A

Mini Project

On

AUTOMATIC KEYWORD EXTRACTION FOR TEXT SUMMARIZATION

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

BANDI SRAVAN KUMAR REDDY (197R1A05P7)

K. GANGA REDDY (197R1A05L8)

P. VIKAS (197R1A05P0)

Under the Guidance of

G.POORNIMA

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi) Recognized Under Section 2(f) & 12(B) of the UGCAct.1956, Kandlakoya (V), Medchal Road, Hyderabad-501401.

2019-2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "AUTOMATIC KEYWORD EXTRACTION FOR TEXT SUMMARIZATION" being submitted by BANDI.SRAVAN KUMAR REDDY (197R1A05P7), KATIPALLY.GANGA REDDY (197R1A05L8) & POILY.VIKAS (197R1A05P0) in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by them under our guidance and supervision during the year 2022-23

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

G.Poornima
(Assistant Professor)
INTERNAL GUIDE

Dr. A. Raji Reddy
DIRECTOR

Dr. K. Srujan Raju
HOD

EXTERNAL EXAMINER

Submitted for viva voice Examination held on _____

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.

We take this opportunity to express my profound gratitude and deep regard to my guide **G.Poornima,** Associate Professor for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to the Project Review Committee (PRC) **Dr. Punyaban Patel, Ms. Shilpa, Dr. M . Subha Mastan Rao & J. Narasimharao** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju,** Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We also express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion ofthis project.

BANDI SRAVAN KUMAR REDDY (197R1A05P7)
KATIPALLY GANGA REDDY (197R1A05L8)
POILY VIKAS (197R1A05P0)

ABSTRACT

There is a need for an automated system that can extract only relevant information from these data sources. To achieve this, one needs to mine the text from the documents. Text mining is the process of extracting large quantities of text to derive high-quality information. Text mining deploys some of the techniques of natural language processing (NLP) such as parts-of-speech (POS) tagging, parsing, N-grams, tokenization, etc., to perform the text analysis. Due to the excessiveness of data, there is a need for an automatic summarizer which will be capable of summarizing the data especially textual data in the original document without losing any critical purposes. Text summarization has emerged as an important research area in the recent past. In this regard, review of existing work on the text summarization process is useful for carrying out further research.

In recent times, data is growing rapidly in every domain such as news, social media, banking, education, etc. In this paper, recent literature on automatic keyword extraction and text summarization are presented since the text summarization process is highly dependent on keyword extraction. This literature includes the discussion about different methodology used for keyword extraction and text summarization. It also discusses different databases used for text summarization in several domains along with evaluation matrices. Finally, it discusses briefly about issues and research challenges faced by researchers along with future direction.

LIST OF FIGURES/TABLES

LIST OF SCREENSHOTS

TABLE OF CONTENTS

TABLE OF CONTENTS

1. INTRODUCTION

1. INTRODUCTION

1.1 PROJECT SCOPE

Due to the excessiveness of data, there is a need for an automatic summarizer which will be capable of summarizing the data especially textual data in the original document without losing any critical purposes.

In recent times, data is growing rapidly in every domain such as news, social media, banking, education, etc. Due to the excessiveness of data, there is a need of automatic summarizer which will be capable to summarize the data especially textual data in original document without losing any critical purposes. Text summarization is emerged as an important research area in recent past.

1.2 PROJECT PURPOSE

Summarization is the process of reducing a text document to create a summary that retains the most important points of the original document. Extractive summarizers work on the given text to extract sentences that best convey the message hidden in the text. Most extractive summarization techniques revolve around the concept of finding keywords and extracting sentences that have more keywords than the rest. Keyword extraction usually is done by extracting relevant words having a higher frequency than others, with stress on important ones'. Manual extraction or annotation of keywords is a tedious process brimming with errors involving lots of manual effort and time. In this paper, we proposed an algorithm to extract keyword automatically for text summarization.

1.3 PROJECT FEATURES

This project contains the literature review of recent work in text summarization from the point of views of automatic keyword extraction, text databases, summarization process, summarization methodologies and evaluation matrices. Some important research issues in the area of text summarization are also highlighted in the project.

2. SYSTEM ANALYSIS

2. SYSTEM ANALYSIS

SYSTEM ANALYSIS

System Analysis is the important phase in the system development process. The System is studied to the minute details and analyzed. The system analyst plays an important role of an interrogator and dwells deep into the working of the present system. In analysis, a detailed study of these operations performed by the system and their relationships within and outside the system is done. A key question considered here is, “what must be done to solve the problem?” The system is viewed as a whole and the inputs to the system are identified. Once analysis is completed the analyst has a firm understanding of what is to be done.

2.1 PROBLEM DEFINITION

A general statement of face recognition problem can be formulated as the given still or video images of a scene, identify or verify one or more persons in the scene or in any live capturing devices using a stored database of those authorised faces.

.

2.2 EXISTING SYSTEM

Methods that automatically extract keywords from the documents use heuristics to select the most used and significant words or phrases from the text document. There are many other methods like Statistical methods, Graph-based methods, Graph ranking, Top score word selection etc… But they are not efficient and complex to implement.

2.2.1 DISADVANTAGES OF EXISTING SYSTEM

Following are the disadvantages of existing system:

- Certain sentences that contribute to the summary might be omitted which in return might affect the generated summary.
- Neural Network-based models require large resources and time to train. The results might not exactly meet the required standards or the level of manual text summarization.
- Abstractive methods rewrite certain portions of sentences to generate the summary. There is a chance that these sentences might contain grammatical errors affecting the overall readability.

2.3 PROPOSED SYSTEM

This project contains the literature review of recent work in text summarization from the point of views of automatic keyword extraction, text databases, summarization process, summarization methodologies and evaluation matrices. Some important research issues in the area of text summarization are also highlighted in the paper.

The input text is processed using natural language processing and processed input is converted into vector form using word embedding. Word embedding is the collective name for a set of language modelling and feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers. Sentence ranking is done between sentences to extract higher ranked sentence, which forms the extractive summary of the input. The summarized text is then analysed using polarity and subjectivity parameters. The summarized text is also subjected to speech conversion.

2.3.1 ADVANTAGES OF THE PROPOSED SYSTEM

- Computers are noticeably faster than humans and are capable of generating summaries faster.
- Automatic text summarization can be scaled to different languages with the adoption of a proper algorithm whereas humans are limited by the extent of their expertise in a particular language.
- Automatic text summarization can be used in different fields as discussed in the overview, thereby enhancing the user's experience while engaging with a product or a service.

2.4 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and a business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. Three key considerations involved in the feasibility analysis:

- Economic Feasibility
- Technical Feasibility
- Social Feasibility

2.4.1 ECONOMIC FEASIBILITY

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on a project, which will give best, return at the earliest. One of the factors, which affect the development of a new system, is the cost it would require.

The following are some of the important financial questions asked during preliminary investigation:

- The costs conduct a full system investigation.
- The cost of the hardware and software.
- The benefits in the form of reduced costs or fewer costly errors.

Since the system is developed as part of project work, there is no manual cost to spend for the proposed system. Also all the resources are already available, it give an indication that the system is economically possible for development.

2.4.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

2.4.3 BEHAVIORAL FEASIBILITY

This includes the following questions:

- Is there sufficient support for the users?
- Will the proposed system cause harm?

The project would be beneficial because it satisfies the objectives when developed and installed. All behavioral aspects are considered carefully and conclude that the project is behaviorally feasible

2.5 HARDWARE & SOFTWARE REQUIREMENTS

2.5.1 HARDWARE REQUIREMENTS:

Hardware interfaces specify the logical characteristics of each interface between the software product and the hardware components of the system. The following are some hardware requirements.

- Processor : Pentium IV and above
- Hard disk : 512MB and above
- RAM : 256MB and above

2.5.2 SOFTWARE REQUIREMENTS:

Software Requirements specifies the logical characteristics of each interface and software components of the system. The following are some software requirements,

- Operating system : Windows 8 and Above
- Languages : Python
- Tools : Python IDEL3.7 version, Anaconda - Jupyter, Pycharm,Flask

3. ARCHITECTURE

3.ARCHITECTURE

3.1 PROJECT ARCHITECTURE

This project architecture shows the procedure followed for classification, starting from input to final prediction.

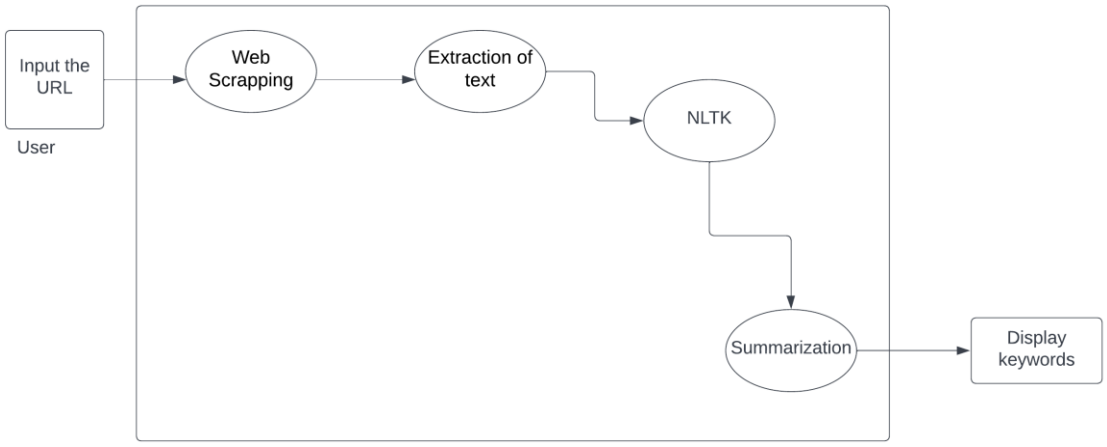


Figure 3.1: Project Architecture of Automatic keyword extraction for text summarization

3.2 DESCRIPTION

Text summarization takes care of choosing the most significant portions of text and generates coherent summaries that express the main intent of the given document. The services offered by our text summarizer is summarizing web articles . Our system does not ask for user details. It provides a platform to get summary without creating an account.

3.3 USE CASE DIAGRAM

In the use case diagram, we have basically one actor who is the user in the trained model.

A use case diagram is a graphical depiction of a user's possible interactions with a system. A use case diagram shows various use cases and different types of users the system has. The use cases are represented by either circles or ellipses. The actors are often shown as stick figures.

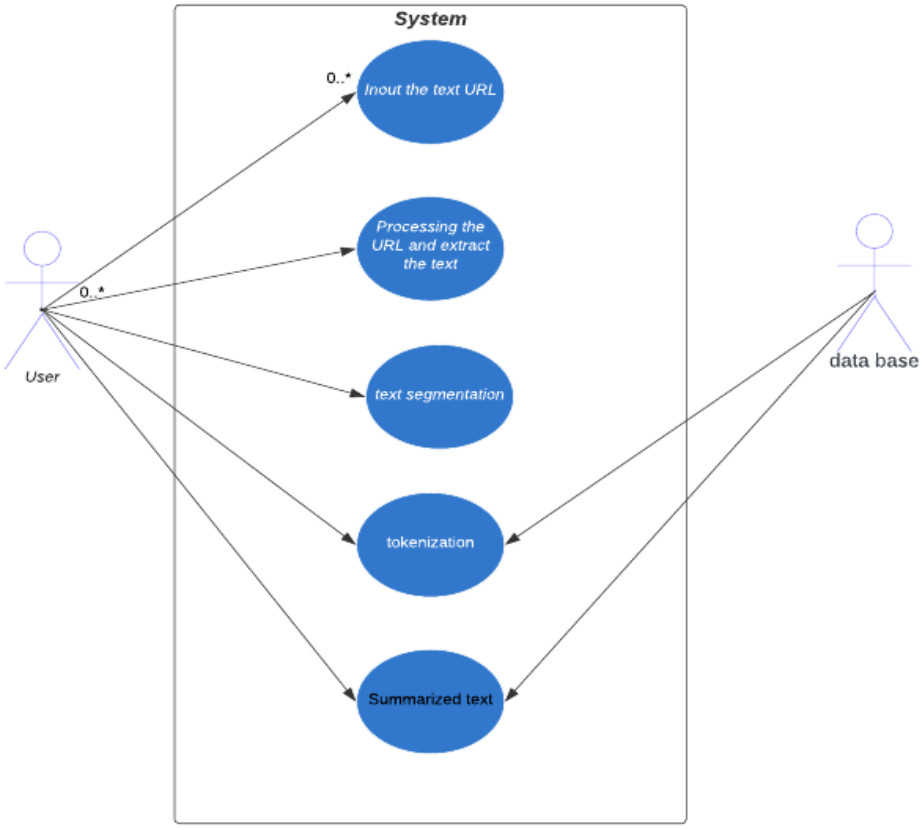


Figure 3.2: Use Case Diagram for Automatic Keyword Extraction for Text Summarization

3.4 CLASS DIAGRAM

Class diagram is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations(or methods), and the relationships among objects.

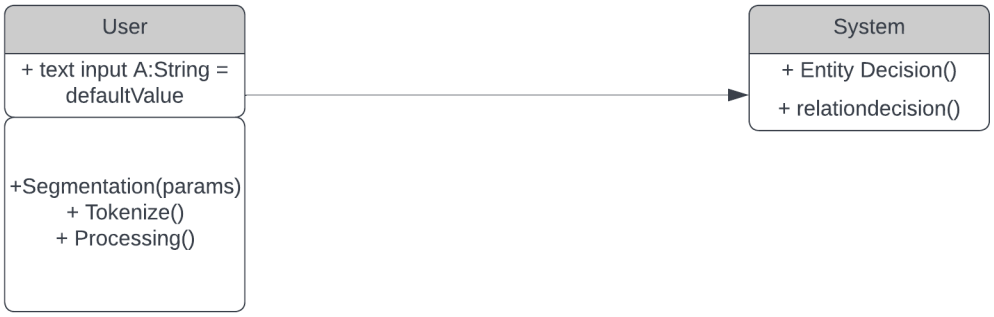


Figure 3.3: Class Diagram for Automatic Keyword Extraction for Text Summarization

3.5 SEQUENCE DIAGRAM

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the logical view of the system under development.

Figure 3.4: Sequence Diagram for Automatic Keyword Extraction for Text Summarization

3.6 ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. They can also include elements showing the flow of data between activities through one or more data stores.

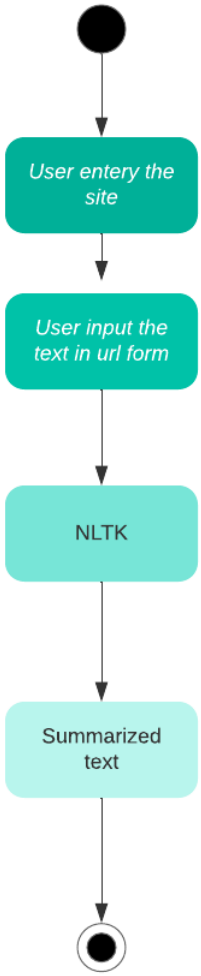


Figure 3.5: Activity Diagram for Automatic Keyword Extraction for Text Summarization

4.IMPLEMENTATION

4.1 SAMPLE CODE

App.py

```
from flask import Flask,
redirect, request
# For flashing a message
from flask import flash
# For rendering html
templates
from flask import
render_template
# For linking files,
generates a URL
from flask import url_for
# Import our forms
from forms import
LinkForm
from summarize import
sum_it_up
# intatiates flask app,
configures the application
app = Flask(__name__)
app.config['SECRET_KE
Y'] = 'This is a secret'
@app.route("/summary",
methods=["GET",
"POST"])
def summary():
form = LinkForm()
# u =
request.args.get("url")
```

```
u = form.link.data
text, keywords =
sum_it_up(u)
if
form.validate_on_submit()
:
flash(f'Summarized!',
'success')
return
render_template('summary
.html', form=form,
text=text, u=u,
keywords=keywords)
@app.route("/",
methods=['GET', 'POST'])
def main():
form = LinkForm()
if
form.validate_on_submit()
:
flash(f'Correct input!',
'success')
return
redirect(url_for('main'))
return
render_template('index.ht
ml', form=form)
if __name__ ==
'__main__':
app.run()
```

Summarize.py

```
# for scraping the webpage
from newspaper import
Article
```

```
# for removing square
brackets
import re
# Gensim summarizer
from
gensim.summarization.su
mmarizer import
summarize
# For extracting the
keywords
from
gensim.summarization
import keywords
# downloads the article
and parses the html, uses
lxml parser
def
downloadwebpage(url):
# downloads the whole
webpage
article = Article(url)
article.download()
# parses the downloaded
html
article.parse()
text = article.text
return text
def sum_it_up(url):
# url =
'https://en.wikipedia.org/w
iki/Elon_Musk'
content =
downloadwebpage(url)
# remove the reference
numbers
```

```
re.sub(r'\[.+\]', '', content)
# finds a list of 10
important keywords, usses
lemmetatization instead of
stemming
k = keywords(content,
words=10,
lemmatize=True).split('\n')
kwords = ', '.join(k)
# computes summary and
reduces size by 20%
return(summarize(content,
0.2), kwords)
```

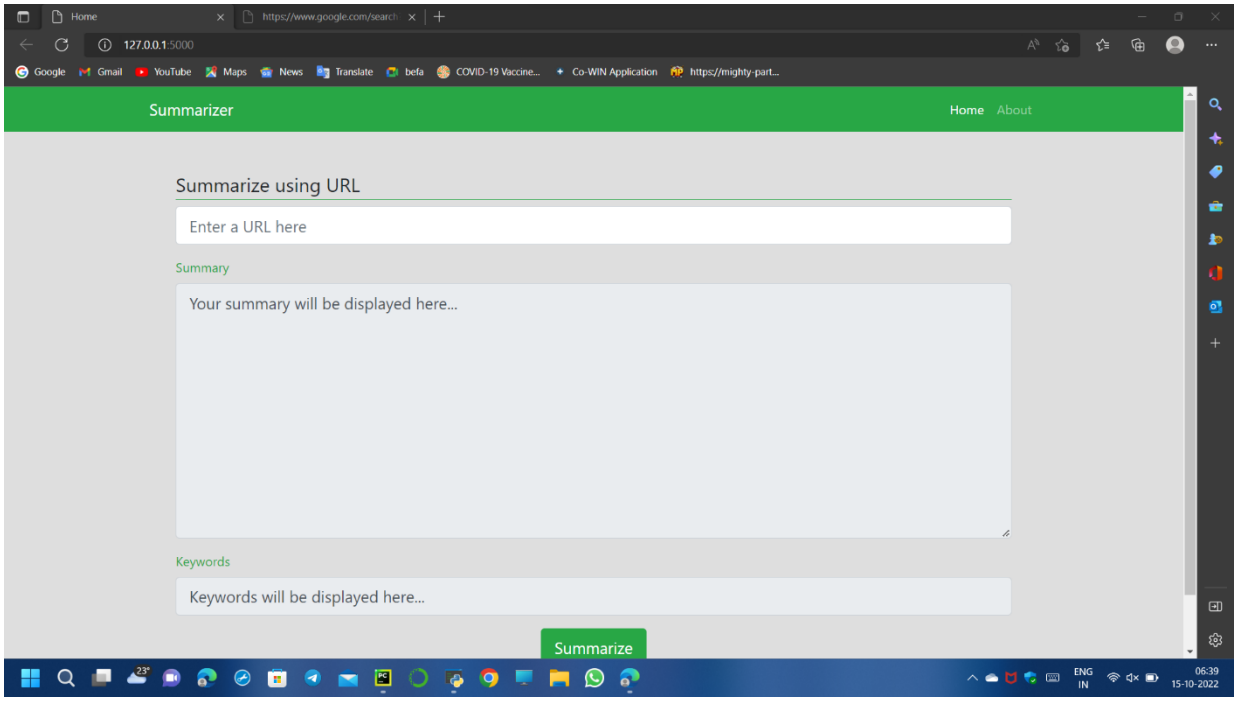
Index.html

```
{% extends "layout.html"
%}
{% block title %}
Home
{% endblock title %}
{% block main %}
<form method="POST"
action="/summary">
<!-- Adds a CSRF token,
sets secret-key for our app
-->
{{ form.hidden_tag() }}
<fieldset class="form-
group">
<legend class="border-
bottom border-
success">Summarize
using
URL</legend>
{% if form.link.errors %}
```

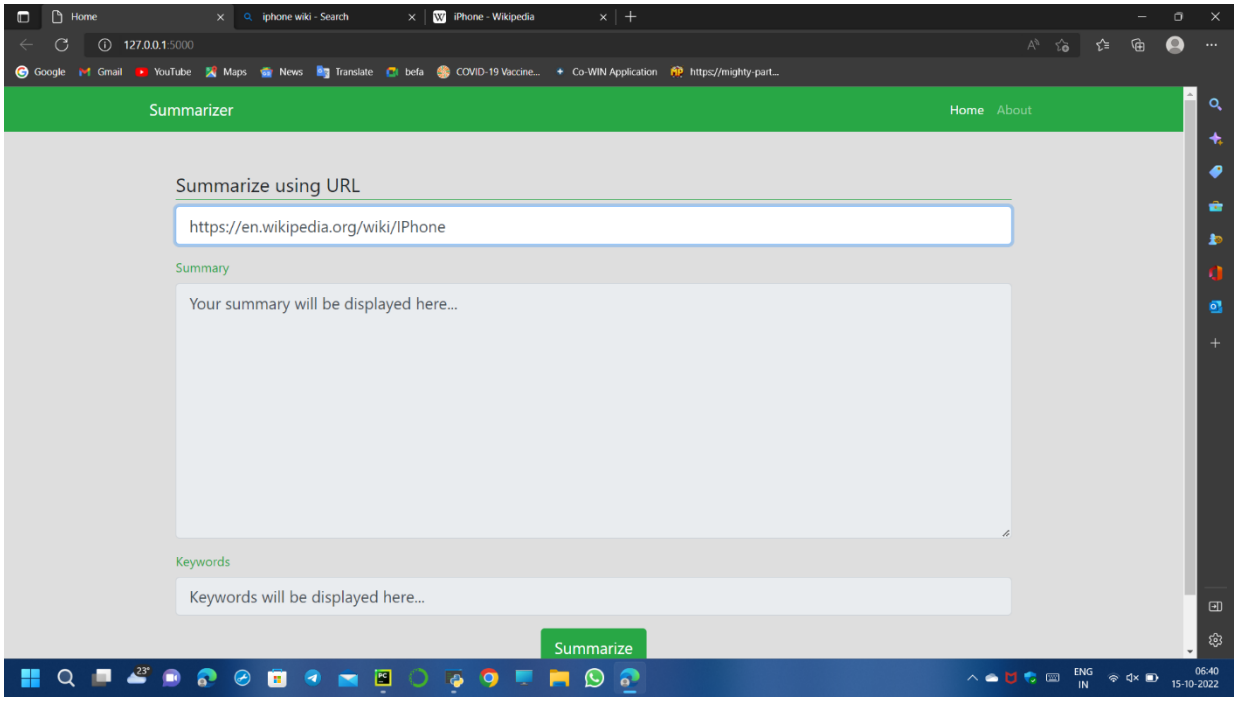
```
{{ form.link(class="form-
control form-control-lg is-
invalid",
placeholder="Enter a URL
here",
autocomplete="off") }}
<div class="invalid-
feedback">
{% for error in
form.link.errors %}
<span>{{ error }}</span>
{% endfor %}
</div>
{% else %}
{{ form.link(id="url",
class="form-control form-
control-lg",
placeholder="Enter a URL
here",
autocomplete="off") }}
{% endif %}
</fieldset>
<div class="form-group">
<label for="summary"
class="text-
success">Summary</label
>
<textarea name=""
id="summary" cols="35"
rows="10" class="form-
control
form-control-lg"
readonly>Your summary
will be displayed here...
</textarea>
```

```
<!-- {{
form.sum(class="form-
control form-control-lg",
cols="35",
rows="10",
placeholder="Your
summary will be displayed
here.....") }} -->
</div>
<div class="form-group">
<label for="keywords"
class="text-
success">Keywords</label
>
<input id="keywords"
type="text"
value="Keywords will be
displayed here..."
class="form-control form-
control-lg"
readonly>
</div>
<div class="form-group
text-center">
{{ form.submit(class="btn
btn-success btn-lg") }}
</div>
</form>
{% endblock main %}
```

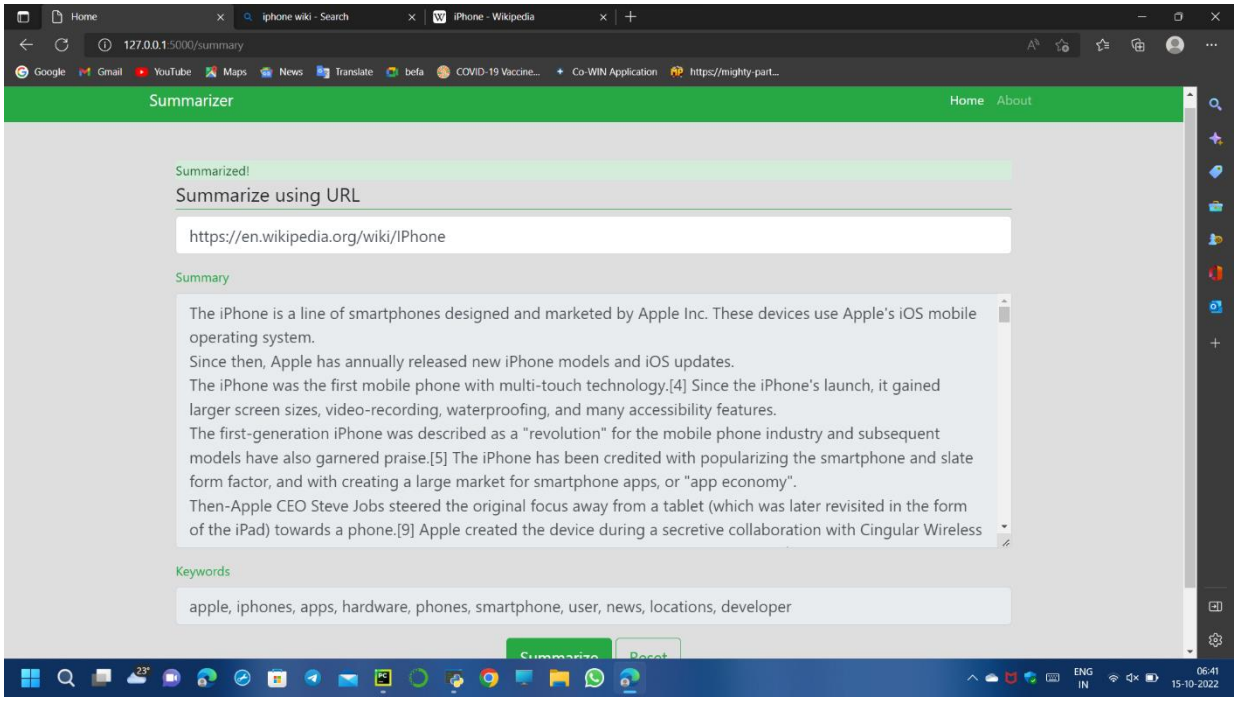
5.SCREENSHOTS



Screenshot 5.1: GUI of project



Screenshot 5.2: Input as URL



Screenshot 5.3: Summarized keyword Output

6.TESTING

6. TESTING

6.1 INTRODUCTION TO TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

6.2 TYPES OF TESTING

6.2.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .It is done after the completion of an individual unit before integration. This is a structural testing that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

6.2.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

6.2.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.
Functional testing is centered on the following items:

Valid Input	: identified classes of valid input must be accepted.
Invalid Input	: identified classes of invalid input must be rejected.
Functions	: identified functions must be exercised.
Output	: identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked. Organization and preparation of functional tests is focused on requirements, key functions, or special test cases.

6.3 TEST CASES

6.3.1 CLASSIFICATION

Test case ID	Test case name	Purpose	Input	Output
1	URL to be summarized	To extract the keywords	The user gives the input in the form URL.	The text will be summarized based on algorithm, Keywords will be displayed.

7. CONCLUSION

7.CONCLUSION & FUTURE SCOPE

7.1 PROJECT CONCLUSION

Text summarization is very helpful for users to extract only needed information in stipulated time. In this area, considerable amount of work has been done in the recent past. Due to lack of information and standardization lot of research overlap is a common phenomenon. Since 2012, exhaustive review paper is not published on automatic keyword extraction and text summarization especially in Indian context. Therefore, we thought that, the survey paper covering recent work in keyword extraction and text summarization may ignite the research community for filling some important research gaps. This paper contains the literature review of recent work in text summarization from the point of views of automatic keyword extraction, text databases, summarization process, summarization methodologies and evaluation matrices. Some important research issues in the area of text summarization are also highlighted in the project.

7.2 FUTURE SCOPE

In future, one can target following direction in the field of summarization:

➢ Text summarization in low resourced languages

➢ especially in Indian language context such as Telugu,

➢ Hindi, Tamil, Bengali, etc.

➢ This work can also be extended to multi-lingual text

➢ summarization.

➢ Multimedia summarization.

➢ Multi-lingual multimedia summarization.

8.BIBLIOGRAPHY

8. BIBLIOGRAPHY

8.1 REFERENCES

[1] J.N Mad-hurt and R. Ganesh Kumar, Extractive Text Summarization Using Sentence Ranking, 2019.

[2] Sethi Prakhar, Sonawane Sameer, Khanwalker Saumitra and R. B. Keskar, Automatic Text Summarization of News Articles ", 2017..

[3] Yuan Hua, Xu Hualin, Qian Yu and Ye Kia, Towards Summarizing Popular Information form massive Tourism Blogs, 2016.

[4] Wu Yutong, Gao Yang, Li Yuefeng and Xu Yue, Mining Topical Relevant Patterns for Multidocument Summarization, 2015.

[5] He Yan-xiang, Liu De-xi, Ji Dong-hong, Yang Hua and Teng Chong, A Multi-document Summarization System Based On Genetic Algorithm, 2006.

8.2 GITHUB LINK

https://github.com/sravanreddy9705/automatic-keyword-extraction-for-text-summarization-using-nlp.git