A

Major Project Report

On

TEXT SUMMARIZATION USING WORD FREQUENCIES

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

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

MOHAMMED SUFIYAN	**(187R1A0598)**
NAMA MEGANA	**(187R1A05B1)**
CHOPPALA ROHIT	**(187R1A0573)**

Under the Guidance of

Dr. Punyaban Patel

(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.

2018-2022

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled **“TEXT SUMMARIZATION USING WORD FREQUENCIES”** being submitted by **MOHAMMED SUFIYAN (187R1A0598), NAMA MEGANA (187R1A05B1), CHOPPALA ROHIT (187R1A0573)** 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 him/her under our guidance and supervision during the year 2021-22.

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

DR. PUNYABAN PATEL
(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 **Dr. Punyaban Patel,** 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 Project Review Committee (PRC) **Mr. A. Uday Kiran, Mr. J. Narasimha Rao, Dr. T. S. Mastan Rao, Mrs. G. Latha, Mr. A. Kiran Kumar,** 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 courseof 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 of this project.

MOHAMMED SUFIYAN	**(187R1A0598)**
NAMA MEGANA	**(187R1A05B1)**
CHOPPALA ROHIT	**(187R1A0573)**

ABSTRACT

Digital data has become an important aspect of machine learning and is present in huge volumes on the internet. To use this data efficiently, data handling and processing techniques are required to filter out information from documents and store them. An application of natural language processing, which helps in handling volumes of data, is text summarization. Text summarization helps in condensing documents, and extract the important facts represented in it. There are two techniques in text summarization: abstractive and extractive summarization. Extractive Summarization extracts keywords from the document and combines them to provide a semantically incorrect summary, whereas, Abstractive Summarization produces a semantically correct summary of the text. In this paper, we compare different techniques to identify low and medium frequency words. We evaluate the techniques based on the correct identification of positiveand negative words.

Keywords: *Text summarization, Abstractive summarization, Extractive Summarization, Semantic, NLP, Word Tokenization, Sentence Ranking.*

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

Page No.

1. INTRODUCTION

1. INTRODUCTION

1.1 PROJECT SCOPE

This project is termed as “Text Summarization Using Word Frequencies”. This provides facility to reduce the large volume of content by reducing it into a precise summary. This project uses Natural language processing to extract the summary from the given input text. We use a NLP toolkit and regular expressions. Millions of web pages and websites exist on the Internet today. Going through a vast amount of content becomes very difficult to extract information on a certain topic. Google will filter the search results and give you the top ten search results, but often you are unable to find the right content that you need. There is a lot of redundant and overlapping data in the articles which leads to a lot of wastage of time. The better way to deal with this problem is to summarize the text data which is available in large amounts to smaller sizes.

1.2 PROJECT PURPOSE

The objective of this project is to understand the concepts of Natural language processing and creating a tool for a text summarization. The concern in an automatic summarization is increasing so the manual work is removed. The project concentrates on creating a tool, which can automatically summarize the document. With the growing amount of data in the world, the interest in the field of automatic summarization generation has been widely increasing so as to reducing the manual effort of a person working on it.

1.3 PROJECT FEATURES

The features of Text Summarization Using Word Frequencies are as follows:

- This Project will reduce huge amount of information into a precise summary by using word tokenization and sentence ranking.
- This is done based on the previous datasets of text summarization so after comparing it can provide up to 80% of accurate results, and the project is still developing further to get the 100% accurate results.
- With the help of text summarization, it can largely reduce the load on the user which can save the time and the user can get interest in his reading.

1.4 GOALS

The main objective of this research work is to create an efficient system that is able to extract the sentences and words that are repeated in the input text and return an effective summary. Thus, the goals below can be derived from this:

- The system should be able to count the word frequencies.
- The system should be able to classify the words based on their ranking.
- The system should be configurable in terms of sentence classes used.
- The system should be able to rank the sentences based on word frequencies.
- The system at the end should produce an efficient summary.

2. SYSTEM ANALYSIS

2. 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 analysis as a firm understanding of what is to be done.

2.1 PROBLEM DEFINITION

As of late, there has been a blast in the measure of text data from an assortment of sources. This volume of text is a priceless source of information and knowledge, which should be effectively summarized to be useful. In this problem, the main objective is to automatictext summarization are described below for lighting more about processes. With the dramatic growth of the Internet, people are overwhelmed by the tremendous amount of online information and documents. This expanding availability of documents has demanded exhaustive research in automatic text summarization [1] [2]. Now days many research is going on for text summarization. Because of increasing information in the internet, these kinds of research are gaining more and more attention among the researchers. Extractive text summarization generates a summary by extracting proper set of sentences from a document or multiple documents by deep learning. The whole concept is to reduce or minimize the valuable information present in the documents.

2.2 EXISTING SYSTEM

Text summarization approach is based on the removal of redundant sentences. A text summarization approach using natural language processing and various extractive summary approaches like statistical based, topic-based, graph-based and machine learning based. The features with better results of extractive summarization can be combined together to make better summarization of the text [4]. A text summarization approach using natural language processing and various extractive summary approaches like statistical based, topic-based, graph-based and machine learning based [9]. The features with better results of extractive summarizationcan be combined together to make better summarization of the text.

2.2.1 LIMITATIONS OF EXISTING SYSTEM

- As the text is in large amount, it makes it quite difficult for us to read the text.
- It is time consuming.

2.3 PROPOSED SYSTEM

Reading the large content of text available online is challenging for the users as it consumes a large amount of time. The proposed system, Automatic Text Summarization is much more practical and applicable in real time. The input text is processed using natural language processing and processed input is converted into vector form using word embedding [3] [4]. 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**.**

2.3.1 TEXT SUMMARIZATION

The practice of breaking down long publications into manageable paragraphs or sentences is known as text summary [5] [6]. The approach extracts vital information while preservingthe meaning of the text. This reduces the time it takes to comprehend large resources, such as research articles, while without omitting important information.

Text summarizing is the process of creating a brief, coherent, and fluent summary of a longer text document, which involves highlighting the text's key points [11].

Text summarization involves a number of challenges, including text detection, interpretation, and summary generation, as well as examination of the final summary [7]. In extraction-based summarizing, identifying essential terms in the document and exploitingthem to unearth useful information to include in the summary are critical tasks.

2.3.2 NATURAL LANGUAGE PROCESSING

Natural Language Processing includes text summary, which is a very useful and vital feature (NLP). Let's start with a definition of text summarization. Assume we have an excessive amount of text data in any form, such as articles, periodicals, or social media posts.Due to a lack of time, we merely want a summary of the material [11]. By deleting unnecessarytext and translating the same text into smaller semantic text form, we can summarize our text in a few lines. In this approach we build algorithms or programs which will reduce thetext size and create a summary of our text data. This is called automatic text summarization inmachine learning [10]. Text summarization is the process of creating shorter text without removing the semantic structure of text.

2.3.3 ALGORITHM

Input: A text in .txt or .rtf format.

Output: A relevant summarized text which is shorter than the original text keeping the concept constant.

- Read a text in .txt or .rtf format and split it into individual tokens.
- Remove the stop words to filter the text.
- Assign a weight value to each individual terms. The weight is calculated as:

$$WT = \frac{Frequency\ of\ the\ term}{Total\ no\ of\ terms\ in\ the\ document}$$

2.3.4 TERM FREQUENCY & INVERSE DOCUMENT FREQUENCY

A High weight in TF-IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents.TF-IDF algorithm is made of two algorithms multiplied together [8].

Term Frequency

Term Frequency (TF) is how often a word appears in a document, divided by howmany words there are.

TF (t) = (Number of times term t appears in document)/ (Total number of terms in thedocument)

Inverse Document Frequency

Term Frequency is how common a word is, inverse document frequency (IDF) [8] is how unique or rare a word is.

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

Example**,**

Consider a document containing 100 words wherein the word ***happy*** appears 5 times. The term frequency (i.e., TF) for ***happy*** is then (5 / 100) = 0.05. Now, assume we have 10 million documents and the word ***happy*** appears in one thousands of these. Then, the inverse document frequency (i.e., IDF) is calculated as log (10,000,000 / 1,000) = 4.

Thus, the TF-IDF weight is the product of these quantities: 0.05 * 4 = 0.20.

2.3.5 ADVANTAGES OF THE PROPOSED SYSTEM

The system is very simple in design and to implement. The system requires verylow system resources, and the system will work in almost all configurations.

- In this approach, more than 60% of the generated sentences match with the original input text.
- It is cost effective and time efficient.
- Helps in better research work.

2.4 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with every 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 are

- Economic Feasibility
- Technical Feasibility
- Behavioural Feasibility.

2.4.1 ECONOMIC FEASIBILITY

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on 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 of 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 BEHAVIOURAL 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 specifies the logical characteristics of each interface between the software product and the hardware components of the system. The following are some hardware requirements.

• System	:	Intel core i3
• Hard disk	:	20 GB
• RAM	:	4GB
• Processor	:	2.7 GHz or more

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 10
• Languages	:	Python
• Tool	:	Anaconda Navigator(Jupyter)

3. ARCHITECTURE

3. ARCHITECTURE

3.1 PROJECT ARCHITECTURE

This architecture shows the procedure followed for Text summarization using word frequencies, where the input is preprocessed further the sentence and word tokenization is done and sentences are joined based on sentence ranking which produces the result as shown in Figure 3.1.

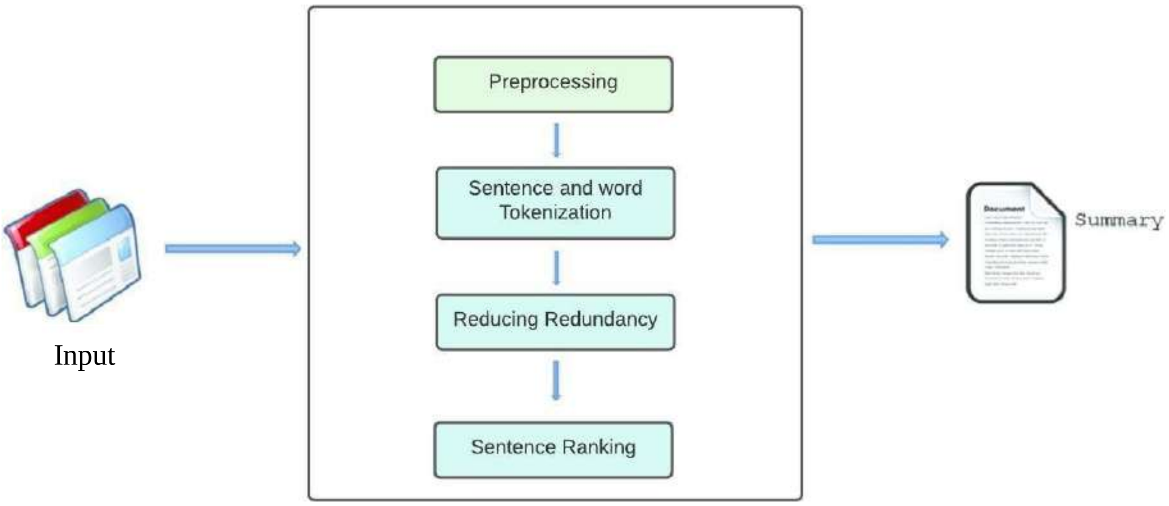


Figure 3.1 Project Architecture for Text Summarization Using Word Frequencies.

3.2 USE CASE DIAGRAM

The use case diagram the user uploads the input file in text format which undergoes text preprocessing (i.e., removal of stop words), then tokenization of words is done, apply frequencies to the words sentence ranking is done for the output is shown in Figure 3.2.

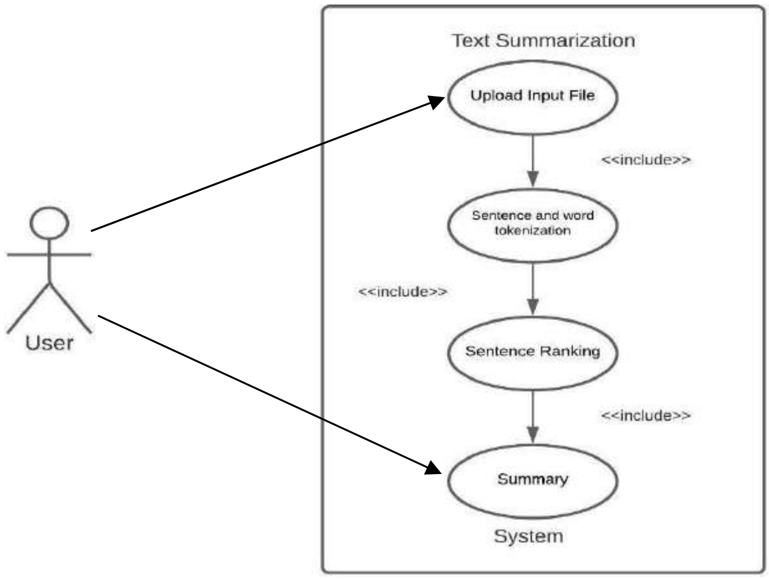


Figure 3.2 Use Case Diagram for Text Summarization Using Word Frequencies.

3.3 CLASS DIAGRAM

Class Diagram is a collection of classes and objects. In this diagram, we have three classes namely, Text, Sentence and Words. Each class diagram has some attributes and some operations. The attributes are made public as shown in Figure 3.3.

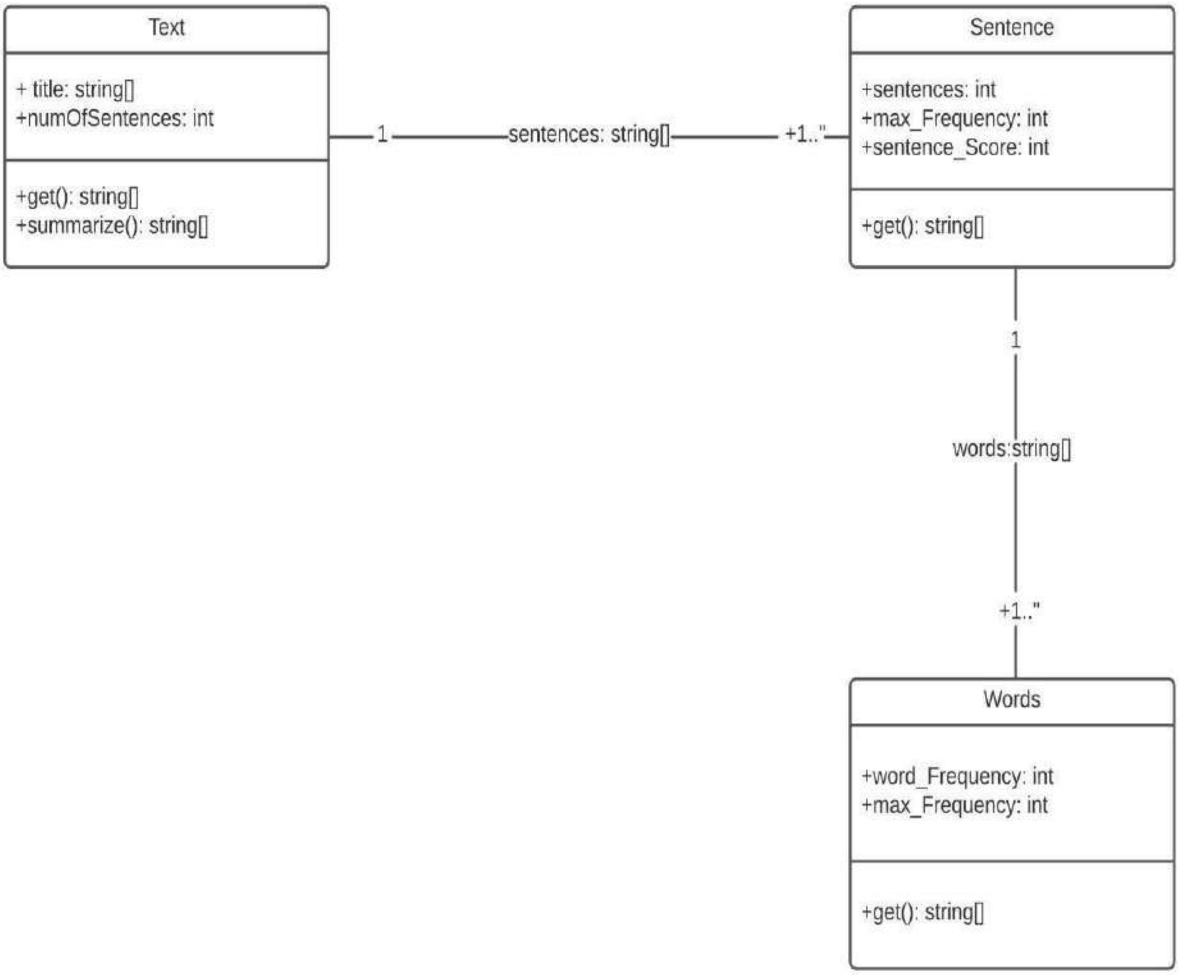


Figure 3.3 Class Diagram for Text Summarization Using Word Frequencies.

3.4 SEQUENCE DIAGRAM

In the Sequence diagram, the interaction between objects is observed i.e. the order in which these interactions takes place. The Figure contains one user and three objects in which sequence of operations takes place starting from uploading the file until the end as shown in Figure 3.4.

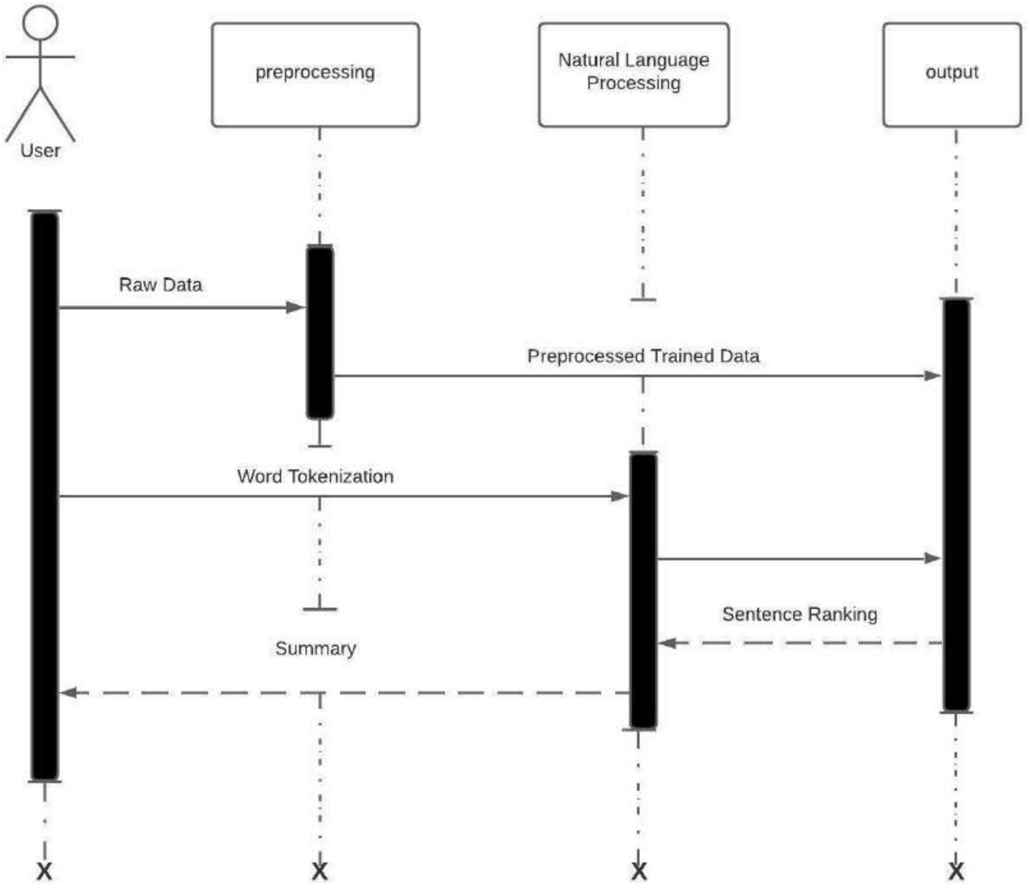


Figure 3.4 Sequence Diagram for Text Summarization Using Word Frequencies.

3.5 ACTIVITY DIAGRAM

An activity diagram depicts the behavior of the system. The starting state before an activity is known as initial state. In the first stage the text document is uploaded then in the next stage the preprocessing takes place further proceeds to sentence and word tokenization, in the third stage sentence scores are calculated based on word frequency, the fourth stage consists of choosing the important sentences and finally combining them to produce the output as shown in Figure 3.5.

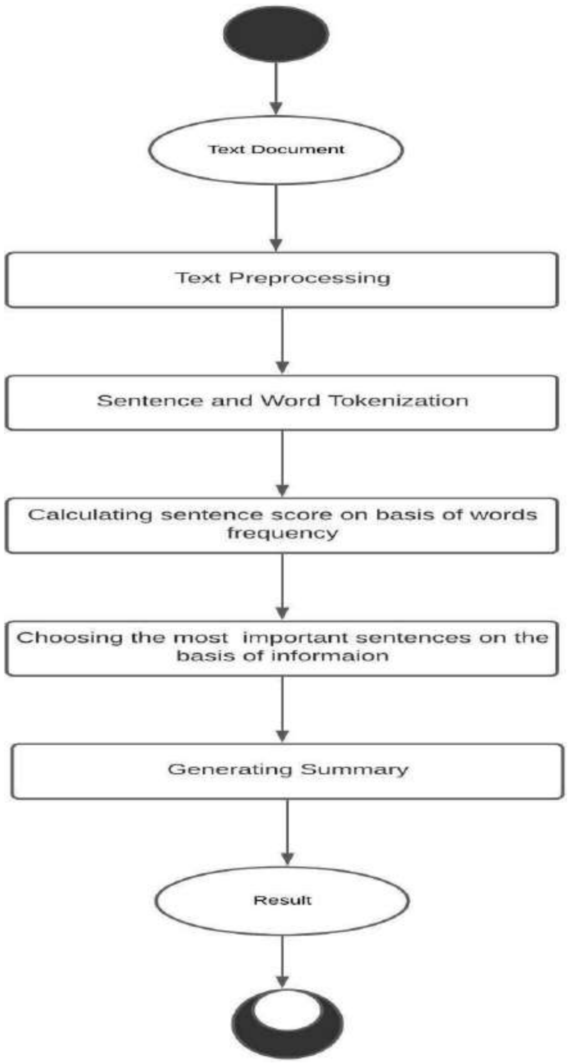

Text Document
Text Preprocessing
Sentence and Word Tokenization
Calculating sentence score on basis of words frequency
Choosing the most important sentences on the basis of informaion
Generating Summary
Result

Figure 3.5 Activity Diagram for Text Summarization Using Word Frequencies.

4. IMPLEMENTATION

4. IMPLEMENTATION

The sample code for Text Summarization using Word Frequencies is shown below:

4.1 SAMPLE CODE

```
 # input text article
 article_text= ' '

 # importing documents
import re
import nltk
nltk.download('stopwords')
article_text = article_text.lower()

# remove spaces, punctuations and numbers
clean_text = re.sub('[^a-zA-Z]', ' ', article_text)
clean_text = re.sub('\s+', ' ', clean_text)

# split into sentence list
sentence_list = nltk.sent_tokenize(article_text)
sentence_list

# calculate the frequency of words in each sentence.
stopwords   =   nltk.corpus.stopwords.words('english')
word_frequencies = {}
for word in nltk.word_tokenize(clean_text): if
     word not in stopwords:
             if   word   not   in   word_frequencies:
                     word_frequencies[word] = 1
             else:
                    word_frequencies[word] += 1

# calculate the term frequency for each word in a paragraph
maximum_frequency   =   max(word_frequencies.values())
for word in word_frequencies:
      word_frequencies[word] = word_frequencies[word] / maximum_frequency

# create a table and calculate documents for words(IDF) matrix.
sentence_scores = {}
for sentence in sentence_list:
     for word in nltk.word_tokenize(sentence):
            if word in word_frequencies and len(sentence.split(' ')) < 30:
                    if sentence not in sentence_scores:
                           sentence_scores[sentence] = word_frequencies[word]
                    else:
                          sentence_scores[sentence] += word_frequencies[word]
```

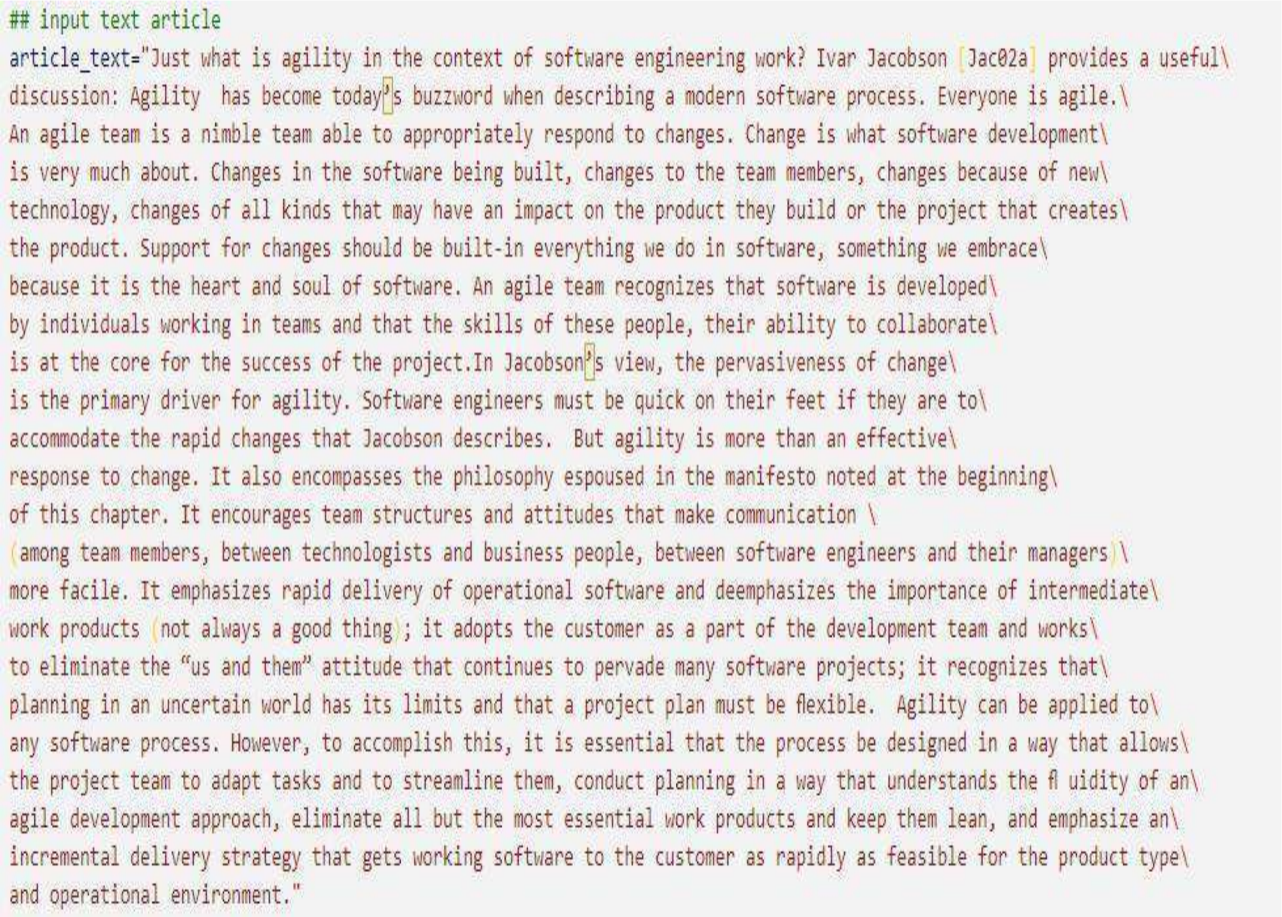
```
word_frequencies
sentence_scores

# get the top sentences

 import heapq

# generate the summary
summary = heapq.nlargest(5, sentence_scores, key=sentence_scores.get)
print(" ".join(summary))
```

5. SCREENSHOTS

5.SCREENSHOTS

5.1 INPUT

The input text for Text Summarization is shown in Figure 5.1.

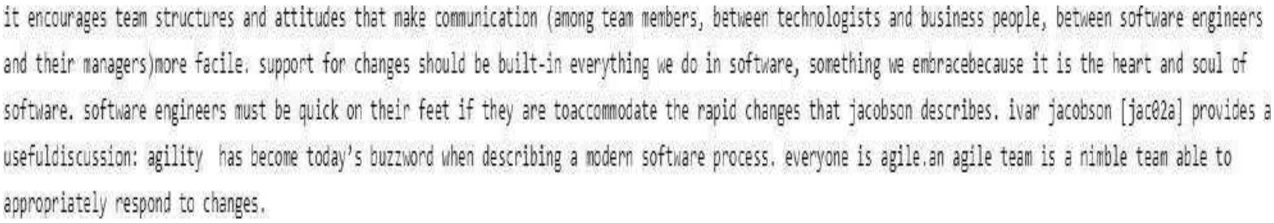


```
## input text article
article_text="Just what is agility in the context of software engineering work? Ivar Jacobson [Jac02a] provides a useful\
discussion: Agility  has become today’s buzzword when describing a modern software process. Everyone is agile.\
An agile team is a nimble team able to appropriately respond to changes. Change is what software development\
is very much about. Changes in the software being built, changes to the team members, changes because of new\
technology, changes of all kinds that may have an impact on the product they build or the project that creates\
the product. Support for changes should be built-in everything we do in software, something we embrace\
because it is the heart and soul of software. An agile team recognizes that software is developed\
by individuals working in teams and that the skills of these people, their ability to collaborate\
is at the core for the success of the project.In Jacobson’s view, the pervasiveness of change\
is the primary driver for agility. Software engineers must be quick on their feet if they are to\
accommodate the rapid changes that Jacobson describes.  But agility is more than an effective\
response to change. It also encompasses the philosophy espoused in the manifesto noted at the beginning\
of this chapter. It encourages team structures and attitudes that make communication \
(among team members, between technologists and business people, between software engineers and their managers)\
more facile. It emphasizes rapid delivery of operational software and deemphasizes the importance of intermediate\
work products (not always a good thing); it adopts the customer as a part of the development team and works\
to eliminate the “us and them” attitude that continues to pervade many software projects; it recognizes that\
planning in an uncertain world has its limits and that a project plan must be flexible.  Agility can be applied to\
any software process. However, to accomplish this, it is essential that the process be designed in a way that allows\
the project team to adapt tasks and to streamline them, conduct planning in a way that understands the fl uidity of an\
agile development approach, eliminate all but the most essential work products and keep them lean, and emphasize an\
incremental delivery strategy that gets working software to the customer as rapidly as feasible for the product type\
and operational environment."
```

Figure 5.1 Input text before Text Summarization Using Word Frequencies.

5.2 OUTPUT

The output text for Text Summarization is shown in Figure 5.2.



```
it encourages team structures and attitudes that make communication (among team members, between technologists and business people, between software engineers
and their managers)more facile. support for changes should be built-in everything we do in software, something we embracebecause it is the heart and soul of
software. software engineers must be quick on their feet if they are toaccommodate the rapid changes that jacobson describes. ivar jacobson [jac02a] provides a
usefuldiscussion: agility  has become today’s buzzword when describing a modern software process. everyone is agile.an agile team is a nimble team able to
appropriately respond to changes.
```

Figure 5.2 Result image after Text Summarization Using Word Frequencies.

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 test. 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 valid. It is the testing of individual softwareunits 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. Unittests 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 run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, 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. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes.

6.3 TESTCASES

6.3.1 UPLOADING TEXT

The text file for Text Summarization using word frequencies uploaded as input.

Table 6.1 Uploading input file for Text Summarization Using Word Frequencies.

Test case ID	Test case name	Purpose	Test Case	Output
1	User uploads the text file	Use it for Summarization.	If the file uploaded is in correct format.	Accurate result
2	User uploads some other file in other format.	Use it for Summarization.	If the file uploaded is not in correct format.	No proper result

7. CONCLUSION

7. CONCLUSION & FUTURE.SCOPE

7.1 PROJECT CONCLUSION

This approach gives the brief account of the ideas and techniques used to produce reliable and meaningful summaries using cosine similarity sentence ranking algorithm. This is worth further exploration in other market domains. Evaluation results will indicate the security of the product. The main aim of an automatic summarization system is to produce a precise meaningful summary for a large volume of information available. But, it still requires a lot of improvement due to the huge amount of data available online in different formats.

7.2 FUTURE SCOPE

The tenet of this research goes a lot further than just evaluation, since summary involves having an accurate description of the content of the documents. The development of more focused summaries may lead to more consistent methods, which is highly desirable. However, automatic text summarization is still a very promising research area with many challenges ahead.

8. BIBLIOGRAPHY

8. BIBLIOGRAPHY

[1] Research Article An Automatic Multi document Text Summarization Approach Based on Naive Bayesian Classifier Using Timestamp Strategy.

[2] Research Article An Automatic Multi document Text Summarization Approach Based on Naive Bayesian Classifier Using Timestamp St.

[3] C. Lakshmi Devasena and M. Hemalatha, "Automatic Text categorization and summarization using rule reduction," IEEE-International Conference On Advances in Engineering, Science And Management (ICAESM -2012), Nagapattinam, Tamil Nadu, 2012, pp. 594-598.

[4] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3.

[5] A. R. Pal and D. Saha, "An approach to automatic text summarization using WordNet," 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, 2014, pp. 1169-1173.

[6] D. Hingu, D. Shah and S. S. Udmale, "Automatic text summarization of Wikipedia articles," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, 2015, pp. 1-4.

[7] Steinberger, J., & Ježek, K. (2008). Automatic Text Summarization (The state of the art 2007 and new challenges). Znalosti, 30(2), 1-12.

[8] Yohei, S. (2002). Sentence extraction by TF/IDF and Position Weighting from newspaper articles. In Proceedings of the Third NTCIR Workshop.

[9] Das, D., & Martins, A. F. (2007). A survey on automatic text summarization. Literature Survey for the Language and Statistics, 3(3), 1-12.

[10] Santosh Kumar Bharti, “Automatic Keyword Extraction for Text Summarization in Multidocument eNewspapers Articles”, European Journal of Advances in Engineering and Technology, Vol. 4, Issue 6, pp. 410-427, 2017.

[11] Roshna Chettri, Udit Kr. Chakraborty, “Automatic Text Summarization”, International Journal of Computer Applications, Vol. 161, Issue 1, pp. 5-7, 2017.

8.1 GITHUB LINK

- https://github.com/sufiyan98/Text_Summarization_Using_WordFrequencies

International Journal of Innovative Research in Engineering
Volume 3, Issue 3 (May-June 2022), PP: 396-398
www.theijire.com
ISSN No: 2582-8746

Text Summarization Using Word Frequencies

Punyaban Patel[1], Mohammed Sufiyan[2], Nama Megana[3], Choppala Rohit[4]
[1,2,3,4]Computer Science and Engineering, CMR Technical Campus, Hyderabad, India)

How to cite this paper:
Punyaban Patel[1], Mohammed Sufiyan[2], Nama Megana[3], Choppala Rohit[4], "Text Summarization Using Word Frequencies", IJIRE-V3I03-396-368

Abstract: *Digital knowledge has become a crucial side of machine learning and is gift in large volumes on the web. To use this knowledge with efficiency, knowledge handling and process techniques are needed to filter data from documents and store them. Associate degree application of linguistic communication process, that helps in handling volumes of knowledge, is text summarization. Text summarization helps in condensation documents, and extract the necessary facts described in it. There are 2 techniques in text summarization: theoretical and extractive summarization. Extractive summarization extracts keywords from the document and combines them to supply a semantically incorrect outline, whereas, theoretical summarization produces a semantically correct outline of the text. During this paper, we have a tendency to compare completely different techniques to spot low and radio frequency words. We have a tendency to measure the techniques supported the right identification of positive and negative words.*

Key Word: *Text summarization; Abstractive summarization; Extractive Summarization; Semantic; NLP; Word Tokenization; Sentence Ranking*

I. INTRODUCTION

This project is termed as "Text Summarization Using Word Frequencies". This provides facility to cut back the massive volume of content by reducing it into a certain outline. This project uses language process to extract the outline from the given input text. We have a tendency to use a IP toolkit and regular expressions. The objective of this project is to grasp the ideas of language process and making a tool for a text account. The priority in Associate in Nursing automatic account is increasing therefore the manual work is removed. The project concentrates on making a tool which may mechanically summarize the document. With the growing quantity of information within the world, the interest within the field of automatic account generation has been wide increasing thus on reducing the manual effort of someone engaged on it.

II. MATERIAL AND METHODS

Reading the big content of text out there on-line is difficult for the users because it consumes an oversized quantity of your time. The projected system, Automatic Text report is way a lot of sensible and applicable in real time. The input text is pro cessed mistreatment linguistic communication process and processed input is born-again into vector kind mistreatment word embedding. Word embedding is that the collective name for a group of language modelling and have learning techniques in informatics wherever words or phrases from the vocabulary area unit mapped to vectors of real numbers. Sentence ranking is completed between sentences to extract higher hierarchic sentence that forms the extractive outline of the input.

Text Summarization: Text summary is the process of cutting down large publications into manageable paragraphs or sentences. The method collects critical information while maintaining the text's meaning. This cuts down on the time it takes to interpret vast amounts of data while avoiding omitting vital information, such as research articles. The process of constructing a concise, cohesive, and fluent summary of a text is known as text summarization. Highlighting the text's essential points is part of a longer text piece. Text summarization presents a number of issues, including text detection, text summarization, and text summarization. Interpretation and creation of a summary, as well as a review of the final summary in extraction-based summarizing, recognizing and leveraging key terms in the document. It's up to them to find helpful material to include in the summary.

Natural Language Processing: Text summarization is a very helpful and important component of Natural Language Processing (NLP). Let's begin with an explanation of what text summarizing is assume we have an abundance of text data in any form, including papers, magazines, and social media posts. We only need a summary of the material due to a shortage of time. We can summarize our text in a few lines by eliminating extraneous text and translating it into smaller semantic text form. In this method, we construct algorithms or programs that minimize the amount of the text and generate a summary of our text data. In machine learning, this is known as automatic text summarization. The practice of reducing the length of a document without losing its semantic structure is known as Text summarization.

Algorithm:

Input: A text in .txt or .rtf format.

Output: A relevant summarized text which is shorter than the original text keeping the theme or concept constant.

1. Read a text in .txt or .rtf format and split it into individual tokens.
2. Remove the stop words to filter the text.
3. Assign a weight value to each individual terms. The weight is calculated as:

$$WT = \frac{Frequency\ of\ the\ term}{Total\ no\ of\ terms\ in\ the\ document}$$

Procedure Methodology:

Text report approach relies on the removal of redundant sentences. A text report approach exploitation language process and varied extractive outline approaches like applied mathematics based mostly, topic-based, graph - based and machine learning based mostly. The options with higher results of extractive report is combined along to form higher report of the text.

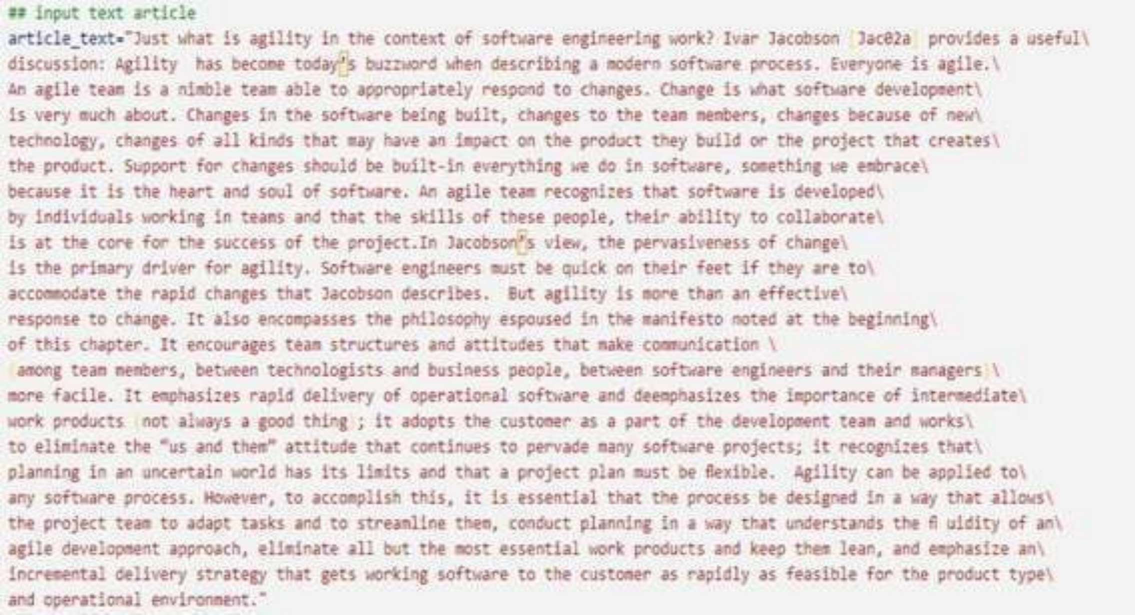
Text Summarization Steps:

1. Make sentences out of paragraphs.
2. Text Preprocessing.
3. Sentence Tokenization.
4. Calculate the Weighted Frequency of Occurrence.
5. In original sentences, replace words with weighted frequencies.
6. Arrange the sentences in descending order of their sum.

III. RESULT

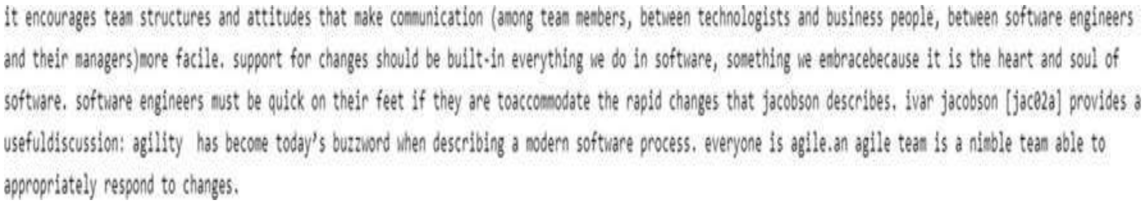
We've submitted the input file and performed data preprocessing, such as removing duplicates. We can observe the outcomes before and after applying the Natural Language Tokenization to sentences and words. Method for language processing wedelete extraneous sentences using NLP techniques. We first count the word frequencies, then rank the sentences based on the maximum word frequencies.

Figure no 1: Input



```
## input text article
article_text="Just what is agility in the context of software engineering work? Ivar Jacobson [Jac02a] provides a useful\
discussion: Agility  has become today's buzzword when describing a modern software process. Everyone is agile.\
An agile team is a nimble team able to appropriately respond to changes. Change is what software development\
is very much about. Changes in the software being built, changes to the team members, changes because of new\
technology, changes of all kinds that may have an impact on the product they build or the project that creates\
the product. Support for changes should be built-in everything we do in software, something we embrace\
because it is the heart and soul of software. An agile team recognizes that software is developed\
by individuals working in teams and that the skills of these people, their ability to collaborate\
is at the core for the success of the project.In Jacobson's view, the pervasiveness of change\
is the primary driver for agility. Software engineers must be quick on their feet if they are to\
accommodate the rapid changes that Jacobson describes.  But agility is more than an effective\
response to change. It also encompasses the philosophy espoused in the manifesto noted at the beginning\
of this chapter. It encourages team structures and attitudes that make communication \
(among team members, between technologists and business people, between software engineers and their managers) \
more facile. It emphasizes rapid delivery of operational software and deemphasizes the importance of intermediate\
work products (not always a good thing); it adopts the customer as a part of the development team and works\
to eliminate the "us and them" attitude that continues to pervade many software projects; it recognizes that\
planning in an uncertain world has its limits and that a project plan must be flexible.  Agility can be applied to\
any software process. However, to accomplish this, it is essential that the process be designed in a way that allows\
the project team to adapt tasks and to streamline them, conduct planning in a way that understands the fl uidity of an\
agile development approach, eliminate all but the most essential work products and keep them lean, and emphasize an\
incremental delivery strategy that gets working software to the customer as rapidly as feasible for the product type\
and operational environment."
```

Figure no 2: Output



```
it encourages team structures and attitudes that make communication (among team members, between technologists and business people, between software engineers
and their managers)more facile. support for changes should be built-in everything we do in software, something we embracebecause it is the heart and soul of
software. software engineers must be quick on their feet if they are toaccommodate the rapid changes that jacobson describes. ivar jacobson [jac02a] provides a
usefuldiscussion: agility  has become today's buzzword when describing a modern software process. everyone is agile.an agile team is a nimble team able to
appropriately respond to changes.
```

IV. DISCUSSION

This paradigm can be applied to situations in which a large amount of data must be examined and a conclusion or output must be formed. Data is growing at an exponential rate nowadays. This massive amount of data creates a dilemma in terms of analyzing it and extracting various beneficial conclusions from it. This is where the concept of summarizing comes into play. Summarization is a technique for condensing a large dataset into little, countable lines of data so that a user may obtain something useful out of it in a short amount of time. Websites, blogs, news articles, webpages, and books can all benefit from this strategy. Aspreviously said, data is created from a variety of sources nowadays.

We can utilize this project to quickly comprehend an entire book and then construct our evaluation based on that summary.Researchers and scientists must read a large number of scientific publications and patents; a summary tool may help them save timeskimming through the articles, increasing productivity and assisting them in new discoveries. Lawyers have a vast number of case files to go through, and sifting through them all is a time-consuming task. This tool will aid in the summarization of these case file data, allowing the lawyer to grasp the case entirely in a short period of time.

V. CONCLUSION

The ideas and strategies utilized to construct trustworthy and understandable summaries using the cosine similarity sentence ranking algorithm are briefly described in this approach. This merits additional investigation in other market segments. The product'ssecurity will be determined by the outcomes of the evaluation. The primary goal of an autonomous summarization system is to provide an accurate and understandable summary from a huge amount of data. However, due to the vast amount of data available online in various formats, it still needs a lot of work.

Text summarization as a branch of NLP is quickly increasing due to the need for compressed and relevant synopses of a topic, as well as the massive volume of information available on the internet. Business analysts, government organizations, teachers, development researchers, marketing executives, and students can all benefit from text summarizing. Precise information improves the accuracy and efficiency of the search process. Users need this to process information in a short amount of time. This method can be employed in both the business and scientific worlds. It is less time consuming and less complicated than abstractive text summarization approaches, however the resulting summary is less accurate and meaningful than Abstractive methods.

References

[1] *Research Article An Automatic Multi document Text Summarization Approach Based on Naive Bayesian Classifier Using Timestamp Strategy.*

[2] *Research Article An Automatic Multi document Text Summarization Approach Based on Naive Bayesian Classifier Using Timestamp St.*

[3] *C. Lakshmi Devasena and M. Hemalatha, "Automatic Text categorization and summarization using rule reduction," IEEE-International Conference On Advances in Engineering, Science And Management (ICAESM -2012), Nagapattinam, Tamil Nadu, 2012, pp. 594-598.*

[4] *J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3.*

[5] *A. R. Pal and D. Saha, "An approach to automatic text summarization using WordNet," 2014 IEEE International Advance Computing Conference (IACC),Gurgaon, 2014, pp. 1169-1173.*

[6] *D. Hingu, D. Shah and S. S. Udmale, "Automatic text summarization of Wikipedia articles," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, 2015, pp. 1-4.*

[7] *Steinberger, J., & Ježek, K. (2008). Automatic Text Summarization (The state of the art 2007 and new challenges). Znalosti, 30(2), 1-12.*

[8] *Yohei, S. (2002). Sentence extraction by TF/IDF and Position Weighting from newspaper articles. In Proceedings of the Third NTCIR Workshop.*

[9] *Das, D., & Martins, A. F. (2007). A survey on automatic text summarization. Literature Survey for the Language and Statistics, 3(3), 1-12.*

[10] *Santosh Kumar Bharti, "Automatic Keyword Extraction for Text Summarization in Multidocument eNewspapers Articles", EuropeanJournal of Advances in Engineering and Technology, Vol. 4, Issue 6, pp. 410-427, 2017.*

[11] *Roshna Chettri, Udit Kr. Chakraborty, "Automatic Text Summarization", International Journal of Computer Applications, Vol. 161, Issue 1, pp. 5-7, 2017.*

Certificate

OF PUBLICATION

IJIRE

THIS CERTIFICATE IS CONFIRM THAT

Mohammed Sufiyan

PUBLISHED FOLLOWING ARTICLE

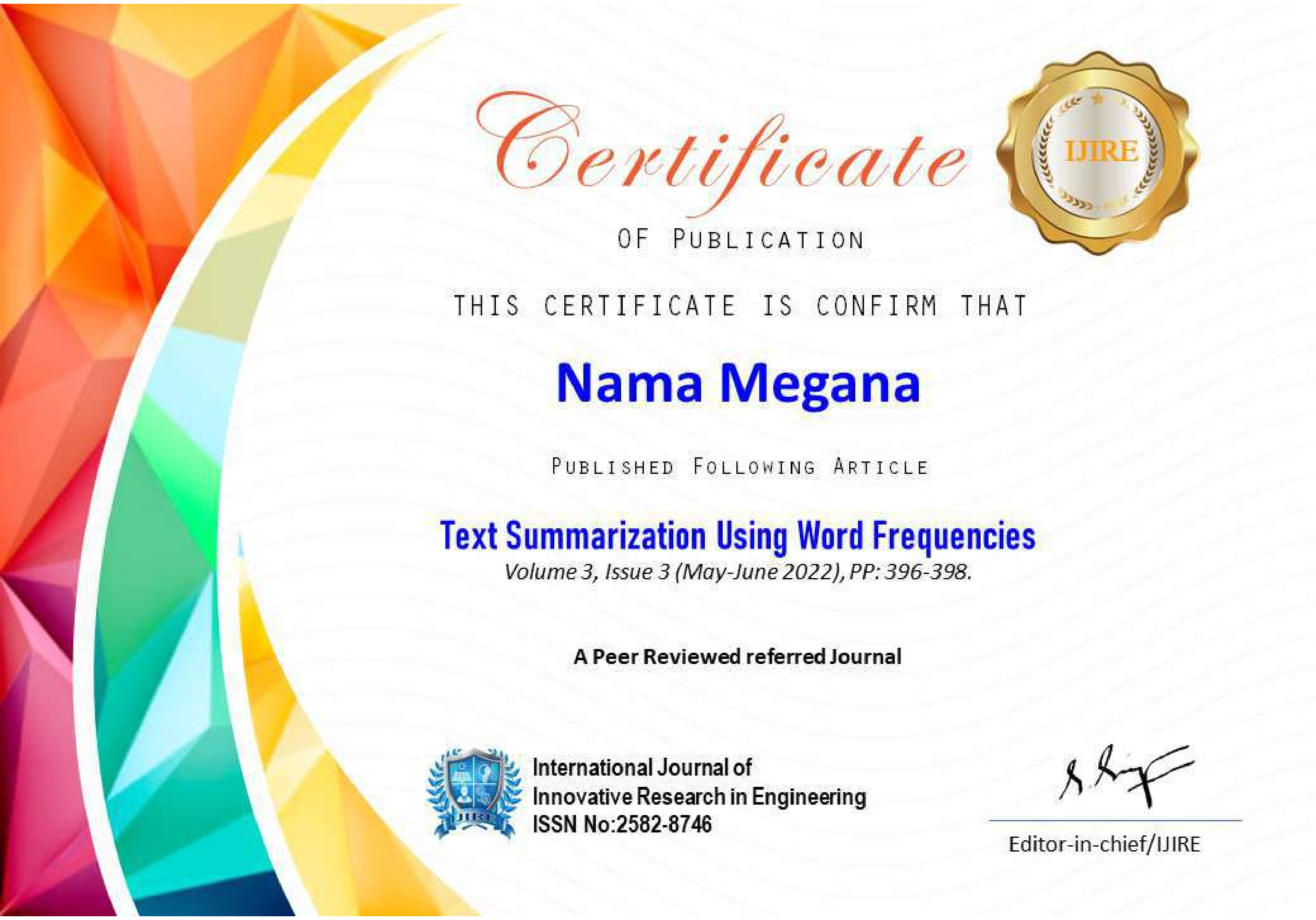
Text Summarization Using Word Frequencies

Volume 3, Issue 3 (May-June 2022), PP: 396-398.

A Peer Reviewed referred Journal

International Journal of
Innovative Research in Engineering
ISSN No:2582-8746

Editor-in-chief/IJIRE



Certificate

OF PUBLICATION

IJIRE

THIS CERTIFICATE IS CONFIRM THAT

Nama Megana

PUBLISHED FOLLOWING ARTICLE

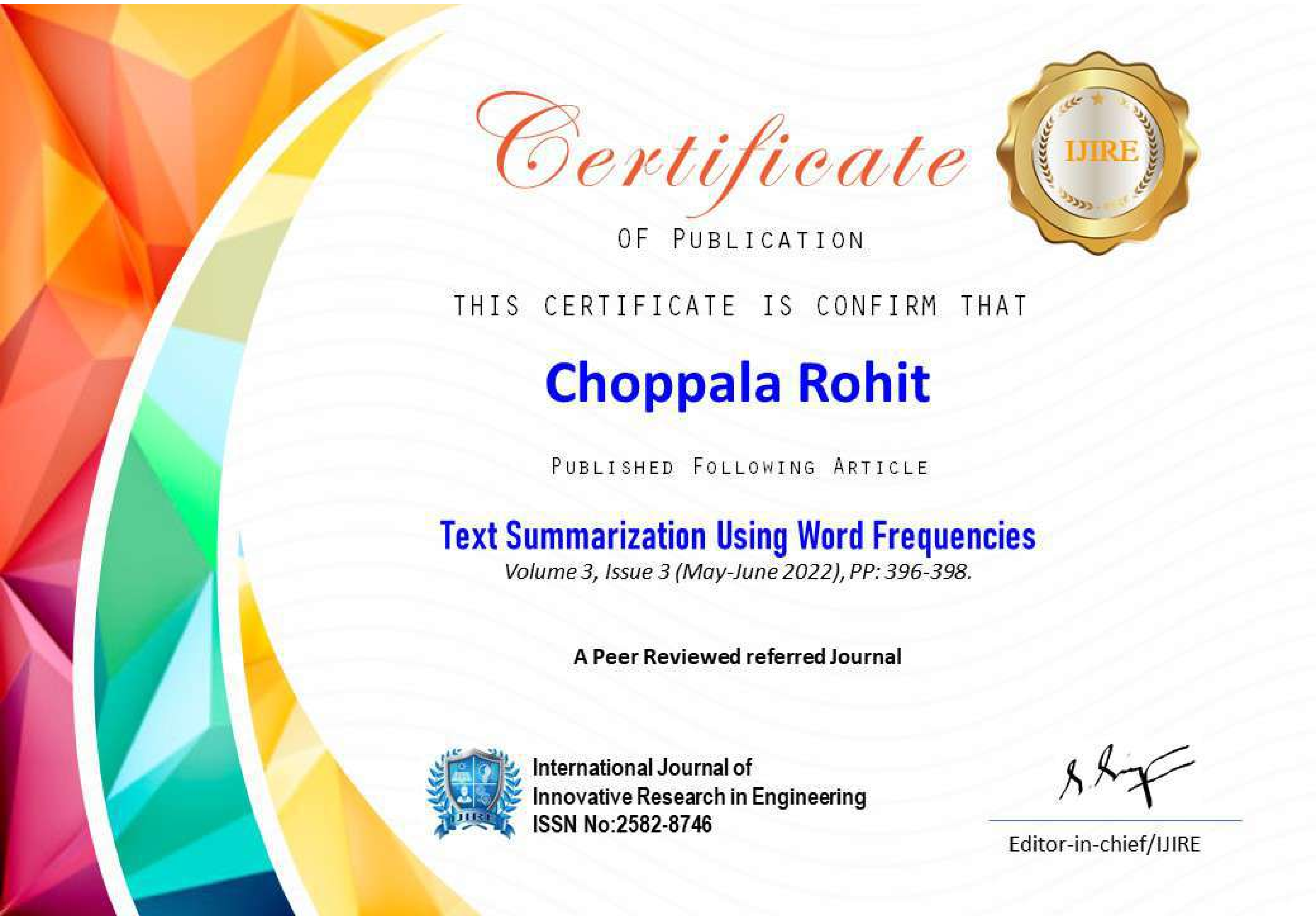
Text Summarization Using Word Frequencies

Volume 3, Issue 3 (May-June 2022), PP: 396-398.

A Peer Reviewed referred Journal

IJIRE
International Journal of
Innovative Research in Engineering
ISSN No:2582-8746

Editor-in-chief/IJIRE



Certificate

OF PUBLICATION

IJIRE

THIS CERTIFICATE IS CONFIRM THAT

Choppala Rohit

PUBLISHED FOLLOWING ARTICLE

Text Summarization Using Word Frequencies

Volume 3, Issue 3 (May-June 2022), PP: 396-398.

A Peer Reviewed referred Journal

International Journal of
Innovative Research in Engineering
ISSN No:2582-8746

Editor-in-chief/IJIRE