GUJARAT TECHNOLOGICAL UNIVERSITY

Chandkheda, Ahmedabad

Affiliated

A.D. Patel Institute of Technology

Project Report On

MultiPurpose Text Summarizer

Under

Final Year Project

B.E. Semester – VIII

Computer Engineering

Sr.no	Name of Student	Enrollment no.
1	Vrundan Patel	160010107040
2	Ankit Sharda	160010107050
3	Dharmesh Vekariya	160010107062

Prof. Gopi Bhatt

Faculty Guide

Prof. Bhagirath Prajapati

Head of the department

Academic Year

2019-20

A.D. Patel Institute of Technology

CERTIFICATE

This is to certify that a Final Year Project Report entitled **"Multipurpose Text Summarizer"** has been carried out by Vrundan Patel (160010107040), Ankit Sharda (160010107050), Dharmesh Vekariya (160010107062) towards the partial fulfilment of the degree of **Bachelor of Engineering of ADIT** the work carried out by them under my guidance and supervision. The work submitted, in my opinion, has reached to a level required for being accepted for the examination.

Date of the submission:

Prof. Gopi Bhatt
Faculty Guide

Prof. Bhagirath Prajapati
Head of the department

GUJARAT TECHNOLOGICAL UNIVERSITY

ACKNOWLEDGMENT

We take this opportunity to express our sincere gratitude to all those who helped us in various capacities in undertaking this project and devising the report.

We are privileged to express our sense of gratitude to respected faculty guide Prof. Gopi Bhatt, who guided us through the difficulties, faced in the project and provided an immense support in completing the subject successfully.

We are also grateful to Prof. Bhagirath Prajapati, Head of Department, Computer Engineering for his much needed guidance.

We take this opportunity also to thank our friends and contemporaries for their co-operation and compliance.

Vrundan Patel

160010107040

Ankit Sharda

160010107050

Dharmesh Vekariya

160010107062

CONTENT

1. Introduction

1.1 Problem Summary

Text Summarization is the process of obtaining salient information from an authentic text document. In this technique, the extracted information is achieved as a summarized report and conferred as a concise summary to the user. It is very crucial for humans to understand and to describe the content of the text. Text Summarization techniques are classified into abstractive and extractive summarization.

There is a wealth of textual content available on the internet. But, usually, the internet contribute more data than is desired. Therefore, a twin problem is detected: Seeking for appropriate documents through an awe-inspiring number of reports offered, and fascinating a high volume of important information.

The objective of automatic text summarization is to condense the original text into a precise version preserves its report content and global denotation. The main advantage of a text summarization is reading time of the user can be reduced. A marvellous text summary system should reproduce the assorted theme of the document even as keeping repetition to a minimum.

1.2 Objective

Purpose of our project is where given a document with huge text content it will be converted into a precise summary. The summary will not only retain the essence of the document, but will also save a lot of time and effort. An effective summary of the document will concise and fluent while preserving key information and overall meaning. Text Summarizer is a web-based application which helps in summarizing the text. We can upload our data and this application gives us the summary of that data in as many numbers of lines as we want. The main purpose is to provide reliable summaries of web pages or uploaded files depends on the user’s choice. The unnecessary sentences will be discarded to obtain the most important sentences.

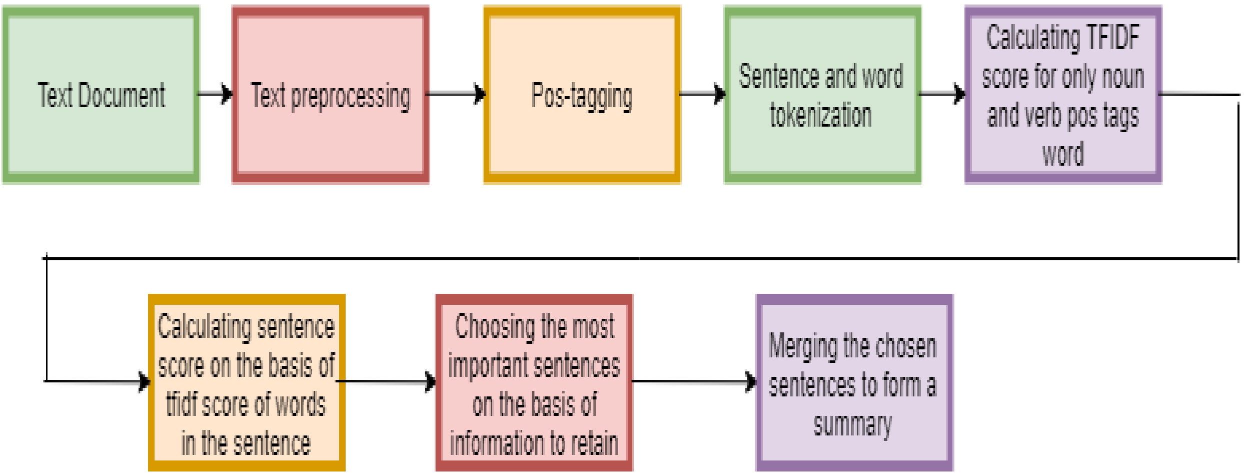
1.3 Problem Specification

The objective of automatic text summarization is to condense the original text into a precise version preserves its report content and global denotation. The main advantage of a text summarization is reading time of the user can be reduced. A marvellous text summary system should reproduce the assorted theme of the document even askeeping repetition to a minimum.

Term frequency–inverse document frequency, is basically a numerical term that is meant to show how important a word is to a document in a collection of documents. In information retrieval systems, it is used as a weighting factor. As the number of times a word appears in the document increases, the tf-idf value increases proportionally. But this tf-idf value is decreased by the frequency of the word in the collection. This helps to take into account the fact that some words appear more frequently in general.

1.4 Plan of Work

When needed to summarize the large document, we used TFIDF algorithm to do some specified number of the steps on the document that need to be summarized. These steps are tokenization of text, creating the frequency matrix of the each word in the sentence, calculate the term frequency and Inverse document frequency for the text and then using TF-IDF score we can summarize the uploaded or given document.


Text Document
Text preprocessing
Pos-tagging
Sentence and word tokenization
Calculating TFIDF score for only noun and verb pos tags word
Calculating sentence score on the basis of tfidf score of words in the sentence
Choosing the most important sentences on the basis of information to retain
Merging the chosen sentences to form a summary

1.5 Tools Required

1) Python

Python is an interpreted, high-level, general-purpose programming language. The biggest strength of the Python is large library. It is our project's base on which we implement all other libraries.

2) NLTK

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

3) NLP Libraries

NLP libraries are used to process the textual data and find some results. The main use of NLP libraries is in text preprocessing step where we perform text cleaning steps like text tokenization, text lemmatization, eliminating stop words and short words, removing duplicate words.

4) HTML

Hypertext Mark-up Language (HTML) is the standard mark-up language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript. We have used HTML to design the front-end of the Text Summarization System. Various major tags of HTML like textbox, button, etc. are used in the system.

5) **CSS**

Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a mark-up language like HTML. CSS is designed to enable the separation of presentation and content, including layout, colours, and fonts. We have used various CSS properties like border, margin, padding, background-image, font related properties, etc. for proper styling of the User Interface.

6) **JavaScript**

Alongside HTML and CSS, JavaScript is one of the core technologies of the World Wide Web. JavaScript enables interactive web pages and is an essential part of web applications. The vast majority of websites use it for client-side page behaviour, and all major web browsers have a dedicated JavaScript engine to execute it. We have used JavaScript to handle onclick event.

7) **Django**

Django is a Python-based free and open-source web framework that follows the model-template-view (MTV) architectural pattern. Django's primary goal is to ease the creation of complex, database-driven websites. The framework emphasizes reusability and "pluggability" of components, less code, low coupling, rapid development, and the principle of don't repeat yourself. We have used Django framework for running python code from frontend webpage.

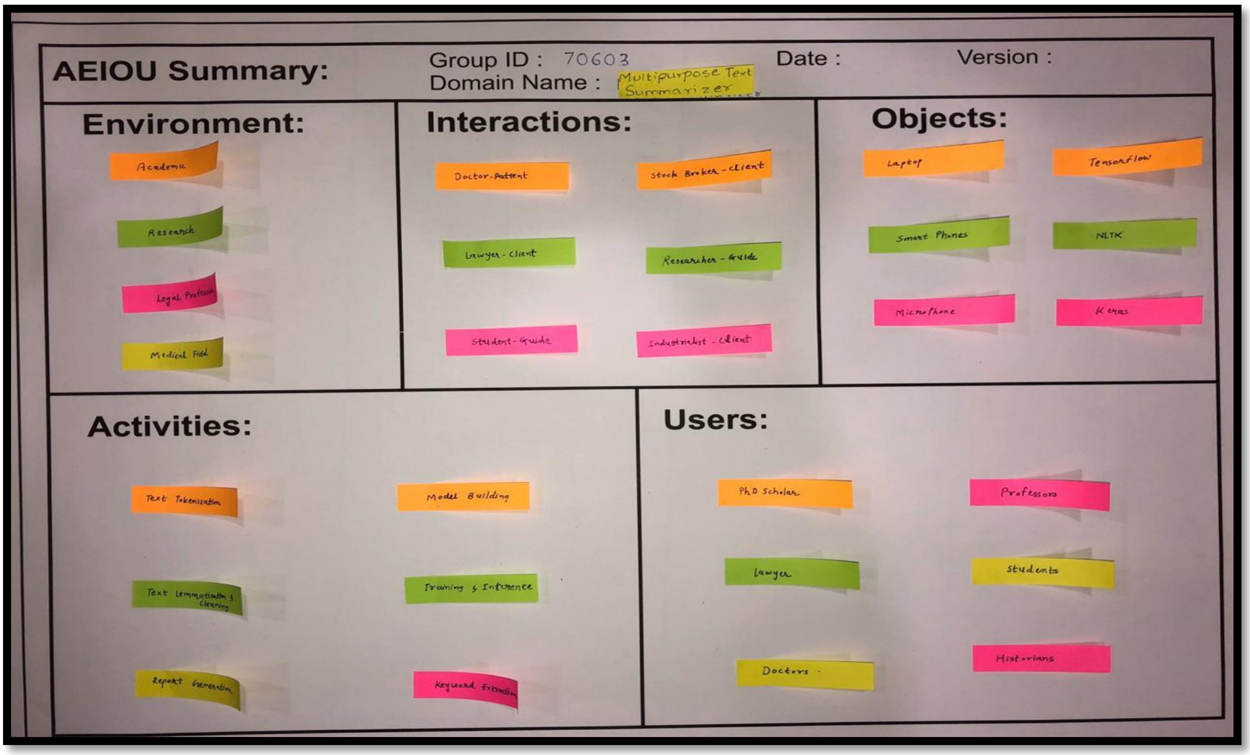
8) **Transcrypt**

Transcrypt gives us a command-line tool which we can run to compile a Python script into a JavaScript file. In Transcrypt, we can interact with the page structure (the DOM) using a toolbox of specialized Python objects and functions. Transcrypt makes the conversion to JavaScript at the earliest possible time - before the browser is even running.

2. Canvases

2.1 AEIOU CANVAS:

In this canvas, activities, environment in which the application works, interactions between users and application, objects and users are defines. Environment includes the tools and packages required for the development of the system. Objects are the items used to provide the functionality and develop the system accordingly. Activities include the functions provided by the system to the user. Users are the people using and interacting with the system.



AEIOU Summary:

Group ID : 70603

Domain Name : Multipurpose Text Summarizer

Date :

Version :

Environment:

Academic

Research

Legal Profession

Medical Field

Interactions:

Doctor-Patient

Stock Broker - client

Lawyer - client

Researcher - Guide

Student- Guide

Industrialist - client

Objects:

Laptop

Tensorflow

Smart Phones

NLTK

Microphone

Keras

Activities:

Text Tokenization

Model Building

Text Lemmatization & Cleaning

Training & Inference

Report Generation

Keyword Extraction

Users:

PhD Scholar

Professors

Lawyer

Students

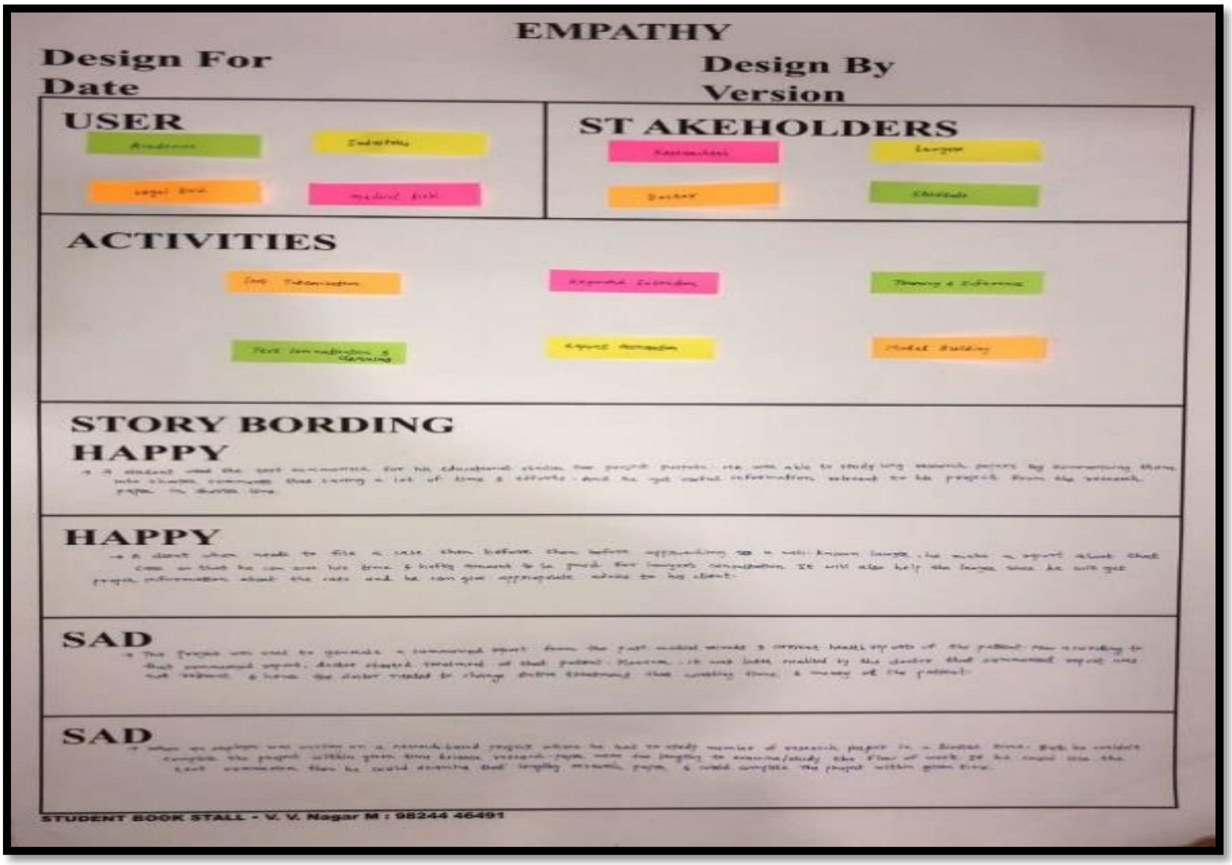
Doctors

Historians

AEOIU Canvas

2.2 EMPATHY MAPPING:

Our project is Multipurpose Text Summarizer. So we have listed the users and stakeholders who are benefitted from this system. We have also mentioned the activities that will be performed in this project. We have also described two happy stories and sad stories related to this project.



EMPATHY

Design For
Date

Design By
Version

USER

Academics
Industries
Legal Field
Medical Field

ST AKEHOLDERS

Researchers
Lawyer
Doctor
Students

ACTIVITIES

Text Tokenization
Keyword Extraction
Training & Inference
Text summarization & cleaning
Layout Recognition
Model Building

STORY BORDING

HAPPY

→ A student used the text summarizer for his educational studies for project purpose. He was able to study long research papers by summarizing them into shorter summaries thus saving a lot of time & efforts. And he got useful information relevant to his project from the research paper in shorter time.

HAPPY

→ A client when needs to file a case then before then before approaching to a well-known lawyer, he makes a report about that case so that he can save his time & hefty amount to be paid for lawyer's consultation. It will also help the lawyer since he will get proper information about the case and he can give appropriate advice to his client.

SAD

→ The project was used to generate a summarized report from the past medical records & ongoing health reports of the patient. Now according to that summarized report, doctor started treatment of that patient. However, it was later realized by the doctor that summarized report was not relevant & hence the doctor needed to change entire treatment thus wasting time & money of the patient.

SAD

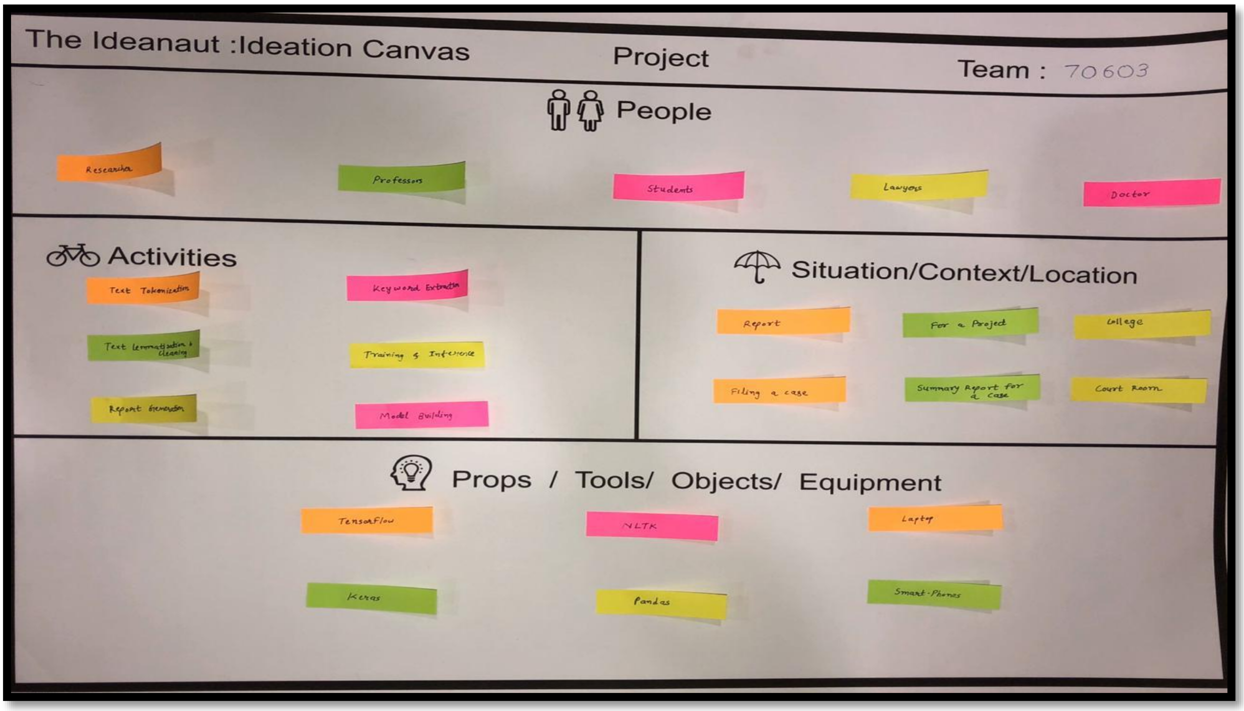
→ When an employee was working on a research-based project where he had to study number of research papers in a limited time. But he couldn't complete the project within given time because research papers were too lengthy to examine/study the flow of work. If he could use the text summarizer then he could examine that lengthy research paper & could complete the project within given time.

STUDENT BOOK STALL - V. V. Nagar M : 98244 46491

Empathy Canvas

2.3 IDEATION CANVAS

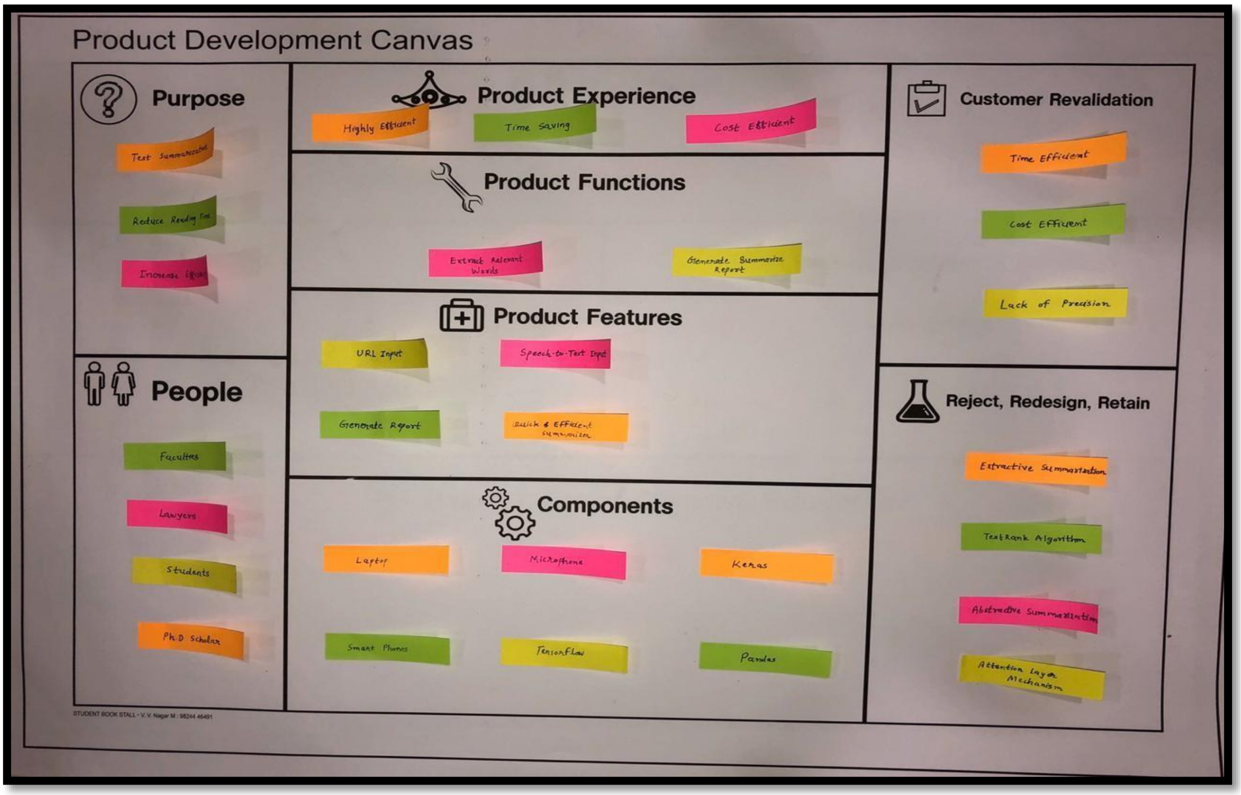
Ideation canvas helped to identify real users other than the users in empathy canvas. Also the situations or problems the users or developer may face were identified. We also listed out the possible solutions to it. We explored various approaches to innovative thinking and techniques for idea generation from a range of sources.



Ideation Canvas

2.4 PRODUCT DEVELOPMENT CANVAS:

In this canvas, we could define the purpose of our system, how it will benefit the users, how it functions, its features, the components used and certain features that should be redesigned and retained. Purpose defines the reason for using the system. People means the main users of the system. Components defines the tools used in building the system. Customer revalidation means what were the reviews of customers after using the system.



Product Development Canvas

Purpose
Text Summarization
Reduce Reading Time
Increase Efficiency

People
Faculties
Lawyers
Students
Ph.D Scholar

Product Experience
Highly Efficient
Time Saving
Cost Efficient

Product Functions
Extract Relevant Words
Generate Summarize Report

Product Features
URL Input
Speech-to-Text Input
Generate Report
Quick & Efficient Summarizer

Components
Laptop
Microphone
Keras
Smart Phones
TensorFlow
Pandas

Customer Revalidation
Time Efficient
Cost Efficient
Lack of Precision

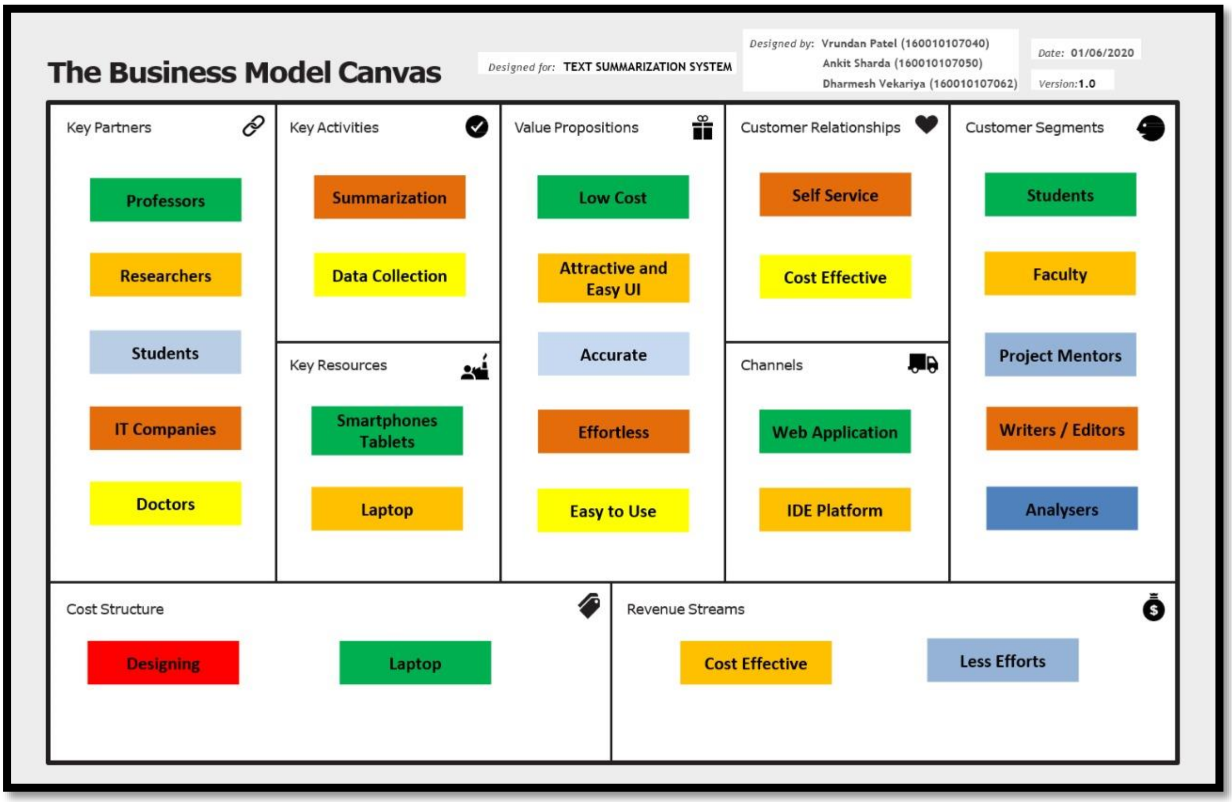
Reject, Redesign, Retain
Extractive Summarization
TextRank Algorithm
Abstractive Summarization
Attention Layer Mechanism

STUDENT BOOK STALL - V. V. Nagar M : 98244 46491

Product Development Canvas

2.5 BUSINESS MODEL CANVAS:

Business Model Canvas is used to validate the market significance of products and services which will be of technology nature in this case. Technology projects are often solutions or processes that solve a technical problem. However the market implementation of such solutions also require that the problem solution is designed to overcome not just the technical barriers but also market and business related barriers of costs and customer. Key partners for this projects are students, researchers, law-firms, etc. Key activities are data retrieval and text summarization. Value proposition of system is its negligible cost, time saving and accurate summary generation capability.



The Business Model Canvas

Designed for: TEXT SUMMARIZATION SYSTEM

Designed by: Vrundan Patel (160010107040)
Ankit Sharda (160010107050)
Dharmesh Vekariya (160010107062)

Date: 01/06/2020

*Version:*1.0

Key Partners
Professors
Researchers
Students
IT Companies
Doctors

Key Activities
Summarization
Data Collection

Key Resources
Smartphones Tablets
Laptop

Value Propositions
Low Cost
Attractive and Easy UI
Accurate
Effortless
Easy to Use

Customer Relationships
Self Service
Cost Effective

Channels
Web Application
IDE Platform

Customer Segments
Students
Faculty
Project Mentors
Writers / Editors
Analysers

Cost Structure
Designing
Laptop

Revenue Streams
Cost Effective
Less Efforts

Business Model Canvas

3. Implementation

Step-1:

Importing necessary libraries and initializing WordNetLemmatizer.

Step-2: Text pre-processing:

The pre-processing steps applied in this algorithm include, removing special characters, digits and one letter word and stop words from the text.

Step-3: Calculating the frequency of each word in the document.

While working with text it becomes important to calculate the frequency of words, to find the most common or least common words based on the requirement of the algorithm.

Step-4: Calculating sentence score.

In this approach, we will be using TFIDF score of each word to calculate the total sentence score.

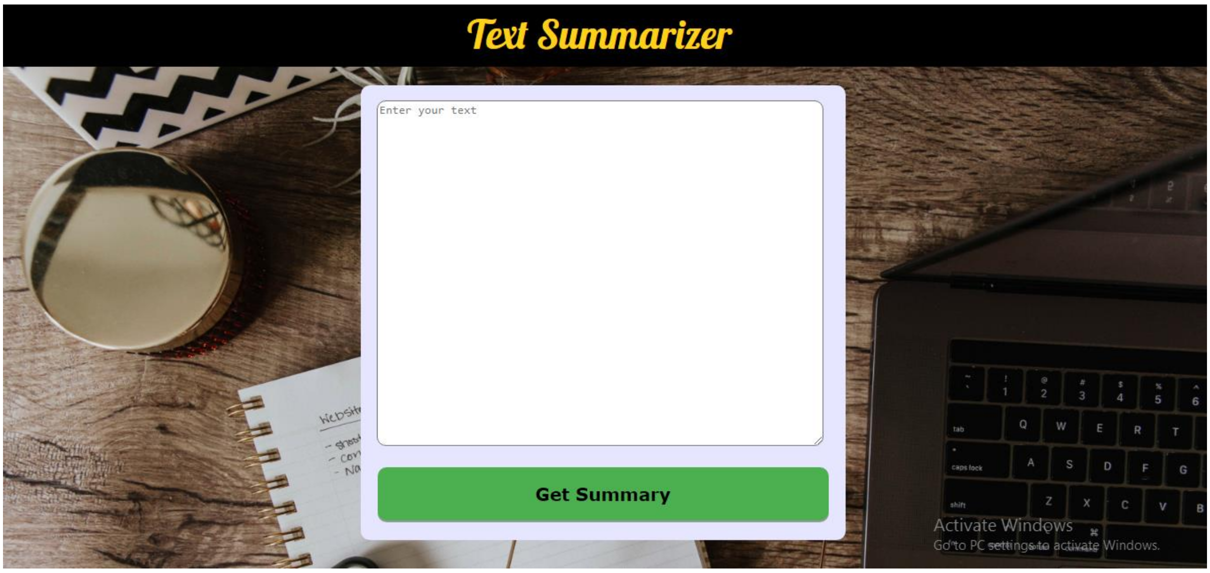
- **TF(w)** = (Number of times term w appears in a document) / (Total number of terms in the document)

- **IDF(w)** = log_e(Total number of documents / Number of documents with term w in it)

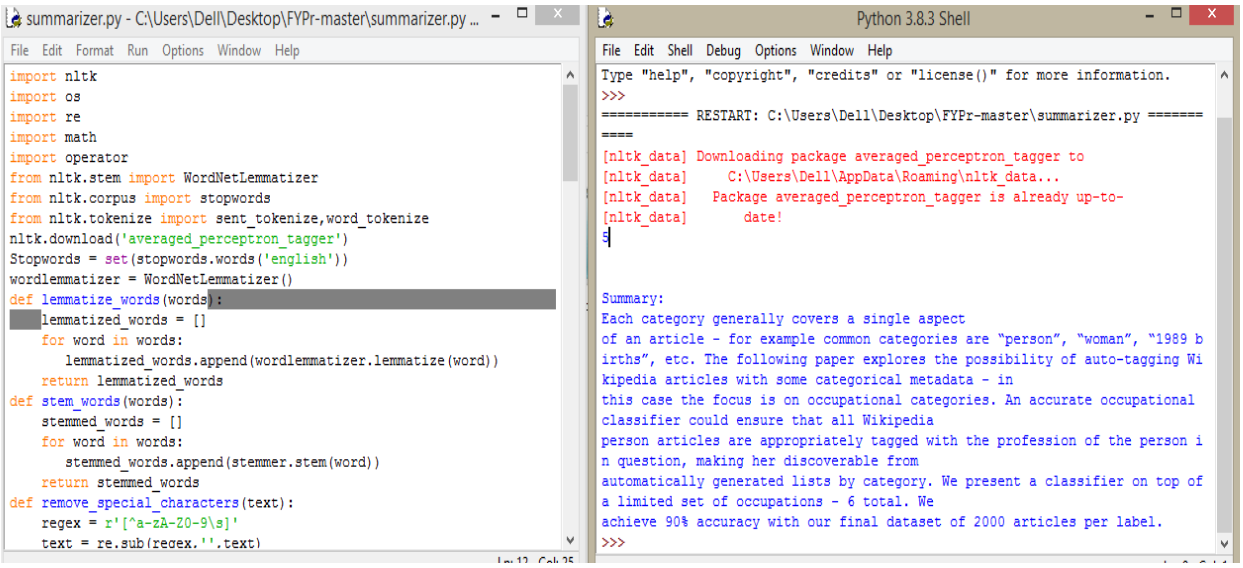
- **TFIDF(w) = TF(w) * IDF(w)**

Step-5: Finding most important sentences

To find the most important sentences, take the individual sentences from tokenized sentences and compute the sentence score. After calculating the scores, the top sentences based on the retention rate provided by the user are included in the summary.



User Interface


summarizer.py - C:\Users\Dell\Desktop\FYPr-master\summarizer.py ...
File Edit Format Run Options Window Help
Python 3.8.3 Shell
File Edit Shell Debug Options Window Help

```
import nltk
import os
import re
import math
import operator
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize,word_tokenize
nltk.download('averaged_perceptron_tagger')
Stopwords = set(stopwords.words('english'))
wordlemmatizer = WordNetLemmatizer()
def lemmatize_words(words):
    lemmatized_words = []
    for word in words:
        lemmatized_words.append(wordlemmatizer.lemmatize(word))
    return lemmatized_words
def stem_words(words):
    stemmed_words = []
    for word in words:
        stemmed_words.append(stemmer.stem(word))
    return stemmed_words
def remove_special_characters(text):
    regex = r'[^a-zA-Z0-9\s]'
    text = re.sub(regex,'',text)
```

```
Type "help", "copyright", "credits" or "license()" for more information.
>>>
=========== RESTART: C:\Users\Dell\Desktop\FYPr-master\summarizer.py =======
====
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\Dell\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
5


Summary:
Each category generally covers a single aspect
of an article - for example common categories are “person”, “woman”, “1989 b
irths”, etc. The following paper explores the possibility of auto-tagging Wi
kipedia articles with some categorical metadata - in
this case the focus is on occupational categories. An accurate occupational
classifier could ensure that all Wikipedia
person articles are appropriately tagged with the profession of the person i
n question, making her discoverable from
automatically generated lists by category. We present a classifier on top of
a limited set of occupations - 6 total. We
achieve 90% accuracy with our final dataset of 2000 articles per label.
>>>
```

Code Output

4. Summary

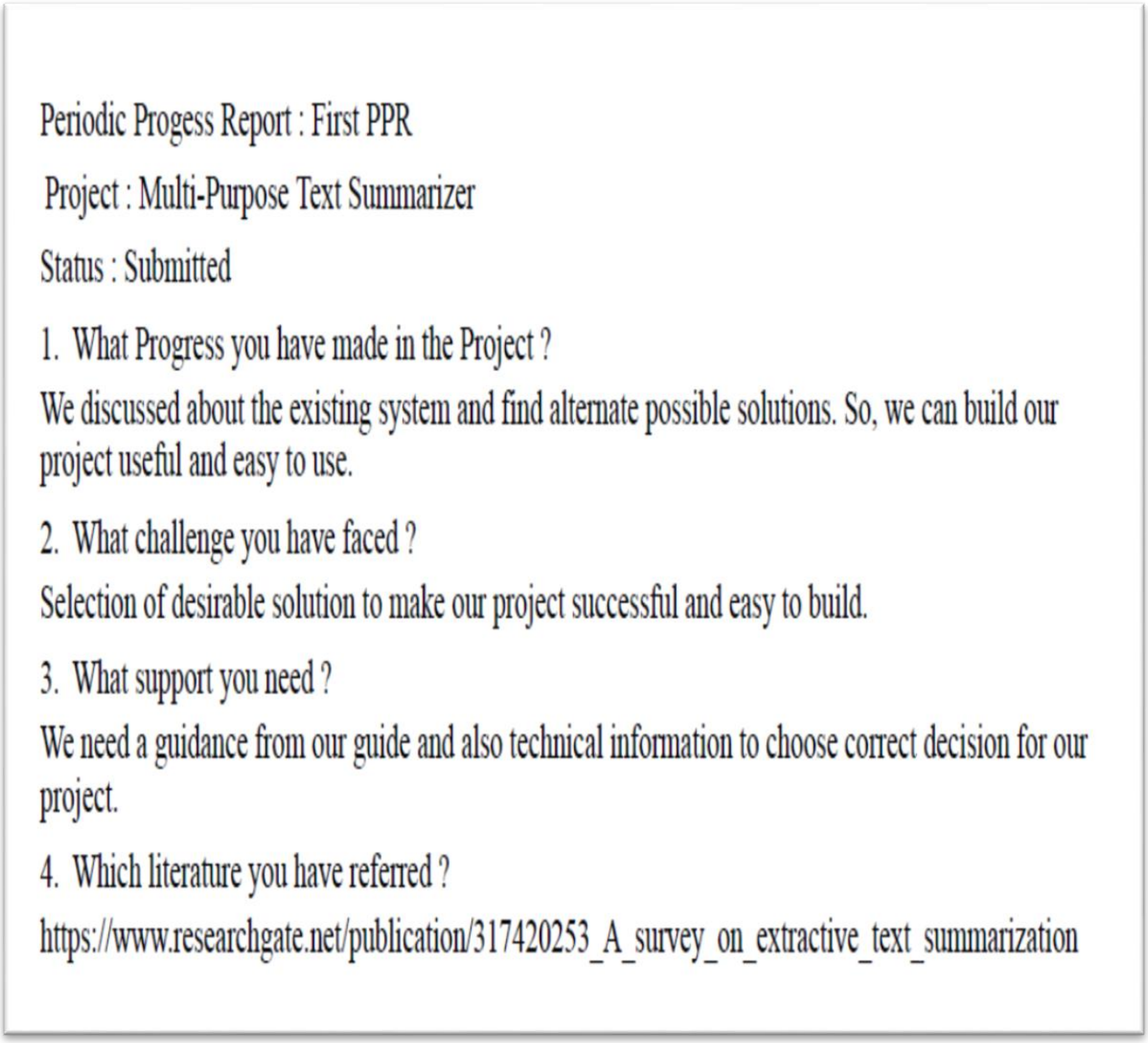
The main aim of this project is to provide concise and to the point summary of large text documents. It will save lot of valuable human hours since they can get the required information from given document in very short time. The project described here uses TF-IDF approach for ranking sentences based on their TF-IDF score of each sentence. TF-IDF, short for term frequency–inverse document frequency, is a numeric measure that is use to score the importance of a word in a document based on how often did it appear in that document and a given collection of documents. After calculating the scores, the top sentences based on the retention rate provided by the user are included in the summary. This is an extractive method of text summarization.

5. References

- https://medium.com/voice-tech-podcast/automatic-extractive-text-summarization-using-tfidf3fc9a7b26f5
- https://www.nltk.org/
- https://towardsdatascience.com/text-summarization-extractive-approach-567fe4b85c23
- https://www.sciencedirect.com/science/article/pii/S0950705119303235
- https://patents.google.com/patent/US20040117725A1/en?q=text+summarization&oq=text+summarization
- https://patents.google.com/patent/US7451395B2/en?q=text+summarization&oq=text+summarization
- https://www.transcrypt.org/
- https://yasoob.me/2019/05/22/running-python-in-the-browser/

APPENDIX:

- **PPR1**



Periodic Progess Report : First PPR

Project : Multi-Purpose Text Summarizer

Status : Submitted

1. What Progress you have made in the Project ?

We discussed about the existing system and find alternate possible solutions. So, we can build our project useful and easy to use.

2. What challenge you have faced ?

Selection of desirable solution to make our project successful and easy to build.

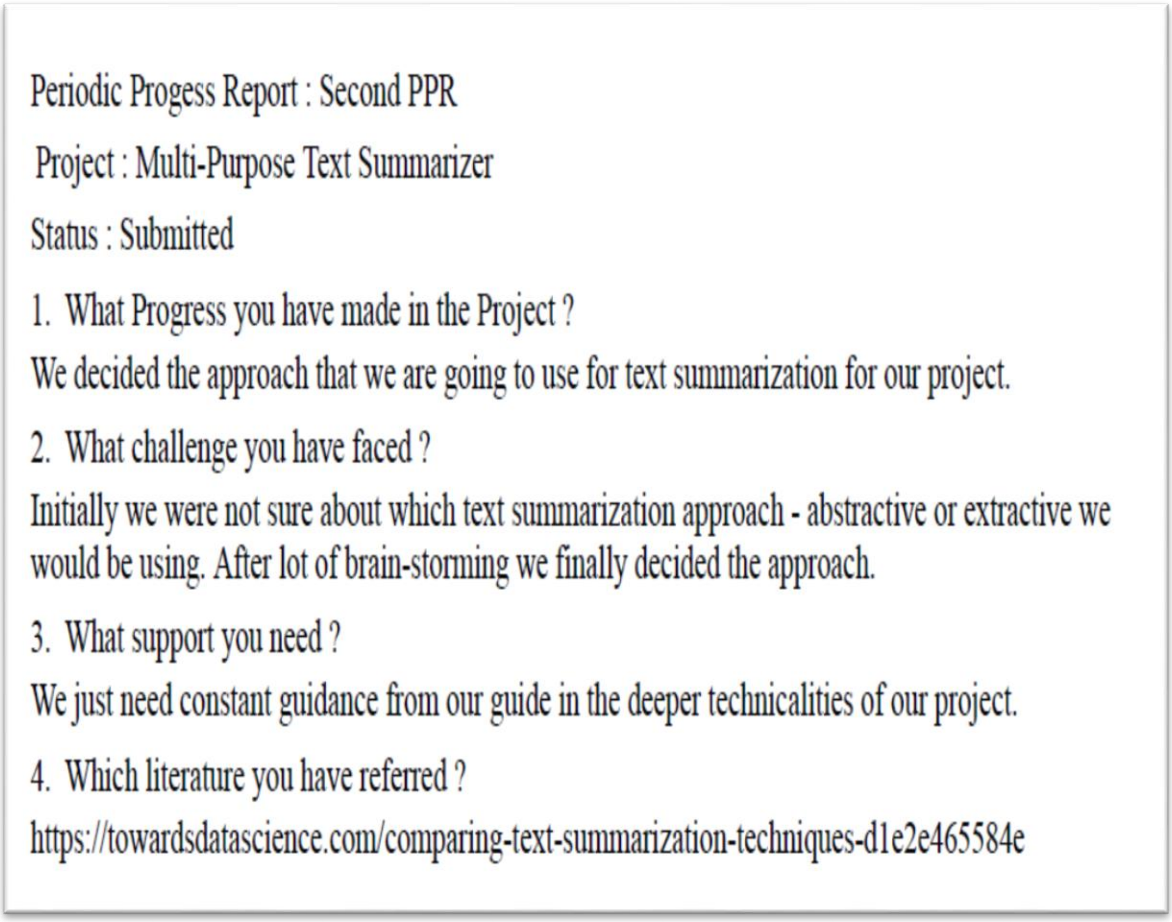
3. What support you need ?

We need a guidance from our guide and also technical information to choose correct decision for our project.

4. Which literature you have referred ?

https://www.researchgate.net/publication/317420253_A_survey_on_extractive_text_summarization

- **PPR2**



Periodic Progess Report : Second PPR

Project : Multi-Purpose Text Summarizer

Status : Submitted

1. What Progress you have made in the Project ?

We decided the approach that we are going to use for text summarization for our project.

2. What challenge you have faced ?

Initially we were not sure about which text summarization approach - abstractive or extractive we would be using. After lot of brain-storming we finally decided the approach.

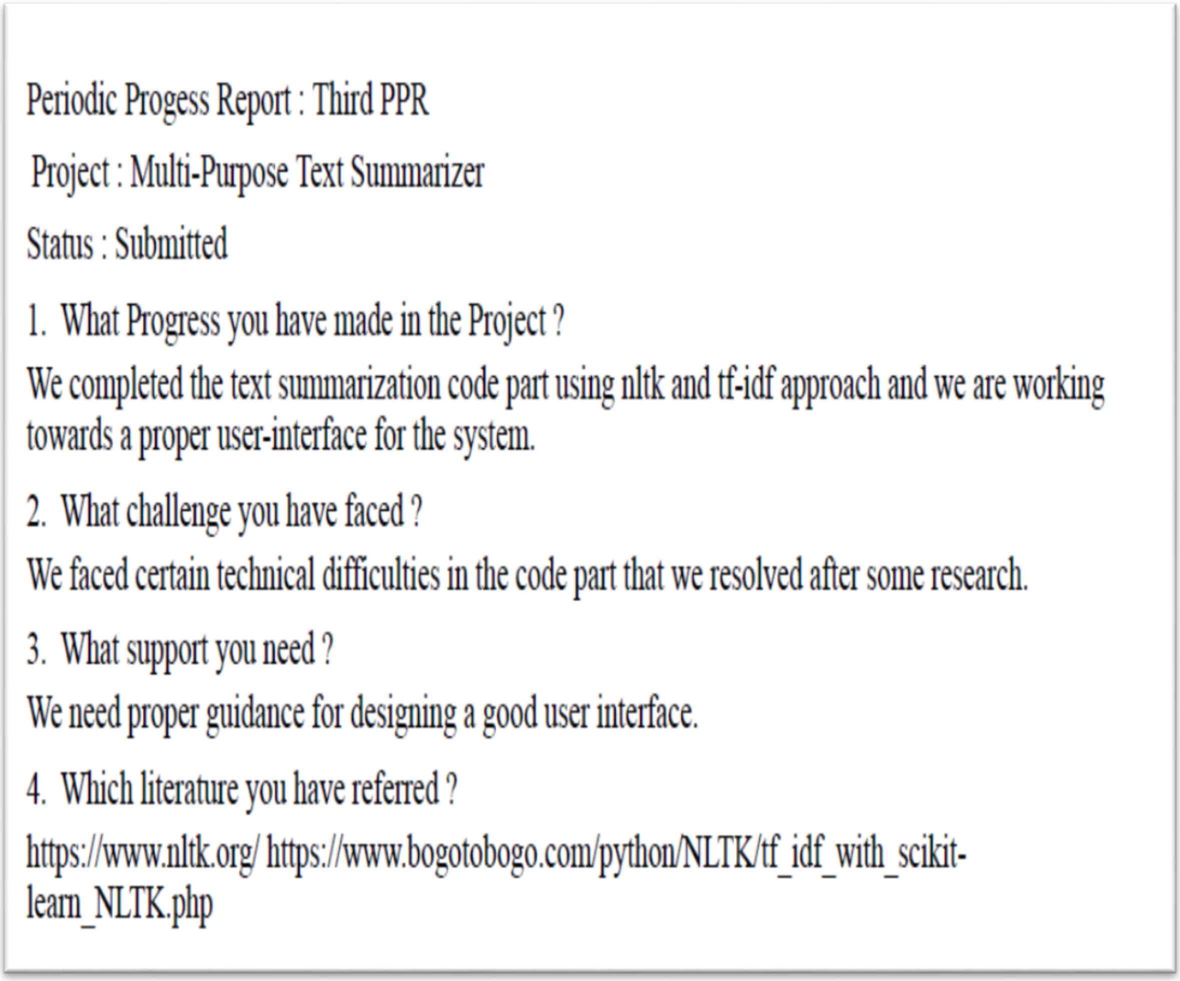
3. What support you need ?

We just need constant guidance from our guide in the deeper technicalities of our project.

4. Which literature you have referred ?

https://towardsdatascience.com/comparing-text-summarization-techniques-d1e2e465584e

- **PPR3**



Periodic Progess Report : Third PPR

Project : Multi-Purpose Text Summarizer

Status : Submitted

1. What Progress you have made in the Project ?

We completed the text summarization code part using nltk and tf-idf approach and we are working towards a proper user-interface for the system.

2. What challenge you have faced ?

We faced certain technical difficulties in the code part that we resolved after some research.

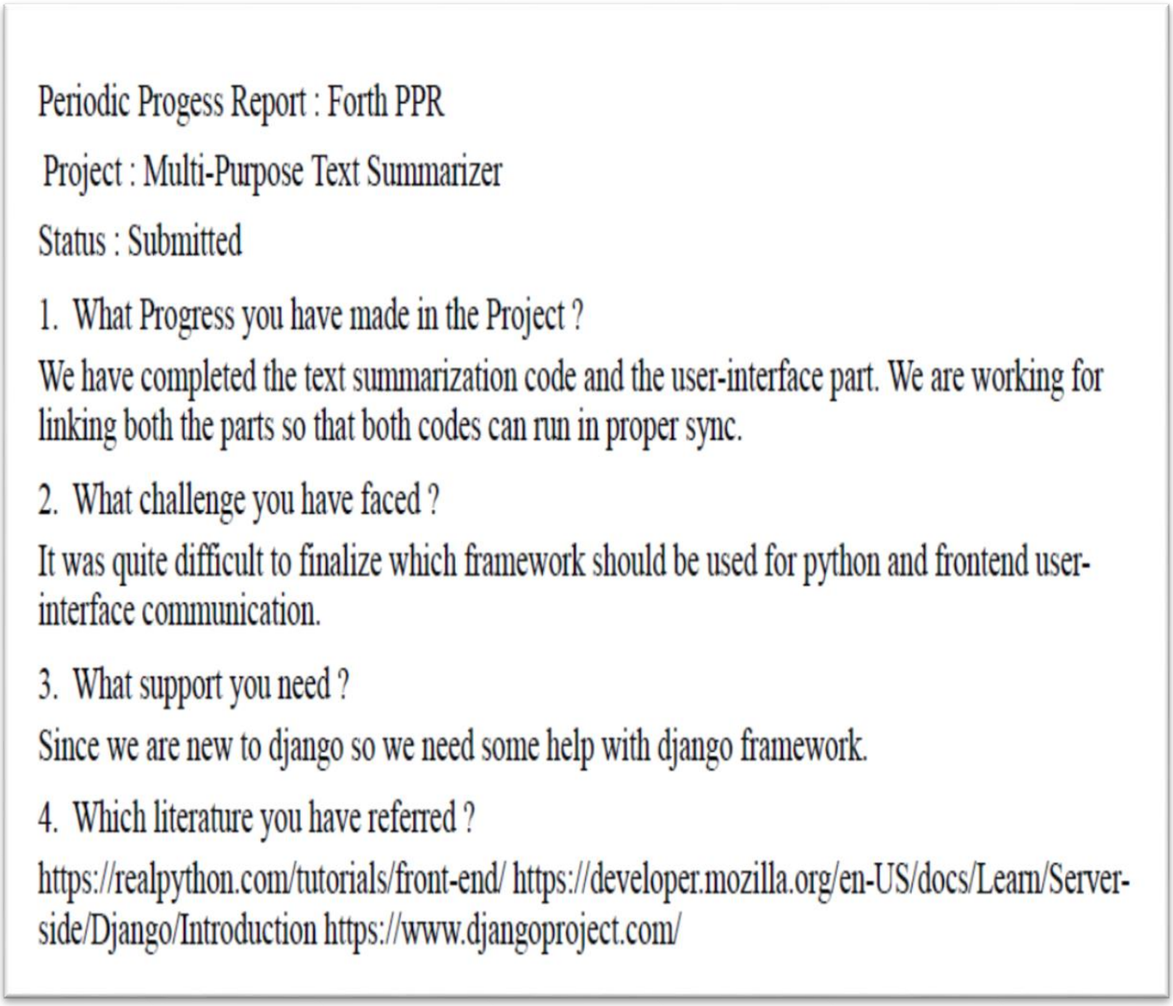
3. What support you need ?

We need proper guidance for designing a good user interface.

4. Which literature you have referred ?

https://www.nltk.org/ https://www.bogotobogo.com/python/NLTK/tf_idf_with_scikit-learn_NLTK.php

- **PPR4**



Periodic Progess Report : Forth PPR

Project : Multi-Purpose Text Summarizer

Status : Submitted

1. What Progress you have made in the Project ?

We have completed the text summarization code and the user-interface part. We are working for linking both the parts so that both codes can run in proper sync.

2. What challenge you have faced ?

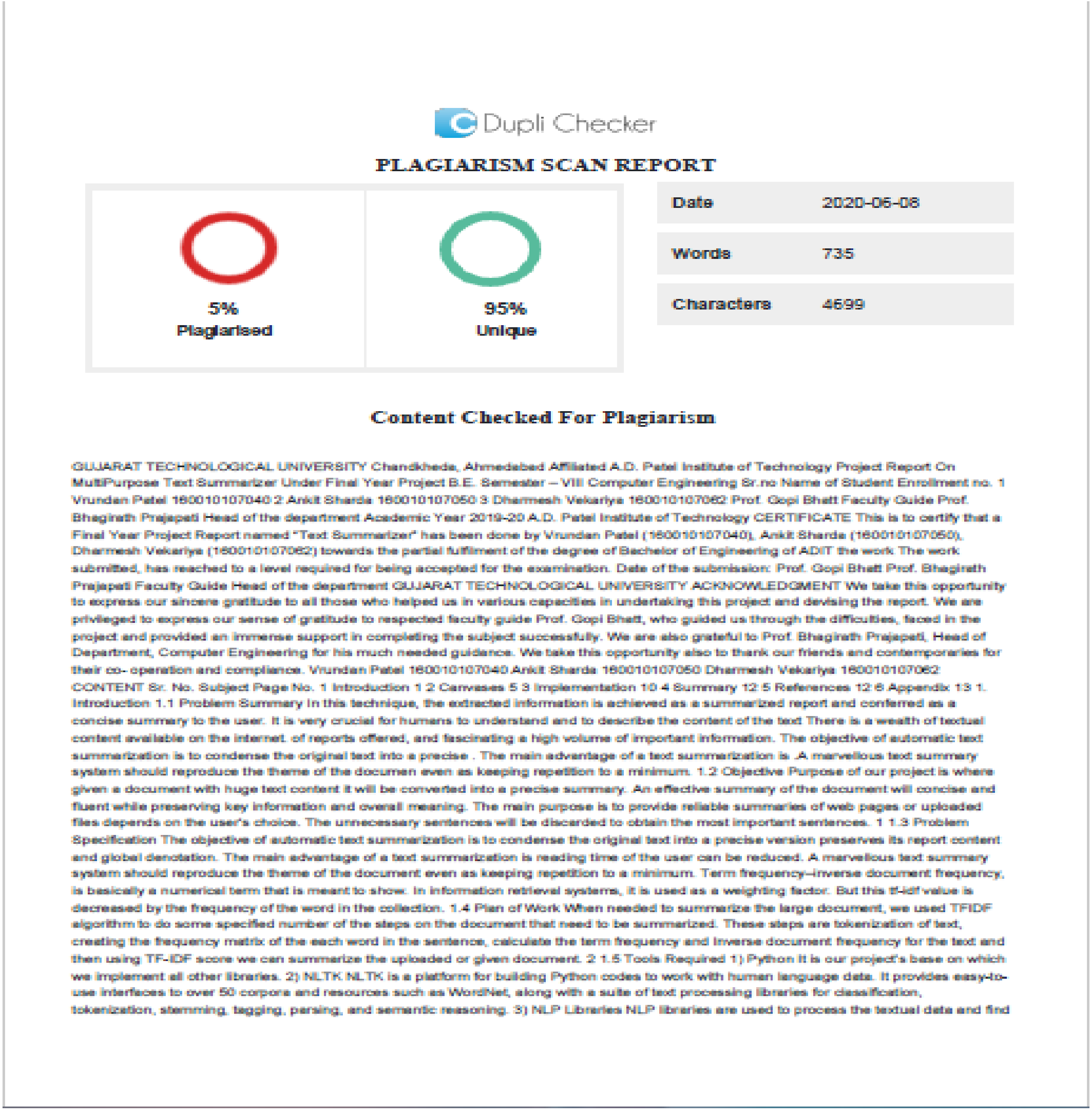
It was quite difficult to finalize which framework should be used for python and frontend user-interface communication.

3. What support you need ?

Since we are new to django so we need some help with django framework.

4. Which literature you have referred ?

https://realpython.com/tutorials/front-end/ https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django/Introduction https://www.djangoproject.com/



Dupli Checker

PLAGIARISM SCAN REPORT

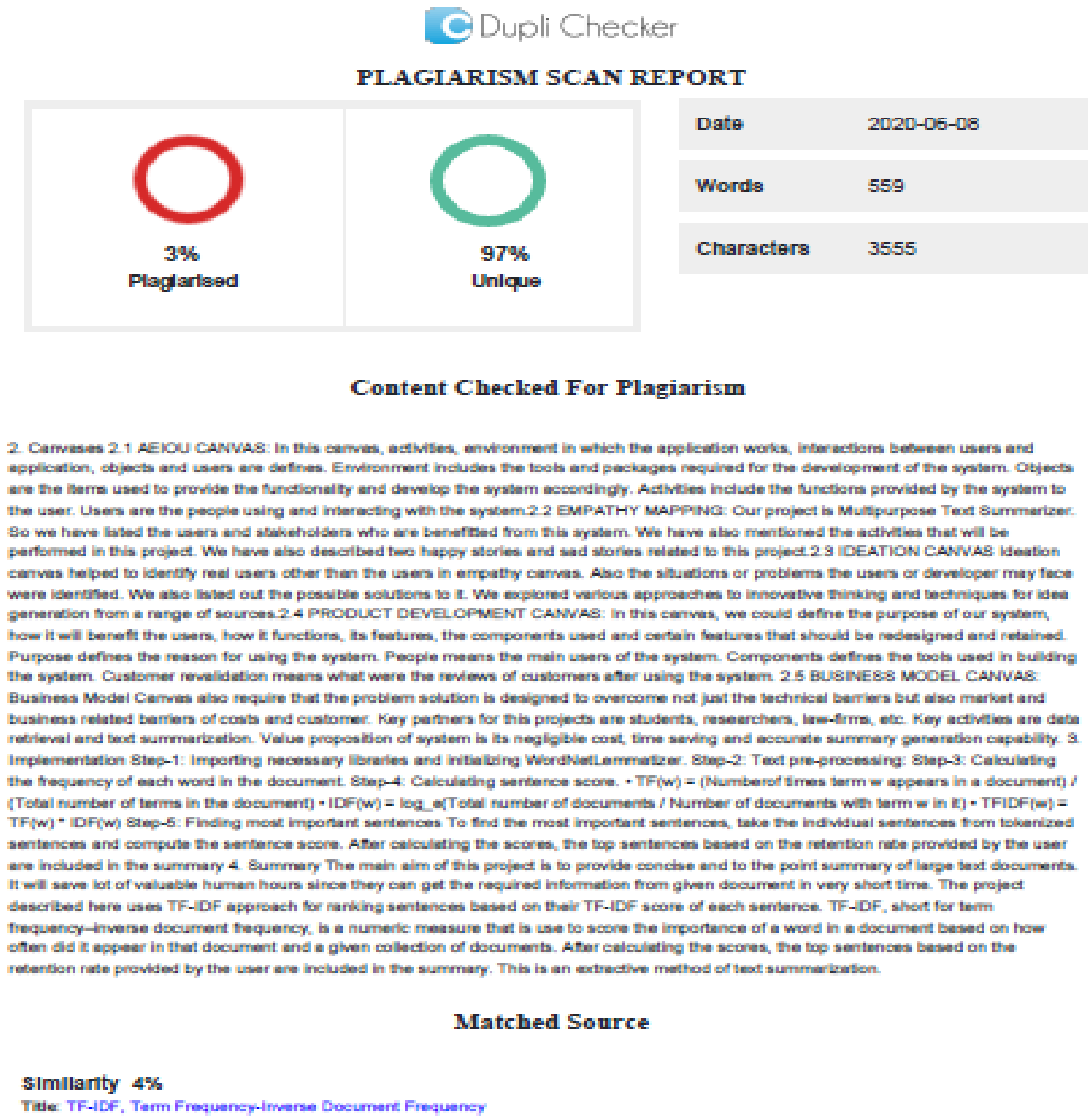
5%
Plagiarised

95%
Unique

Date	2020-06-08
Words	735
Characters	4699

Content Checked For Plagiarism

GUJARAT TECHNOLOGICAL UNIVERSITY Chandkheda, Ahmedabad Affiliated A.D. Patel Institute of Technology Project Report On MultiPurpose Text Summarizer Under Final Year Project B.E. Semester – VIII Computer Engineering Sr.no Name of Student Enrollment no. 1 Vrundan Patel 160010107040 2 Ankit Sharda 160010107050 3 Dharmesh Vekariya 160010107062 Prof. Gopi Bhatt Faculty Guide Prof. Bhagirath Prajapati Head of the department Academic Year 2019-20 A.D. Patel Institute of Technology CERTIFICATE This is to certify that a Final Year Project Report named "Text Summarizer" has been done by Vrundan Patel (160010107040), Ankit Sharda (160010107050), Dharmesh Vekariya (160010107062) towards the partial fulfilment of the degree of Bachelor of Engineering of ADIT the work The work submitted, has reached to a level required for being accepted for the examination. Date of the submission: Prof. Gopi Bhatt Prof. Bhagirath Prajapati Faculty Guide Head of the department GUJARAT TECHNOLOGICAL UNIVERSITY ACKNOWLEDGMENT We take this opportunity to express our sincere gratitude to all those who helped us in various capacities in undertaking this project and devising the report. We are privileged to express our sense of gratitude to respected faculty guide Prof. Gopi Bhatt, who guided us through the difficulties, faced in the project and provided an immense support in completing the subject successfully. We are also grateful to Prof. Bhagirath Prajapati, Head of Department, Computer Engineering for his much needed guidance. We take this opportunity also to thank our friends and contemporaries for their co- operation and compliance. Vrundan Patel 160010107040 Ankit Sharda 160010107050 Dharmesh Vekariya 160010107062 CONTENT Sr. No. Subject Page No. 1 Introduction 1 2 Canvases 5 3 Implementation 10 4 Summary 12 5 References 12 6 Appendix 13 1. Introduction 1.1 Problem Summary In this technique, the extracted information is achieved as a summarized report and conferred as a concise summary to the user. It is very crucial for humans to understand and to describe the content of the text There is a wealth of textual content available on the internet. of reports offered, and fascinating a high volume of important information. The objective of automatic text summarization is to condense the original text into a precise . The main advantage of a text summarization is .A marvellous text summary system should reproduce the theme of the documen even as keeping repetition to a minimum. 1.2 Objective Purpose of our project is where given a document with huge text content it will be converted into a precise summary. An effective summary of the document will concise and fluent while preserving key information and overall meaning. The main purpose is to provide reliable summaries of web pages or uploaded files depends on the user's choice. The unnecessary sentences will be discarded to obtain the most important sentences. 1 1.3 Problem Specification The objective of automatic text summarization is to condense the original text into a precise version preserves its report content and global denotation. The main advantage of a text summarization is reading time of the user can be reduced. A marvellous text summary system should reproduce the theme of the document even as keeping repetition to a minimum. Term frequency–inverse document frequency, is basically a numerical term that is meant to show. In information retrieval systems, it is used as a weighting factor. But this tf-idf value is decreased by the frequency of the word in the collection. 1.4 Plan of Work When needed to summarize the large document, we used TFIDF algorithm to do some specified number of the steps on the document that need to be summarized. These steps are tokenization of text, creating the frequency matrix of the each word in the sentence, calculate the term frequency and Inverse document frequency for the text and then using TF-IDF score we can summarize the uploaded or given document. 2 1.5 Tools Required 1) Python It is our project's base on which we implement all other libraries. 2) NLTK NLTK is a platform for building Python codes to work with human language data. It provides easy-to-use interfaces to over 50 corpora and resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. 3) NLP Libraries NLP libraries are used to process the textual data and find



Dupli Checker

PLAGIARISM SCAN REPORT

3%
Plagiarised

97%
Unique

Date	2020-06-08
Words	559
Characters	3555

Content Checked For Plagiarism

2. Canvases 2.1 AEIOU CANVAS: In this canvas, activities, environment in which the application works, interactions between users and application, objects and users are defines. Environment includes the tools and packages required for the development of the system. Objects are the items used to provide the functionality and develop the system accordingly. Activities include the functions provided by the system to the user. Users are the people using and interacting with the system.2.2 EMPATHY MAPPING: Our project is Multipurpose Text Summarizer. So we have listed the users and stakeholders who are benefitted from this system. We have also mentioned the activities that will be performed in this project. We have also described two happy stories and sad stories related to this project.2.3 IDEATION CANVAS: Ideation canvas helped to identify real users other than the users in empathy canvas. Also the situations or problems the users or developer may face were identified. We also listed out the possible solutions to it. We explored various approaches to innovative thinking and techniques for idea generation from a range of sources.2.4 PRODUCT DEVELOPMENT CANVAS: In this canvas, we could define the purpose of our system, how it will benefit the users, how it functions, its features, the components used and certain features that should be redesigned and retained. Purpose defines the reason for using the system. People means the main users of the system. Components defines the tools used in building the system. Customer revalidation means what were the reviews of customers after using the system. 2.5 BUSINESS MODEL CANVAS: Business Model Canvas also require that the problem solution is designed to overcome not just the technical barriers but also market and business related barriers of costs and customer. Key partners for this projects are students, researchers, law-firms, etc. Key activities are data retrieval and text summarization. Value proposition of system is its negligible cost, time saving and accurate summary generation capability. 3. Implementation Step-1: Importing necessary libraries and initializing WordNetLemmatizer. Step-2: Text pre-processing: Step-3: Calculating the frequency of each word in the document. Step-4: Calculating sentence score. • TF(w) = (Numberof times term w appears in a document) / (Total number of terms in the document) • IDF(w) = log_e(Total number of documents / Number of documents with term w in it) • TFIDF(w) = TF(w) * IDF(w) Step-5: Finding most important sentences To find the most important sentences, take the individual sentences from tokenized sentences and compute the sentence score. After calculating the scores, the top sentences based on the retention rate provided by the user are included in the summary 4. Summary The main aim of this project is to provide concise and to the point summary of large text documents. It will save lot of valuable human hours since they can get the required information from given document in very short time. The project described here uses TF-IDF approach for ranking sentences based on their TF-IDF score of each sentence. TF-IDF, short for term frequency–inverse document frequency, is a numeric measure that is use to score the importance of a word in a document based on how often did it appear in that document and a given collection of documents. After calculating the scores, the top sentences based on the retention rate provided by the user are included in the summary. This is an extractive method of text summarization.

Matched Source

Similarity 4%

Title: TF-IDF, Term Frequency-Inverse Document Frequency