initiating the improvement project and the goals that it is meant to accomplish.

Another important function of the problem statement is to be used as a communication device. Problem statement helps with obtaining buy-in from those involved in the project, before the project begins, the stakeholders verify the problem and goals are accurately described in the problem statement. Once this approval is received, the project team reviews it to ensure everyone understands the issue at hand and what they are trying to accomplish. This also helps define the project scope, which keeps the project concentrated on the overall goal.

The problem statement is referenced throughout the project to establish focus within the project team and verify they stay on track. At the end of the project, it is revisited to confirm the implemented solution indeed solves the problem. A well-defined problem statement can also aid in performing root-cause analysis to understand why the problem occurred and ensure measures can be taken to prevent it from happening in the future.

It is important to note that the problem statement does *not* define the solution or methods of reaching the solution. The problem statement simply recognizes the gap between the problem and goal states. It can be said that, “a problem well stated is half solved.” However, there are often multiple, viable solutions to a problem. Only after the problem statement is written and agreed upon should the solution(s) be discussed, and the resulting course of action determined.

The problem statement is the initial starting point for a project. It is basically a one to three-page statement that everyone on the project agrees with that describes what will be done at a high level. The problem statement is intended for a broad audience and should be written in non-technical terms. It helps the non- technical and technical personnel communicate by providing a description of a problem. It doesn’t describe the solution to the problem.

The input to requirement engineering is the problem statement prepared by customer. It may give an overview of the existing system along with broad expectations from the new system.

The first phase of requirements engineering begins with requirements elicitation i.e., gathering of information about requirements. Here, requirements are identified with the help of customer and existing system processes. So, from here begins the preparation of problem statement. So, basically a problem statement describes what needs to be done without describing how. To tell user the best and top-rated movies, apps, devices that fits in his/her range or fulfil his/her desired needs. Text summarization is the creation of a shortened version of a text by a computer program.

The text summariser will generate a summary of quarter of the total number of sentences in the document by semantic analysis and NLP. It will be deployed using PyQt5.

**1.4 Scope of Study**

Text summarization is one of the major problems in the field of Natural Language Processing, and yet it is even after years of research and implementations, fraught with complications. However, there have been some major breakthroughs in the past. Methods such as Deep Understanding, Sentence Extraction, Paragraph Extraction, Machine Learning, and even some which employ all these methods along with Traditional NLP Techniques (Semantic Analysis, etc.). As such, keeping these accomplishments in mind, there is still ample amount of research left in the domain of Text Summarization, as a meaningful summary is still difficult to attain in all domains and languages.

**Chapter 2**

**System Analysis**

**2.1 Introduction**

Systems development is systematic process which includes phases such as planning, analysis, design, deployment, and maintenance. Here, in this tutorial, we will primarily focus on −

* Systems analysis
* Systems design

## Systems Analysis

It is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components.

System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem-solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose.

Analysis specifies what the system should do.

## Systems Design

It is a process of planning a new business system or replacing an existing system by defining its components or modules to satisfy the specific requirements. Before planning, you need to understand the old system thoroughly and determine how computers can best be used in order to operate efficiently.

System Design focuses on how to accomplish the objective of the system.

System Analysis and Design (SAD) mainly focuses on −

* Systems
* Processes
* Technology

## What is a System?

The word System is derived from Greek word Systema, which means an organized relationship between any set of components to achieve some common cause or objective.

*A system is “an orderly grouping of interdependent components linked together according to a plan to achieve a specific goal.”*

Constraints of a System

A system must have three basic constraints −

* A system must have some structure and behaviour which is designed to achieve a predefined objective.
* Interconnectivity and interdependence must exist among the system components.
* The objectives of the organization have a higher priority than the objectives of its subsystems.

For example, traffic management system, payroll system, automatic library system, human resources information system.

## Properties of a System

A system has the following properties −

### Organization

Organization implies structure and order. It is the arrangement of components that helps to achieve predetermined objectives.

### Interaction

It is defined by the manner in which the components operate with each other.

For example, in an organization, purchasing department must interact with production department and payroll with personnel department.

### Interdependence

Interdependence means how the components of a system depend on one another. For proper functioning, the components are coordinated and linked together according to a specified plan. The output of one subsystem is the required by other subsystem as input.

### Integration

Integration is concerned with how a system component are connected together. It means that the parts of the system work together within the system even if each part performs a unique function.

### Central Objective

The objective of system must be central. It may be real or stated. It is not uncommon for an organization to state an objective and operate to achieve another.

The users must know the main objective of a computer application early in the analysis for a successful design and conversion

### Outputs and Inputs

* The main aim of a system is to produce an output which is useful for its user.
* Inputs are the information that enters into the system for processing.
* Output is the outcome of processing.

### Processor(s)

* The processor is the element of a system that involves the actual transformation of input into output.
* It is the operational component of a system. Processors may modify the input either totally or partially, depending on the output specification.
* As the output specifications change, so does the processing. In some cases, input is also modified to enable the processor for handling the transformation.

### Control

* The control element guides the system.
* It is the decision–making subsystem that controls the pattern of activities governing input, processing, and output.
* The behaviour of a computer System is controlled by the Operating System and software. In order to keep system in balance, what and how much input is needed is determined by Output Specifications.

### Feedback

* Feedback provides the control in a dynamic system.
* Positive feedback is routine in nature that encourages the performance of the system.
* Negative feedback is informational in nature that provides the controller with information for action.

### Environment

* The environment is the “supersystem” within which an organization operates.
* It is the source of external elements that strike on the system.
* It determines how a system must function. For example, vendors and competitors of organization’s environment, may provide constraints that affect the actual performance of the business.

### Boundaries and Interface

* A system should be defined by its boundaries. Boundaries are the limits that identify its components, processes, and interrelationship when it interfaces with another system.
* Each system has boundaries that determine its sphere of influence and control.
* The knowledge of the boundaries of a given system is crucial in determining the nature of its interface with other systems for successful design.

## Types of Systems

The systems can be divided into the following types −

### Physical or Abstract Systems

* Physical systems are tangible entities. We can touch and feel them.
* Physical System may be static or dynamic in nature. For example, desks and chairs are the physical parts of computer centre which are static. A programmed computer is a dynamic system in which programs, data, and applications can change according to the user's needs.
* Abstract systems are non-physical entities or conceptual that may be formulas, representation or model of a real system.

Open or Closed Systems

* For example, an information system which must adapt to the changing environmental conditions.
* A closed system does not interact with its environment. It is isolated from environmental influences. A completely closed system is rare in reality.

### Adaptive and Non-Adaptive System

* Adaptive System responds to the change in the environment in a way to improve their performance and to survive. For example, human beings, animals.
* Non-Adaptive System is the system which does not respond to the environment. For example, machines.

### Permanent or Temporary System

* Permanent System persists for long time. For example, business policies.
* Temporary System is made for specified time and after that they are demolished. For example, A DJ system is set up for a program and it is dissembled after the program.

### Natural and Manufactured System

* Natural systems are created by the nature. For example, Solar system, seasonal system.
* Manufactured System is the man-made system. For example, Rockets, dams, trains.

### Deterministic or Probabilistic System

* Deterministic system operates in a predictable manner and the interaction between system components is known with certainty. For example, two molecules of hydrogen and one molecule of oxygen makes water.
* Probabilistic System shows uncertain behaviour. The exact output is not known. For example, Weather forecasting, mail delivery.

### Social, Human-Machine, Machine System

* Social System is made up of people. For example, social clubs, societies.
* In Human-Machine System, both human and machines are involved to perform a particular task. For example, Computer programming.
* All the tasks are performed by the machine. For example, an autonomous robot.

### Man–Made Information Systems

* It is an interconnected set of information resources to manage data for particular organization, under Direct Management Control (DMC).
* This system includes hardware, software, communication, data, and application for producing information according to the need of an organization.

Man-made information systems are divided into three types −

* Formal Information System − It is based on the flow of information in the form of memos, instructions, etc., from top level to lower levels of management.
* Informal Information System − This is employee-based system which solves the day to day work related problems.
* Computer Based System − This system is directly dependent on the computer for managing business applications. For example, automatic library system, railway reservation system, banking system, etc.

## Systems Models

### Schematic Models

* A schematic model is a 2-D chart that shows system elements and their linkages.
* Different arrows are used to show information flow, material flow, and information feedback.

### Flow System Models

* A flow system model shows the orderly flow of the material, energy, and information that hold the system together.
* Program Evaluation and Review Technique (PERT), for example, is used to abstract a real-world system in model form.

### Static System Models

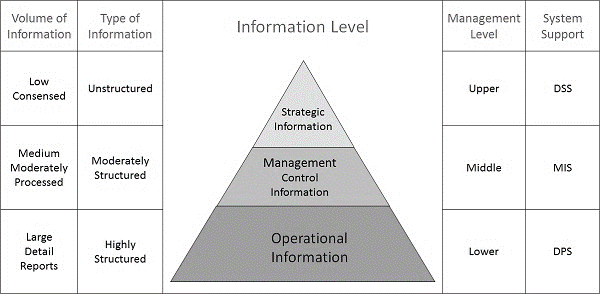
* They represent one pair of relationships such as *activity–time* or *cost–quantity*.
* The Gantt chart, for example, gives a static picture of an activity-time relationship.

### Dynamic System Models

* Business organizations are dynamic systems. A dynamic model approximates the type of organization or application that analysts deal with.
* It shows an ongoing, constantly changing status of the system. It consists of −
  + Inputs that enter the system
  + The processor through which transformation takes place
  + The program(s) required for processing

## Categories of Information

There are three categories of information related to managerial levels and the decision managers make.



### Strategic Information

* This information is required by topmost management for long range planning policies for next few years. For example, trends in revenues, financial investment, and human resources, and population growth.
* This type of information is achieved with the aid of Decision Support System (DSS).

### Managerial Information

* This type of Information is required by middle management for short and intermediate range planning which is in terms of months. For example, sales analysis, cash flow projection, and annual financial statements.
* It is achieved with the aid of Management Information Systems (MIS).

### Operational information

* This type of information is required by low management for daily and short-term planning to enforce day-to-day operational activities. For example, keeping employee attendance records, overdue purchase orders, and current stocks available.
* It is achieved with the aid of Data Processing Systems (DPS).

**2.2 SRS**

A software requirements specification (SRS) is a description of a software system to be developed. It is modelled after business requirements specification (CONOPS), also known as a stakeholder requirements specification (SRS). The software requirements specification lays out functional and non-functional requirements, and it may include a set of use cases that describe user interactions that the software must provide to the user for perfect interaction.

Software requirements specification establishes the basis for an agreement between customers and contractors or suppliers on how the software product should function (in a market-driven project, these roles may be played by the marketing and development divisions). Software requirements specification is a rigorous assessment of requirements before the more specific system design stages, and its goal is to reduce later redesign. It should also provide a realistic basis for estimating product costs, risks, and schedules. Used appropriately, software requirements specifications can help prevent software project failure. [

The software requirements specification document lists sufficient and necessary requirements for the project development. To derive the requirements, the developer needs to have clear and thorough understanding of the products under development. This is achieved through detailed and continuous communications with the project team and customer throughout the software development process.

The SRS may be one of a contract's deliverable data item descriptions or have other forms of organizationally mandated content.

Table of Contents for an SRS Document:

1. Introduction
   1. Purpose
   2. Document Conventions
   3. Intended Audience and Reading Suggestions
   4. Project Scope
   5. References
2. Overall Description
   1. Product Perspective
   2. Product Features
   3. User Classes and Characteristics
   4. Operating Environment
   5. Design and Implementation Constraints
   6. Assumptions and Dependencies
3. System Features
   1. Functional Requirements
4. External Interface Requirements
   1. User Interfaces
   2. Hardware Interfaces
   3. Software Interfaces
   4. Communication Interfaces
5. Non-Functional Requirements
   1. Security Requirements
   2. Software Quality Attributes

SRS Document for the Application Software – REVEIWER:

1. Introduction
   1. Purpose: -

The purpose of this document is to build a Text summariser to provide data summary of intended document in different formats.

* 1. Document Convention: -

For containing the data, the user will upload the file from the local system.

* 1. Intended Audience and Reading Suggestion: -

This project is a prototype for the Text summarisation. This has been implemented under the guidance of college teachers

Project scope: -

The purpose of the SUMMARISER is to ease the filtering of important information of user’s document. The system is based on NLP and semantic analysis of data from different category. Above all, we hope to provide a comfortable user experience by deploying the text summariser using PyQt5.

* 1. References
* https://www.imdb.com/
* https://www.gsmarena.com/
* https://pandas.pydata.org/docs/reference/frame.html
* https://docs.python.org/3.3/library/collections.html
* https://docs.python.org/3/library/urllib.request.html
* Pratiksha Ashiwal, S.R.Tandan, Priyanka Tripathi, Rohit Miri4, “Web Information Retrieval Using Python and BeautifulSoup”.
* Mini Singh Ahuja, Dr Jatinder Singh, Bal Varnica, ”Web Crawler: Extracting the Web Data”.
* “DATA MINING TECHNIQUES AND APPLICATIONS” Mrs. Bharati M. Ramageri, Lecturer Modern Institute of Information Technology and Research
* “Node.js Challenges in Implementation” , By Hezbullah Shah & Tariq Rahim Soomro.
* “Python as a Tool for Web Server Application Development”, Sheetal Taneja, Pratibha R. Gupta.
* “Summary of web crawler technology research”.
* “The Applications of Importance Performance Analysis in Higher Education Institutions: A Text Mining of Literature”,Afifah Hanim Md Pazila, Razli Che Razak.
* “Web 2.0: A Definition, Literature Review, and Directions for Future Research”.

## Overall Description

2.1 Product Perspective

* A distributed review portal stores the following information.
* Details:

It includes the display all the data based on different category and having all parts in it.

2.2 Product Features

* The major features of reviewer are that its interface is user friendly and good looking just like a premium application.
* User can upload, view and save the summarised document no matter what the length is.
* The person does not have to go through the complete document to analyse the important aspects of the report.
* NLP and semantic analysis is used to generate summary.
* Backend is kept in simple format, .txt files are used.

2.3 USER CLASS And CHARACTERISTICS

We have made a class team in which we have created the data member and member functions.

2.4 Operating Environment

* Distributed Database
* Client/Server System
* Operating System: Windows.
* Platform: MIX

2.5 DESIGN and IMPLEMENTATION CONSTRAINTS

* The global schema, fragmentation schema, and allocation schema.
* Implement the database at least using a centralized database management system.

## System Features

## Functional Requirements

Distributed Database: - It implies that a single application should be able to operate transparently on data that is spread across a variety of different databases.

1. External Interface Requirements
   1. User Interfaces

Front-end software: PyQt5

Back-end software: Python

* 1. Hardware Interfaces

Processor : Intel(R) Core(TM) i3 CPU @ 1.4GHz or above

RAM : 2 GB or more

Hard Disk : 70 GB or more

Screen Resolution : 800 x 600 or above

Keyboard : Compatible Keyboard

Mouse : Compatible Mouse

Driver : CD – ROM

* 1. Software interface

Operating system : Windows 7 or above

4.4 Communication Interfaces

The software has online support and can only be used if there is an access to internet.

1. Non-Functional Requirements

5.1 Security Requirements

Security systems need database storage just like many other applications. However, the special requirements of the security market mean that vendors must choose their database partner carefully.

5.2 Software Quality Attributes

* Filtering/recommendation on the basis of work done by actors, genre and top-rated movies.
* Correctness: The user will search for the details of movies then it will display the appropriate suggestions.
* Maintainability: The details of guests or customer will store in well-organized manner.
* Usability: The users can use this for recommendation of movie, tv series and mobile of their personal choice.

SRS END

**2.3 Feasibility Study**

Economic Feasibility – Identifying & Forecasting Costs & Benefits

Costs and benefits of the proposed computer system must always be considered together, because they are interrelated and often interdependent. Although the systems analyst is trying to propose a system that fulfil various information requirements, decisions to continue with the proposed system will be based on a cost-benefit analysis, not on information requirements. In many ways, benefits are measured becomes apparent in the next section.

Forecasting

Systems analysts are required to predict certain key variables before the proposal is submitted to the client. To some degree, a systems analyst will rely on a what-if analysis, such as, “What if labour costs rise only 5 percent per year for the next three years, rather than 10 percent?” The systems analyst should realize, however, that he or she cannot rely on what-if analysis for everything if the proposal is to be credible, meaningful, and valuable.

The systems analyst has many forecasting models available. The main condition for choosing a model is the availability of historical data. If they are unavailable, the analyst must turn to one of the judgment methods: estimates from the sales force, surveys to estimate customer demand, Delphi studies (a consensus forecast developed independently by a group of experts through a series of iterations), creating scenarios, or drawing historical analogies.

If historical data are available, the next differentiation between classes of techniques involves whether the forecast is conditional or unconditional. Conditional implies that there is an association among variables in the model or that such a causal relationship exists. Common methods in this group include correlation, regression, leading indicators, econometrics, and input/output models.

Unconditional forecasting means the analyst isn’t required to find or identify any causal relationships. Consequently, systems analysts find that these methods are low-cost, easy-to-implement alternatives. Included in this group are graphical judgment, moving averages, and analysis of time-series data. Because these methods are simple, reliable, and cost effective, the remainder of the section focuses on them.

### Estimation of Trends

Trends can be estimated in a number of different ways. One way to estimate trends is to use a moving average. This method is useful because some seasonal, cyclical, or random patterns may be smoothed, leaving the trend pattern. The principle behind moving averages is to calculate the arithmetic mean of data from a fixed number of periods; a three-month moving average is simply the average of the last three months. For example, the average sales for January, February, and March are used to predict the sales for April.

Then the average sales for February, March, and April are used to predict the sales for May, and so on.

When the results are graphed, it is easily noticeable that the widely fluctuating data are smoothed. The moving average method is useful for its smoothing ability, but at the same time it has many disadvantages. Moving averages are more strongly affected by extreme values than by using graphical judgment or estimating using other methods such as least squares.

### Intangible Benefits

Some benefits that accrue to the organization from the use of the information system are difficult to measure but are important, nonetheless. They are known as intangible benefits.

Intangible benefits include improving the decision-making process, enhancing accuracy, becoming more competitive in customer service, maintaining a good business image, and increasing job satisfaction for employees by eliminating tedious tasks. As you can judge from the list given, intangible benefits are extremely important and can have far-reaching implications for the business as it relates to people both outside and within the organization.

Although intangible benefits of an information system are important factors that must be considered when deciding whether to proceed with a system, a system built solely for its intangible benefits will not be successful. You must discuss both tangible and intangible benefits in your proposal, because presenting both will allow decision makers in the business to make a well-informed decision about the proposed system.

### Tangible Costs

The concepts of tangible and intangible costs present a conceptual parallel to the tangible and intangible benefits discussed already. Tangible costs are those that can be accurately projected by the systems analyst and the business’s accounting personnel.

Included in tangible costs are the cost of equipment such as computers and terminals, the cost of resources, the cost of systems analysts’ time, the cost of programmers’ time, and other employee salaries. These costs are usually well established or can be discovered quite easily, and are the costs that will require a cash outlay of the business.

### Intangible Costs

Intangible costs are difficult to estimate and may not be known. They include losing a competitive edge, losing the reputation for being first with an innovation or the leader in a field, declining company image due to increased customer dissatisfaction, and ineffective decision making due to untimely or inaccessible information.

**Chapter 3**

**System Design**

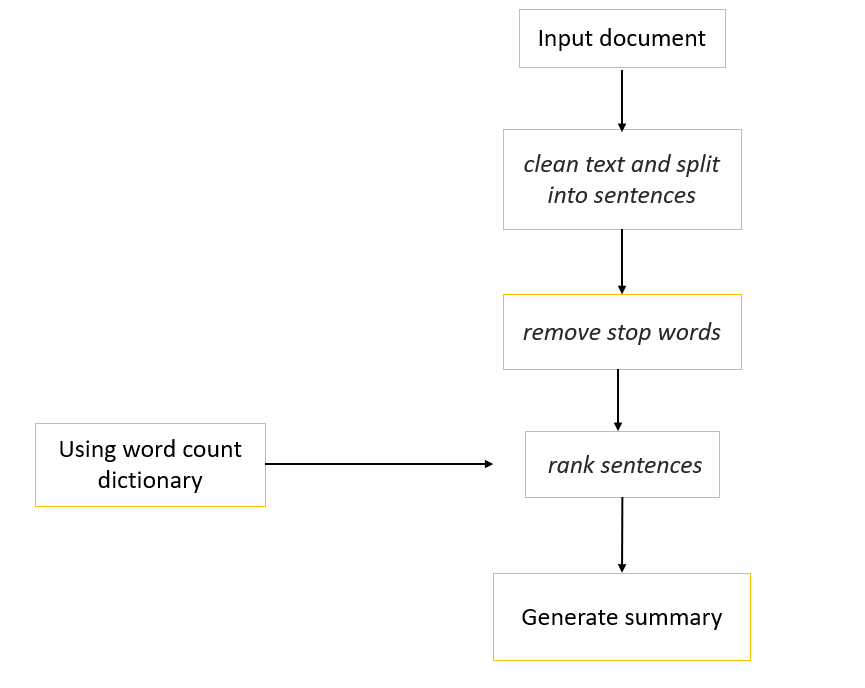
**3.1 Design methods**

Design methods are procedures, techniques, aids, or tools for designing. They offer a number of different kinds of activities that a designer might use within an overall design process. Conventional procedures of design, such as drawing, can be regarded as design methods, but since the 1950s new procedures have been developed that are more usually grouped together under the name of "design methods". What design methods have in common is that they "are attempts to make public the hitherto private thinking of designers; to *externalise* the design process".

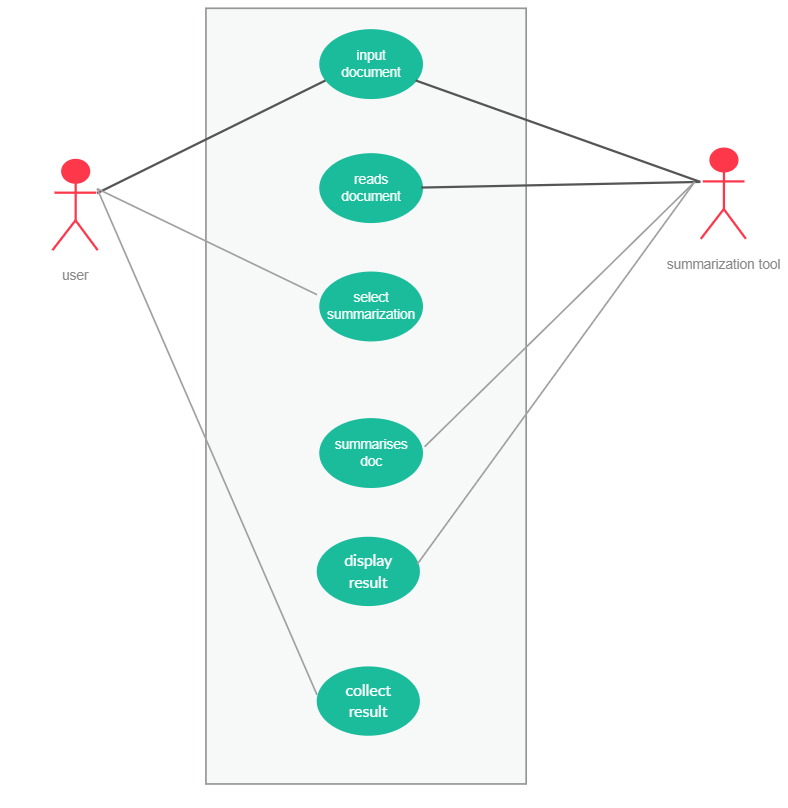
*Design methodology* is the broader study of method in design: the study of the principles, practices and procedures of designing.

Design methods originated in new approaches to problem solving developed in the mid-20th Century, and also in response to industrialisation and mass-production, which changed the nature of designing. A "Conference on Systematic and Intuitive Methods in Engineering, Industrial Design, Architecture and Communications", held in London in 1962is regarded as a key event marking the beginning of what became known within design studies as the "design methods movement", leading to the founding of the Design Research Society and influencing design education and practice.

**Workflow of the Project**



USE CASE FOR TEXT SUMMARISER



**3.2 Various development approaches**

Software Development Approach

Software development is the process of creating new software solutions or modifying existing software solutions. Software can be designed and developed in many different ways. The approach taken to design and develop a software package can vary from the very simple, limited planning approach to a very detailed, formal and structured approach. In this topic you will study four commonly used approaches to software development.

* Structured Approach
* Rapid Application Development Approach
* End-User Approach

Selecting an appropriate development approach.

Why develops software? There are obviously millions of reasons to develop a new program; however, the concept of this topic looks at the method. Which approach and language is best? What happens if you get it wrong?

Structured approach

Definition

“Structured Approach to developing a software application involves the use of the program development cycle:

* Defining the problem
* Planning the solution
* Building the solution
* Checking the solution
* Modifying the solution”.

The stages of the development cycle are:

Defining the problem: - It is necessary to understand the problem in as much detail as possible.

Planning the solution: - It is important to begin the planning and design of a solution to the problem. Planning solution involves determining data types, developing algorithms and diagrams, whether the ‘start from scratch approach' needed, determine the programming language used.

Building the solution: - Building the solution involves converting the specifications and algorithms developed during the planning stage into code.

Checking the solution: - Once the software solution has been coded it is necessary to check that is operates correctly.

Modifying the solution: - Modification of code may be necessary if errors are detected, or the users' feedback means that the program doesn't meet their needs, or the underlying processes change, or the existing system changes, etc. Accurate documentation is extremely important because the modifications may be made by programmer who were not part of the original team.

**3.3 Various development approache**

**3.3.1 PYTHON:**

**3.3.1.1 INTRODUCTION TO PYTHON:**

Python is powerful, fast, plays well with others, runs everywhere, is open, is friendly & easy to learn. Python (Foundation, 2019)is an interpreted, object-oriented, high-level programming language with dynamic semantics. Python's simple, easy to learn syntax emphasizes 31 readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. Python is a general purpose and high level programming language. You can use Python for developing desktop GUI applications, websites and web applications. Also, Python, as a high level programming language, allows you to focus on core functionality of the application by taking care of common programming tasks.

**3.3.1.2 OPEN SOURCE:**

Python is developed under an OSI-approved open source license, making it freely usable and distributable, even for commercial use. Python's license is administered by the Python Software Foundation.

**3.3.2 NLP, Semantic Analysis and Syntactic Analysis:**

**Text is at the heart of how we communicate with everyone.** Each type of communication, whether it’s a tweet, a post on LinkedIn or a review in the comments section of a website, contains potentially relevant, even valuable information that must be captured and understood by companies who want to stay ahead. Capturing the information isn’t the hard part. **What’s really difficult is understanding what is being said, and doing it at scale.**

For humans, the way we understand what’s being said is almost an unconscious process. To understand what a text is talking about, we rely on what we already know about language itself and about the concepts present in a text. Machines can’t rely on these same techniques.

Some technologies only make you think they understand text. An approach based on keywords or statistics, or even pure machine learning, may be using a matching or frequency technique for clues as to what a text is “about.” These methods can only go so far because they are not looking at meaning.

**Semantic analysis describes the process of understanding natural language**–the way that humans communicate–based on meaning and context. Let’s look at **how a cognitive technology like expert.ai performs semantic analysis**.

The **semantic analysis of natural language content starts by reading all of the words in content** to capture the real meaning of any text. It identifies the text elements and assigns them to their logical and grammatical role. It analyzes context in the surrounding text and it analyzes the text structure to accurately disambiguate the proper meaning of words that have more than one definition.

**Chapter 4**

**Testing**

**4.1 Testing techniques and strategies**

To perform testing in a planned and systematic manner, software testing strategy is developed. A testing strategy is used to identify the levels of testing which are to be applied along with the methods, techniques, and tools to be used during testing. This strategy also decides test cases, test specifications, test case decisions, and puts them together for execution.

Developing a test strategy, which efficiently meets the requirements of an organization, is critical to the success of software development in that organization. Therefore, a software testing strategy should contain complete information about the procedure to perform testing and the purpose and requirements of testing.

The choice of software testing strategy is highly dependent on the nature of the developed software. For example, if the software is highly data intensive then a strategy that checks structures and values properly to ensure that all inputs given to the software are correct and complete should be developed. Similarly, if it is transaction intensive then the strategy should be such that it is able to check the flow of all the transactions. The design and architecture of the software are also useful in choosing testing strategy. A number of software testing strategies are developed in the testing process. All these strategies provide the tester a template, which is used for testing. Generally, all testing strategies have following characteristics.

Testing proceeds in an outward manner. It starts from testing the individual units, progresses to integrating these units, and finally, moves to system testing.

Testing techniques used during different phases of software development are different.

Testing is conducted by the software developer and by an ITG.

Testing and debugging should not be used synonymously. However, any testing strategy must accommodate debugging with itself.

Types of Software Testing Strategies

There are different types of software testing strategies, which are selected by the testers depending upon the nature and size of the software. The commonly used software testing strategies are listed below.

Analytic testing strategy: This uses formal and informal techniques to access and prioritize risks that arise during software testing. It takes a complete overview of requirements, design, and implementation of objects to determine the motive of testing. In addition, it gathers complete information about the software, targets to be achieved, and the data required for testing the software.

Model-based testing strategy: This strategy tests the functionality of the software according to the real-world scenario (like software functioning in an organization). It recognizes the domain of data and selects suitable test cases according to the probability of errors in that domain.

Methodical testing strategy: It tests the functions and status of software according to the checklist, which is based on user requirements. This strategy is also used to test the functionality, reliability, usability, and performance of the software.

Process-oriented testing strategy: It tests the software according to already existing standards such as the IEEE standards. In addition, it checks the functionality of the software by using automated testing tools.

Dynamic testing strategy: This tests the software after having a collective decision of the testing team. Along with testing, this strategy provides information about the software such as test cases used for testing the errors present in it.

Philosophical testing strategy: It tests the software assuming that any component of the software can stop functioning anytime. It takes help from software developers, users and systems analysts to test the software.

**4.2 Debugging & Improvement**

Debugging, in computer programming and engineering, is a multistep process that involves identifying a problem, isolating the source of the problem, and then either correcting the problem or determining a way to work around it. The final step of debugging is to test the correction or workaround and make sure it works.

The debugging process: -

In software development, debugging involves locating and correcting code errors in a computer program. Debugging is part of the software testing process and is an integral part of the entire software development lifecycle. The debugging process starts as soon as code is written and continues in successive stages as code is combined with other units of programming to form a software product. In a large program that has thousands and thousands of lines of code, the debugging process can be made easier by using strategies such as unit tests, code reviews and pair programming.

Once an error has been identified, it is necessary to actually find the error in the code. At this point, it can be useful to look at the code's logging and use a stand-alone debugger tool or the debugging component of an integrated development environment (IDE). Invariably, the bugs in the functions that get most use are found and fixed first. In some cases, the module that presents the problem is obvious, while the line of code itself is not. In that case, unit tests – such as JUnit and xUnit, which allow the programmer to run a specific function with specific inputs -- can be helpful in debugging.

The standard practice is to set up a "breakpoint" and run the program until that breakpoint, at which time program execution stops. The debugging component of an IDE typically provides the programmer with the capability to view memory and see variables, run the program to the next breakpoint, execute just the next line of code, and, in some cases, change the value of variables or even change the contents of the line of code about to be executed.

Common debugging tools

Source code analysers, which include security, common code errors and complexity analysers, can also be helpful in debugging. A complexity analyser can find modules that are so intricate as to be hard to understand and test. Some tools can actually analyse a test run to see what lines of code were not executed, which can aid in debugging. Other debugging tools include advanced logging and simulators that allow the programmer to model how an app on a mobile device will display and behave.

*Finding and removing software errors.*

Some tools, especially open-source tools and scripting languages, do not run in an IDE and require a more manual approach to debugging. Such techniques include dropping values to a log, extensive "print" statements added during code execution or hard-coded "wait" commands that simulate a breakpoint by waiting for keyboard input at specific times.

The use of the word *bug* as a synonym for *error* originated in engineering. The term's application to computing and the inspiration for using the word *debugging* as a synonym for *Troubleshooting* has been attributed to Admiral Grace Hopper, a pioneer in computer programming, who was also known for her dry sense of humour. When an actual bug (a moth) got caught between electrical relays and caused a problem in the U.S. Navy's first computer, Admiral Hopper and her team "debugged" the computer and saved the moth. It now resides in the Smithsonian Museum.

**4.3 Code**

Following is the code for reading the file in .pdf or .docx or .txt format:

def read\_content(self, path):

        try:

            if path.endswith('.docx'):

                doc\_text = doctext.DocFile(doc=path)

                text = doc\_text.get\_text()

                return text

            elif path.endswith('.pdf'):

                text=self.pdf\_content\_reader(path)

                print("in pdf")

                return text

            elif path.endswith('.txt'):

                with open(path,'r',encoding="mbcs") as f:

                    return f.read()

            else:

                return "FILE CANNOT OPEN!!!\nINVALID FORMAT FILE... \nPLEASE OPEN A WORD, PDF OR A TXT FILE."

        except:

            pass

    def pdf\_content\_reader(self,path):

        # creating a pdf file object

        pdfFileObj = open(path, 'rb')

        # creating a pdf reader object

        pdfReader = PyPDF2.PdfFileReader(pdfFileObj)

        # creating a page object

        text=""

        for i in range(pdfReader.numPages):

            pageObj = pdfReader.getPage(i)

            text=text+" " +pageObj.extractText()

        # extracting text from page

        #print(pageObj.extractText())

        # closing the pdf file object

        pdfFileObj.close()

        return text

Following is the logic behind the whole text summariser:

    def push\_button\_Submit(self):

        sent\_word\_length = '30'

        path = self.file\_name[0]

        file = self.read\_content(path)

        text = file

        text = re.sub(r'\[[0-9]\*\]',' ',text)

        text = re.sub(r'\s+',' ',text)

        clean\_text = text.lower()

        regex\_patterns = [r'\W',r'\d',r'\s+']

        for regex in regex\_patterns:

            clean\_text = re.sub(regex,' ',clean\_text)

        sentences = nltk.sent\_tokenize(text)

        top\_n= math.sqrt(len(sentences))

        stop\_words = nltk.corpus.stopwords.words('english')

        word\_count = {}

        for word in nltk.word\_tokenize(clean\_text):

            if word not in stop\_words:

                if word not in word\_count.keys():

                    word\_count[word] = 1

                else:

                    word\_count[word] += 1

        sentence\_score = {}

        for sentence in sentences:

            for word in nltk.word\_tokenize(sentence.lower()):

              if word in word\_count.keys():

                if len(sentence.split(' ')) < int(sent\_word\_length):

                     if sentence not in sentence\_score.keys():

                         sentence\_score[sentence] = word\_count[word]

                      else:

                        sentence\_score[sentence] += word\_count[word]

        best\_sentences = heapq.nlargest(int(top\_n), sentence\_score, key=sentence\_score.get)

        summarized\_text = []

        sentences = nltk.sent\_tokenize(text)

        for sentence in sentences:

            if sentence in best\_sentences:

                summarized\_text.append(sentence)

        summarized\_text = "\n".join(summarized\_text)

        self.lblSubmit.setText(summarized\_text)

        savepath=self.path.split('/')

        savepath=savepath[:-1]

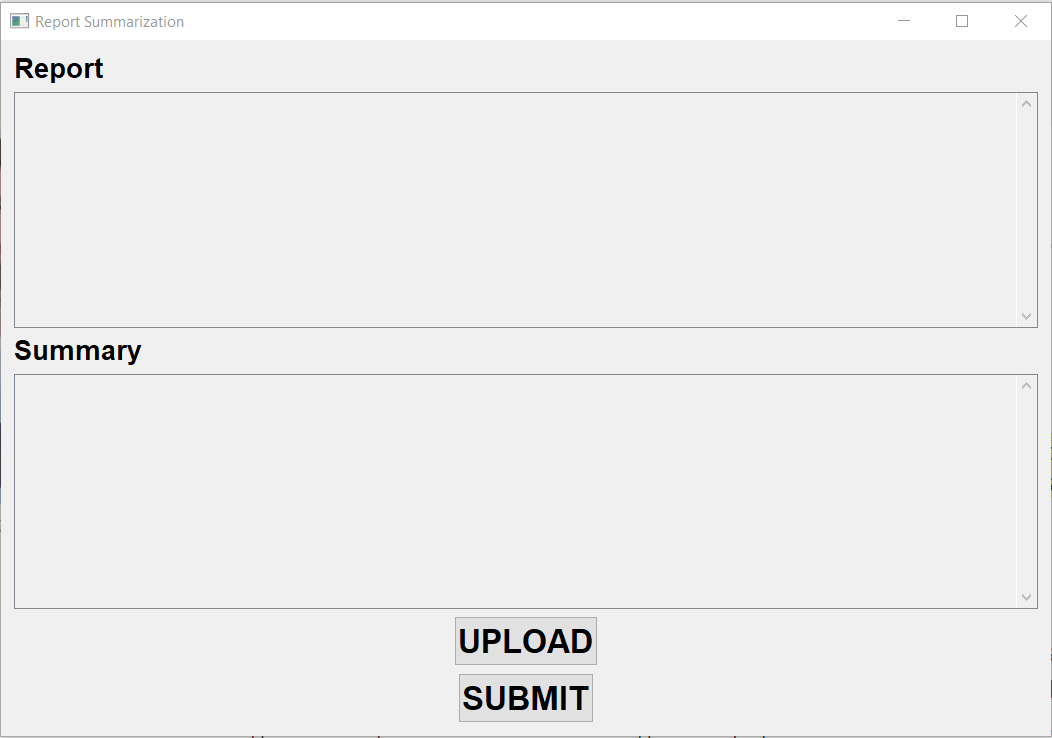
        savepath='/'.join(savepath)+'/'+'summary.txt'

        text\_file = open(savepath, "w")

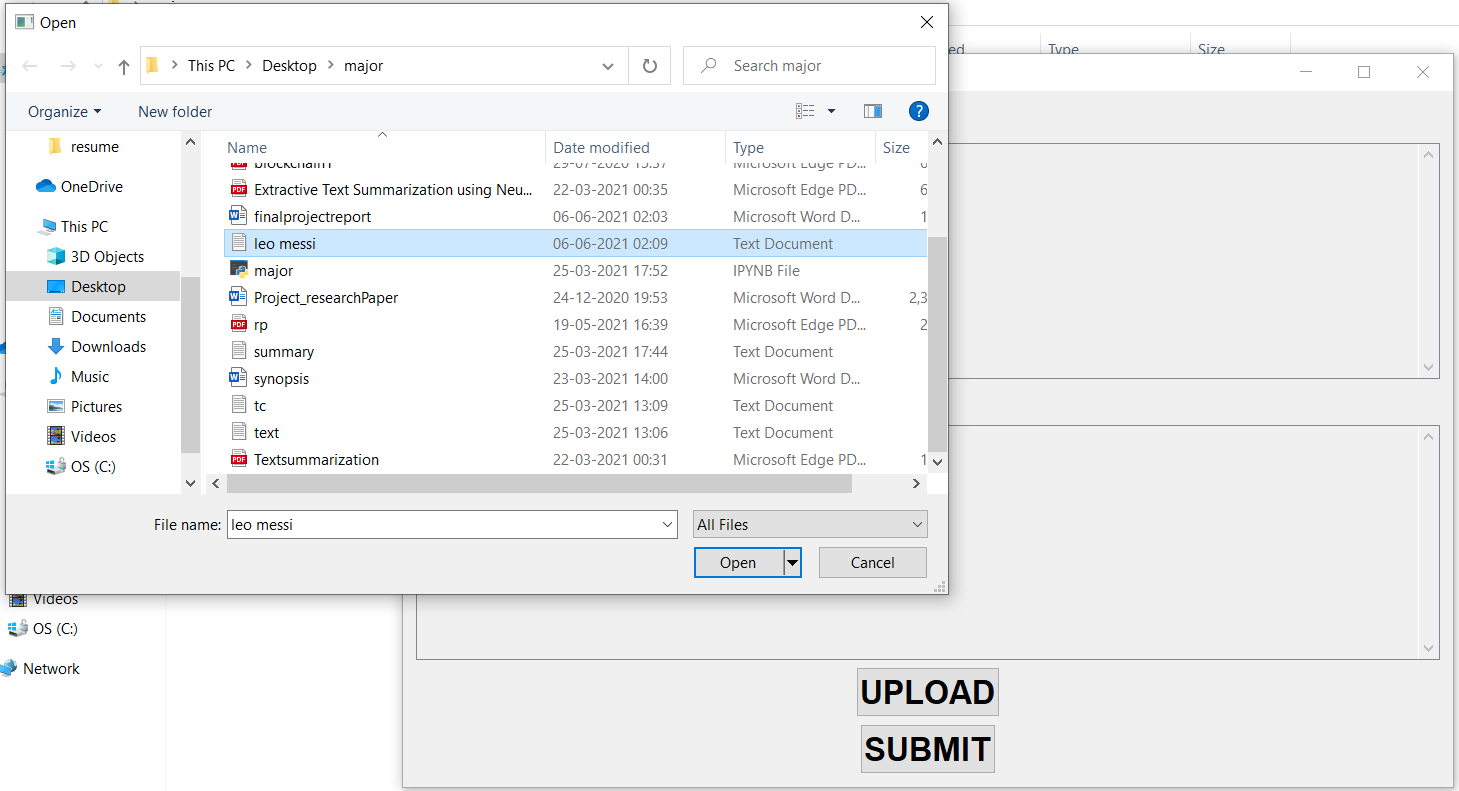
        text\_file.write(summarized\_text)

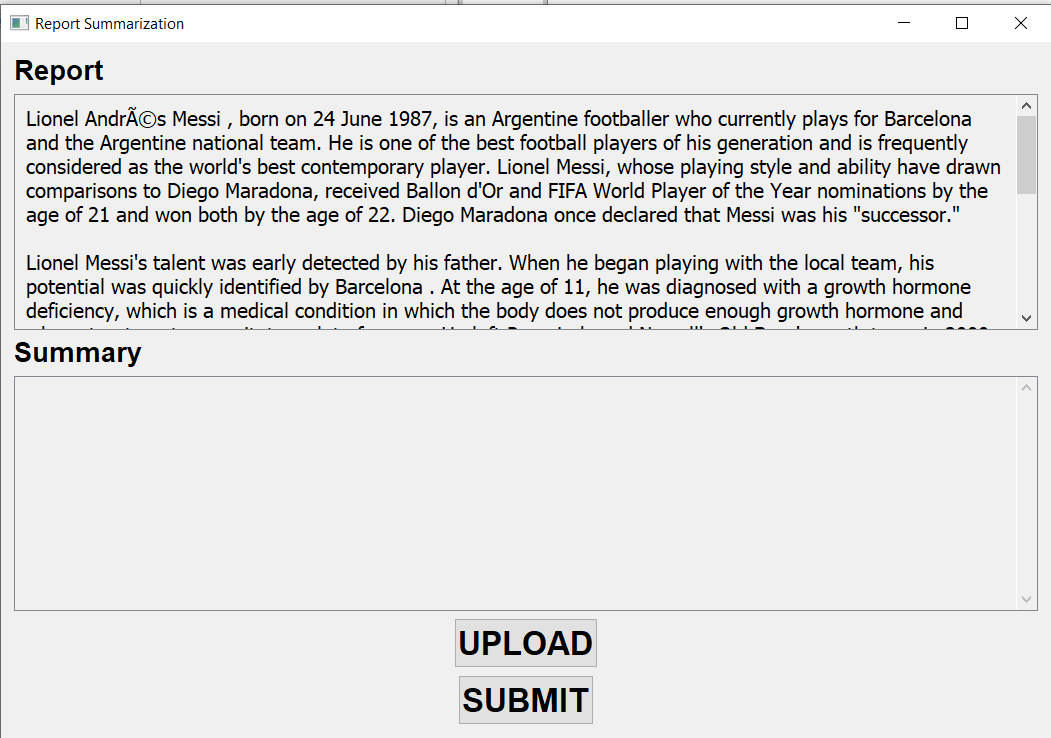
        text\_file.close()

**4.4 Outputs**

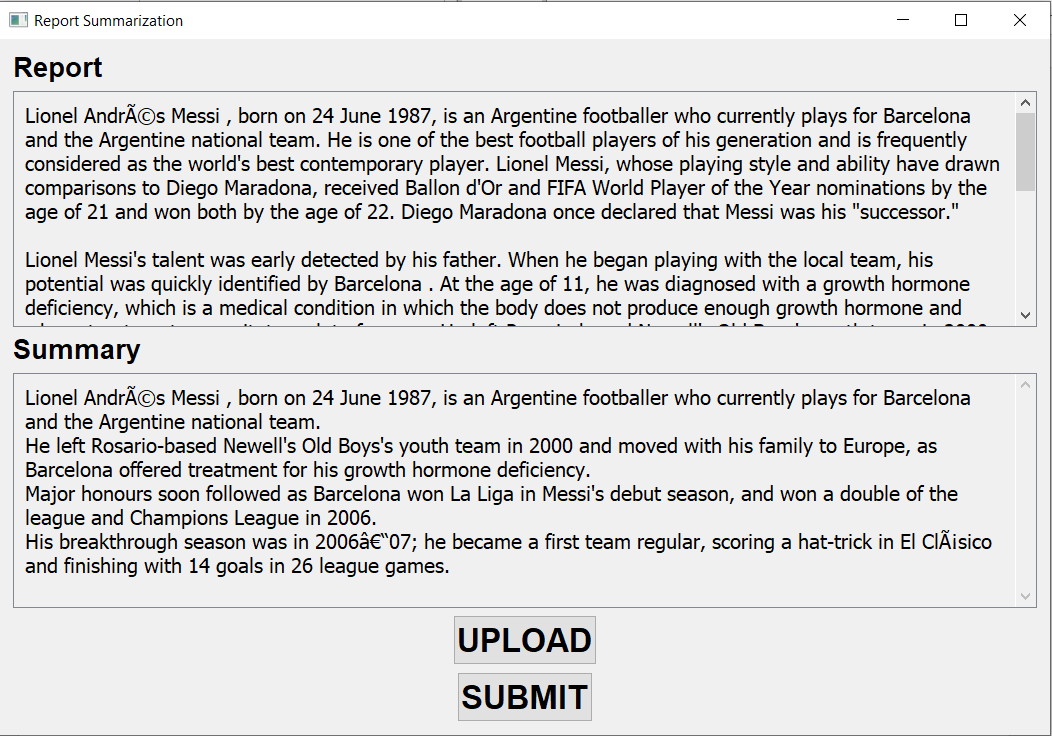
**User interface**

**After clicking the upload button**

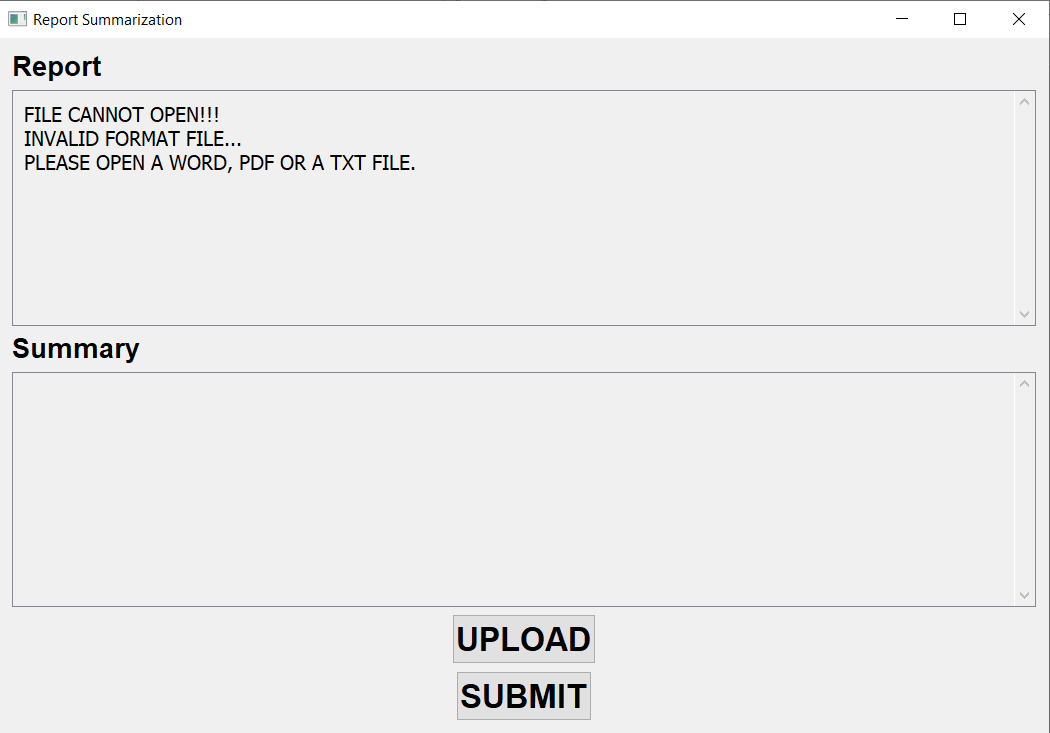




**After clicking the submit button**



**If a wrong format file is selected:**



**Chapter 5**

**Implementation**

**5.1 System Implementation**

In pursuit of our number one project goal – to achieve exceptional results for our customers – we follow a standard system implementation framework, to ensure that the Lucas system is configured and implemented quickly and cleanly. The Lucas project management approach and implementation process have been developed and refined over the course of hundreds of projects.

First, we work with you to determine the exact scope and plan for the project. Next, we configure the system to meet the needs that we jointly identify and test it to make sure that it will. Finally, we install the system and train your team to guarantee a successful go-live. While we have completed projects in as few as 30 days, a typical implementation timeline is between 60 and 90 days, depending on the complexity of the system and the availability of the customer team.

## Project Initiation and System Design

Following operational analysis and process design, and upon agreement to proceed with the system implementation, Lucas will assemble a project team and begin the system design process. During initial project kick-off meetings we will discuss project communications, and provide an initial project plan and milestone schedule. We will also begin to discuss on-site installation and training plans.

The Lucas project team will work with your team to create detailed functional requirements for your mobile work execution system, including the mobile user workflows, management reporting screens, interfaces, business logic, and any customer-specific functionality that will be incorporated in the system.

Upon completion of the functional requirements phase we will provide an updated process flow document and final interface specification. Our engineering team will then complete a detailed system design, which will determine the specific business rules and configuration options that will be incorporated into your system, and develop test plans and initial system documentation.

## Configuration and Testing

Following detailed design, the Lucas team will assemble your system using our pre-built functional blocks that are designed to accommodate customer-specific business rules and configurations without custom development.

As the functional blocks are configured and assembled, Lucas will unit test the individual modules. We will develop integration test plans in consultation with your team, and we will jointly define a final installation and start up plan. The customer project team will continue to prepare the site for installation as the Lucas system is transferred to the Lucas QA department for functional and performance testing. Upon completion of remote testing, the system is ready for installation and roll out.

**5.2 Software implementation**

## Structured Programming

In the process of coding, the lines of code keep multiplying, thus, size of the software increases. Gradually, it becomes next to impossible to remember the flow of program. If one forgets how software and its underlying programs, files, procedures are constructed it then becomes very difficult to share, debug and modify the program. The solution to this is structured programming. It encourages the developer to use subroutines and loops instead of using simple jumps in the code, thereby bringing clarity in the code and improving its efficiency Structured programming also helps programmer to reduce coding time and organize code properly.

Structured programming states how the program shall be coded. Structured programming uses three main concepts:

* Top-down analysis - A software is always made to perform some rational work. This rational work is known as problem in the software parlance. Thus it is very important that we understand how to solve the problem. Under top-down analysis, the problem is broken down into small pieces where each one has some significance. Each problem is individually solved and steps are clearly stated about how to solve the problem.
* Modular Programming - While programming, the code is broken down into smaller group of instructions. These groups are known as modules, subprograms or subroutines. Modular programming based on the understanding of top-down analysis. It discourages jumps using ‘goto’ statements in the program, which often makes the program flow non-traceable. Jumps are prohibited and modular format is encouraged in structured programming.
* Structured Coding - In reference with top-down analysis, structured coding sub-divides the modules into further smaller units of code in the order of their execution. Structured programming uses control structure, which controls the flow of the program, whereas structured coding us
* control structure to organize its instructions in definable patterns.

## Functional Programming

Functional programming is style of programming language, which uses the concepts of mathematical functions. A function in mathematics should always produce the same result on receiving the same argument. In procedural languages, the flow of the program runs through procedures, i.e. the control of program is transferred to the called procedure. While control flow is transferring from one procedure to another, the program changes its state. In procedural programming, it is possible for a procedure to produce different results when it is called with the same argument, as the program itself can be in different state while calling it. This is a property as well as a drawback of procedural programming, in which the sequence or timing of the procedure execution becomes important.

Functional programming provides means of computation as mathematical functions, which produces results irrespective of program state. This makes it possible to predict the behaviour of the program.

## Programming style

Programming style is set of coding rules followed by all the programmers to write the code. When multiple programmers work on the same software project, they frequently need to work with the program code written by some other developer. This becomes tedious or at times impossible, if all developers do not follow some standard programming style to code the program.

An appropriate programming style includes using function and variable names relevant to the intended task, using well-placed indentation, commenting code for the convenience of reader and overall presentation of code. This makes the program code readable and understandable by all, which in turn makes debugging and error solving easier. Also, proper coding style helps ease the documentation and updating.

**5.2.1 Current uses**

The current uses of this embodies the apparent use and approaches of the metrological department.

**5.2.2 Ongoing research and development**

Automatic text summarization is an old challenge but the current research direction diverts towards emerging trends in biomedicine, product review, education domains, emails and blogs. This is due to the fact that there is information overload in these areas, especially on the World Wide Web.Automated summarization is an important area in NLP (Natural Language Processing) research. It consists of automatically creating a summary of one or more texts. The purpose of extractive document summarization is to automatically select a number of indicative sentences, passages, or paragraphs from the original document .

**CONCLUSION AND FUTURE USES**

Text summarization is one of the major problems in the field of Natural Language Processing, and yet it is even after years of research and implementations, fraught with complications. However, there have been some major breakthroughs in the past. Many different methods were used to arrive at the final summary, whether that summary was abstractive or extractive. Methods such as Deep Understanding, Sentence Extraction, Paragraph Extraction, Machine Learning, and even some which employ all these methods along with Traditional NLP Techniques(Semantic Analysis, etc.). As such, keeping these accomplishments in mind, there is still ample amount of research left in the domain of Text Summarization, as a meaningful summary is still difficult to attain in all domains and languages.

**REFRENCES**

1. https://www.imdb.com/
2. https://www.gsmarena.com/
3. https://gadgets.ndtv.com/
4. https://play.google.com/store/apps
5. https://getbootstrap.com/
6. https://pandas.pydata.org/docs/reference/frame.html
7. https://docs.python.org/3.3/library/collections.html
8. https://docs.python.org/3/library/urllib.request.html
9. https://nodejs.org/en/
10. https://www.selenium.dev/
11. Pratiksha Ashiwal, S.R.Tandan, Priyanka Tripathi, Rohit Miri4, “Web Information Retrieval Using Python and BeautifulSoup” International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 4 Issue VI, June 2016 IC Value: 13.98 ISSN: 2321-9653.
12. Mini Singh Ahuja, Dr Jatinder Singh, Bal Varnica, ”Web Crawler: Extracting the Web Data” International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – Jul 2014.
13. Mrs. Bharati M. Ramageri, Lecturer Modern Institute of Information Technology and Research “DATA MINING TECHNIQUES AND APPLICATIONS” Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305.
14. “Node.js Challenges in Implementation” , By Hezbullah Shah & Tariq Rahim Soomro Global Journal of Computer Science and Technology: E-Network, Web & Security Volume 17 Issue 2 Version 1.0 Year June 2017.
15. Sheetal Taneja, Pratibha R. Gupta “Python as a Tool for Web Server Application Development”. JIMS 8i-International Journal of Information, Communication and Computing Technology(IJICCT).
16. “Summary of web crawler technology research”. Linxuan Yu et al 2020 J. Phys.: Conf. Ser. 1449 012036.
17. Afifah Hanim Md Pazila, Razli Che Razak “The Applications of Importance Performance Analysis in Higher Education Institutions: A Text Mining of Literature”. International Journal of Innovation, Creativity and Change.Volume 10, Issue 9, 2019
18. “Web 2.0: A Definition, Literature Review, and Directions for Future Research”.