A

PROJECT REPORT

FOR

SUBJECT: LAB II- PROJECT PHASE II

ON

"Relevance ranked QnA system for live YouTube's embedded stream on an RWA"

Submitted in partial fulfillment of the requirement for the award of

Bachelor of Engineering

In

Computer Science and Engineering

Punyashlok Ahilyadevi Holkar Solapur University

By

Name	Roll. No.	Exam Seat No.
Anjali Chougule	BE CSE-40	
Janvi Pampattiwar	BE CSE-41	
Pragati Phand	BE CSE-42	
DurreAfshan Shaikh	BE CSE-43	

Under Guidance Of

Mr. P.S.R.Patnaik

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

WALCHAND INSTITUTE OF TECHNOLOGY

SOLAPUR - 413006

(2020-2021)

Certificate

This is to certify that the project entitled

"Relevance ranked QnA system for live YouTube's embedded stream on an RWA"

Is submitted by

Name	Roll. No.	Exam Seat No.
Anjali Chougule	BE CSE-40	
Janvi Pampattiwar	BE CSE-41	
Pragati Phand	BE CSE-42	
DurreAfshan Shaikh	BE CSE-43	

Mr.P.S.R.Patnaik   **Dr. Mrs. A.M.Pujar**
Project Guide   **Head CSE Dept**

Dr. S. A. Halkude
Principal

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
WALCHAND INSTITUTE OF TECHNOLOGY
SOLAPUR - 413006
(2020-2021)

Project Approval Sheet

The Project Entitled

"Relevance ranked QnA system for live YouTube's embedded stream on an RWA"

Submitted by

Name	Roll. No.	Exam Seat No.
Anjali Chougule	BE CSE-40	
Janvi Pampattiwar	BE CSE-41	
Pragati Phand	BE CSE-42	
DurreAfshan Shaikh	BE CSE-43	

"Is hereby approved in partial fulfillment for the degree of
Bachelor of Computer Science and Engineering"

Mr.P.S.R.Patnaik
Project Guide

Dr. Mrs. A.M.Pujar
Head CSE Dept

Dr. S. A. Halkude
Principal

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
WALCHAND INSTITUTE OF TECHNOLOGY
SOLAPUR - 413006
(2020 - 2021)

Acknowledgment

At the outset, we would like to take this opportunity to express our deep gratitude to our guide **Mr. P. S. R. Patnaik** of the CSE Department for his guidance and moral support throughout this successful completion of our project.

We heartily thank **Dr. Mrs. A. M. Pujar**, Head of CSE Dept for her moral support and for promoting us through the completion of our project.

We would also like to thank our Principal **Dr. S. A. Halkude** and all staff members for their wholehearted co-operation in completing this project.

UNDERTAKING

We solemnly declare that project work presented in the report titled
"..

Relevance ranked QnA system for live YouTube's embedded stream on an RWA
...."
is solely my project work with no significant contribution from any other person except for the project guide. Small contribution/help wherever taken has been duly acknowledged and that complete report has been written by the members of the project group.

We understand the zero-tolerance policy of the WIT, Solapur, and University towards plagiarism. Therefore, we as Authors of the above-titled report declare that no portion of the report has been plagiarized and any material used as a reference is properly referred/cited.

We undertake that if found guilty of any formal plagiarism in the above-titled report even after award of the degree, WIT, Solapur and Solapur University reserves the rights to withdraw/revoke the degree granted and that WIT, Solapur and the University have the right to publish our name on the website on which names of students are placed who submitted plagiarized report.

Name	Exam Number	University PRN Number	Signature
Anjali Chougule			
Janvi Pampattiwar			
Pragati Phand			
DurreAfshan Shaikh			

Date: / /

Index

Abstract

Nowadays, all over the planet embedding live streaming YouTube video in our own responsive web site. Embedding videos is simply like making backlinks to your website. Like in SEO (Search Engine Optimization), embedding your videos in a very website behaves specifically sort of a back link and so facultative your videos to urge placed in program results and find a lot of views. A lot of number of views our videos receive, our video quality and complete image increase too. Not solely video quality, your product and website quality too increase so increasing your sales and profits.

Video embedding is the method of adding a video player to our web site victimisation of an internet video platform. There are several web sites that are building their own social media platforms, so it is as easy as copying and pasting a link. Video inserting works by adding associate degree embed code from your video hosting platform to the code of your web site. It permits you to integrate live streaming on our website.

Relevance ranking is to be provided for live chat of the live video stream. Relevance ranking is the process of sorting the chats so that those Questions which are most likely to be relevant to the topic are shown at the top of the chat window with the answer.

Introduction

In the real time all over the world live streaming video is embedded into their own responsive website. Since recent years have brought an increase in the popularity of video-sharing across hundreds of different platforms. This means that a number of people are sharing live and on demand videos regularly. Knowing how to broadcast live and embed live streaming video on your website is also becoming increasingly important for all kinds of broadcasters. People can not only watch live streaming video and live chat, but they can also leave blog post comments which mainly remain beyond the broadcast. People can also post supplemental information and links related to our livestream for viewers. If we stream using multiple platforms our blog is the place where fans can always find our latest broadcast.

As we know that many people make their own live streaming video. And the viewers who are watching their streaming they put or raise their question into the comment box. Since the questions which are put by the viewers are not in the relevance ranking so the host is not able to give each and every question in minimum time span.

So, we are developing our own responsive website so that we will provide a chat window which will display all the relevance ranked questions which will make the work of the host easier for answering the questions in minimum time span and also with the chat window we are also displaying the live streaming video.

Thus, this system aims to build a machine learning model that predicts the relevance of ranked questions.

Our system aims to give relevance ranked questions to the user on the topic related to streaming so that it makes the work of the host easier and also save his time from answering the same meaning questions.

To give a grouped and unique questions based on the relevance ranking, it is necessary to keep track of data such as:
- What type of data is taken from the comment box of the live streaming?
- What are the different types of questions asked by the viewer and there should be no emojis and different languages used?

Therefore, we have built a machine learning model which performs the task of giving relevance ranked questions based on the topic. So, we can build machine learning models for that. The built-in libraries used by the ML model for training are:

1]NLTK:

It is used for removing stop words.

2] Sklearn:

It is used for clustering.

3] Pytchat:

It is used for retrieving live chat from streaming.

2.1 Front End

2.1.1 HTML:

HTML is a standard markup language for creating web pages and web applications. It can be assisted by Cascading Style Sheets (CSS) and scripting languages like JavaScript. The front end of a project is built using HTML and CSS. The browser does not display the HTML tags but is utilized to build content of the page also used in plug-in development.

2.1.2 CSS:

CSS is used for styling the content of web pages including colors, layout, fonts. It will allow adapting the presentation of different CSS styles of documents written in a markup language like HTML.

2.1.3 jQuery:

jQuery is a fast, cross-platform, and feature-rich JavaScript library. Its main purpose is to provide an easy way to JavaScript on websites to make it more interactive, attractive, and also add animation. jQuery simplifies HTML document traversing, event handling, and Ajax interactions for rapid web development.

2.1.4 Bootstrap:

Bootstrap is a free front-end open-source HTML, CSS, JavaScript framework used for developing responsive web sites. It quickly designs and customizes mobile-first websites. It contains extensive prebuilt components, JavaScript plugins. It used to make the plugin front end more responsive.

2.2 Back End

2.2.1 Flask:

Flask is a popular lightweight python web application framework and baseband on the WSGI toolkit. Flask provides a simple template to build web applications. Flask can be used to save time building web applications after being imported into python. It has no database, abstraction layer, or form validation or any other components but flask supports extensions. Flask is used to build the REST API of the project with Oauth1. Oauth1 is Authentication level 1 used for authentication purposes.

Features of Flask:
- Integrated support for unit testing.
- Extensive documentation
- Unicode based

2.2.2 Python:

Python is a powerful, easy to learn programming language which contains high-level data structures. It can also be used to create web applications. It runs on different platforms friendly, has a simple syntax, and allows developers to write a program with fewer lines. It makes code short and versatile.

2.2.3 NumPy:

Numerical Python (NumPy) is a library consisting of multidimensional array objects and collections of routines used to manipulate those arrays. NumPy is the most fundamental and effective package deal for running with facts in python. It is handy for mathematical and logical operations and manipulation of arrays. And also used to perform operations on data.

2.2.4 PHP:

Personal Home Page (PHP) is a server-side scripting language used to develop static or dynamic web pages. PHP code may be embedded with HTML code. Used in combination with various web templates, web frameworks. PHP inbuilt has support for working hand in hand with MySQL. PHP is a cross-platform scripting language so it can work on different operating systems. Used for WordPress and WooCommerce Plugin development.

2.2.5 firebase(database):

Store and sync data with our NoSQL cloud database. Data is synced across all clients in real time, and remains available when your app goes offline.

The Firebase Realtime Database is a cloud-hosted database. Data is stored as JSON and synchronized in real-time to every connected client. When you build cross-platform apps with our iOS, Android, and JavaScript SDKs, all of your clients share one Realtime Database instance and automatically receive updates with the newest data.

Methodology

3.1 System Architecture Overview (High-level Architecture):
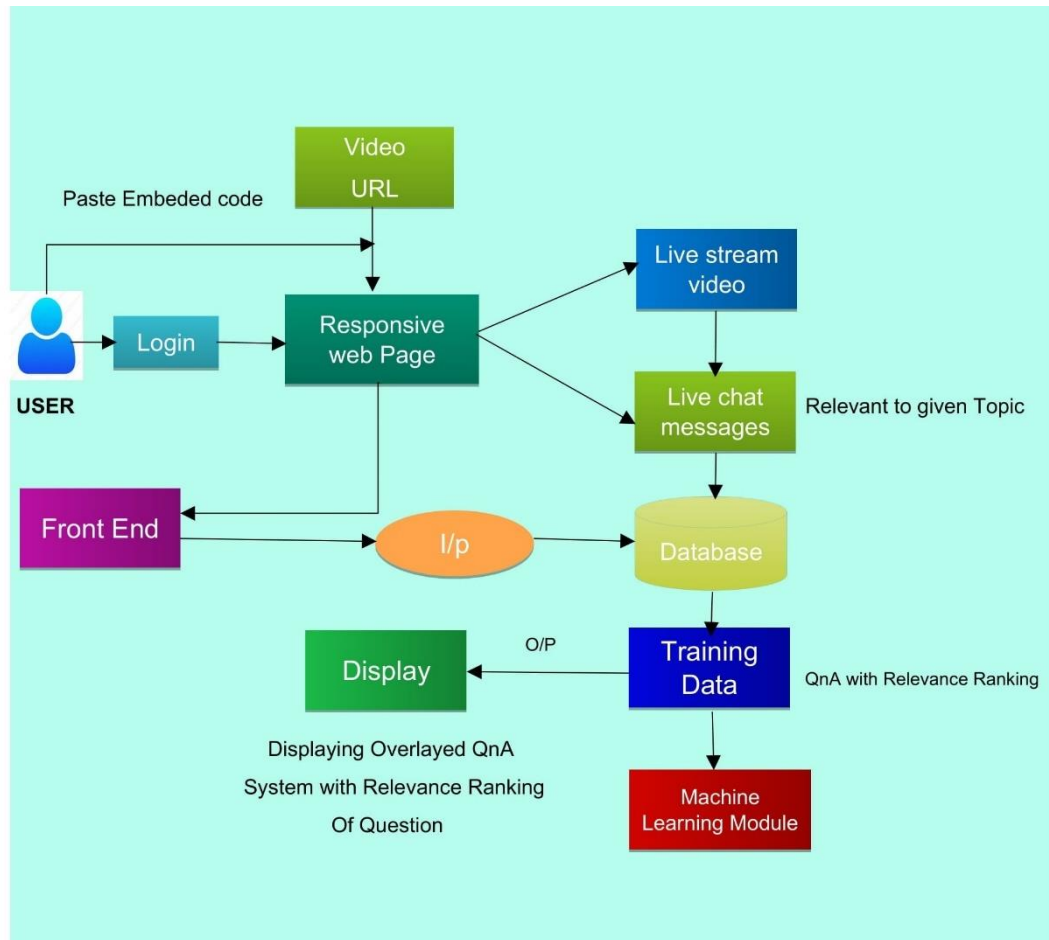
3.1.1 Data Collection:

Data should be collected from YouTube's live chat for pre-processing. The messages are collected from YouTube in systematic manner.

3.1.2 Data Pre-processing:

Real-world data is often incomplete, inconsistent, and lacking in certain behaviour trends and likely contains many errors. Data processing includes transforming raw data into an understandable format.

3.1.3 Prepared Data:

After data collection and data pre-processing the prepared data saved in the required format.

3.1.4 Splitting Data:

Partition of data into training data and test data. Splitting training data in
Training data and validation data.

3.1.5 Machine Learning Model:

ML model built using the ML algorithm. The different algorithms are suitable to solve the problems used and a model trained on the training dataset using that algorithm.

The algorithm giving a good performance is chosen.

3.1.6 NLTK (Natural Language ToolKit):

This library has been used in our project for removing stop words from the questions which have been retrieved from live chat. Stop words refers to the most common words that appear in our language. We see them often and are redundant, so it provides no real information. Hence, we remove them completely.

3.1.7 Ski kit learn (Sklearn):

This library has been used in our project for clustering. For clustering we have K-mean clustering algorithm. The k-means clustering algorithm is an unsupervised clustering algorithm which determines the optimal number of clusters using the elbow method. It works iteratively by selecting a random coordinate of the cluster center and assign the data points to a cluster.

3.2 Low-Level Architecture

3.2.1 Machine Learning Model

3.2.1.1 Data Exploration:

As the data required to train machine learning module is live chat fetched from YouTube stream. This fetched data consists of the emoji's, symbols, notations and so on. To improve the model efficiency, we need to remove all these unwanted data. Also, there is need to remove the stop words like and, or, if, of, etc which may change the meaning of the sentences.

3.2.1.2 Data Pre-processing:

The data provided to machine learning model is chats fetched from live YouTube video. This fetched data contains unwanted symbols, emoji's, notations and stop words which make model less efficient. To remove these unwanted data, we have used the machine learning algorithm.
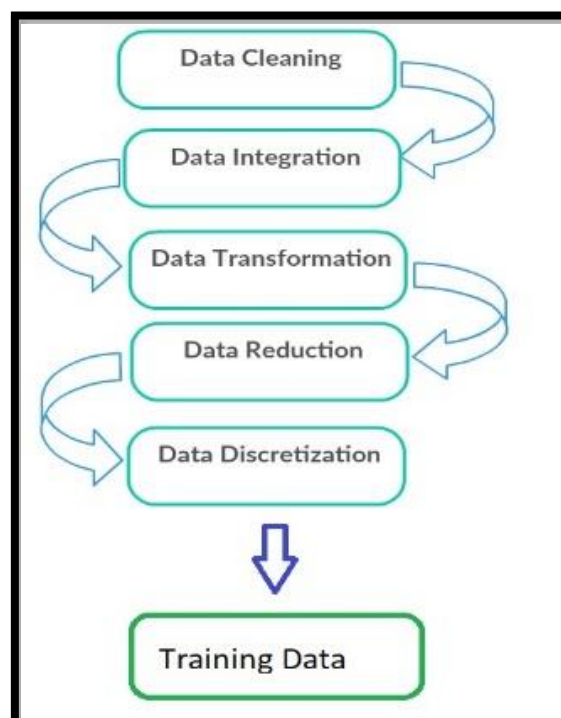
Fig. Data Pre-processing

3.2.1.3 Model Building:

Starting with model building, pytchat library is used for data fetching and different algorithms are applied over data and checked for accuracy of each. Different algorithms are considered and applied over it and checked for accuracy.
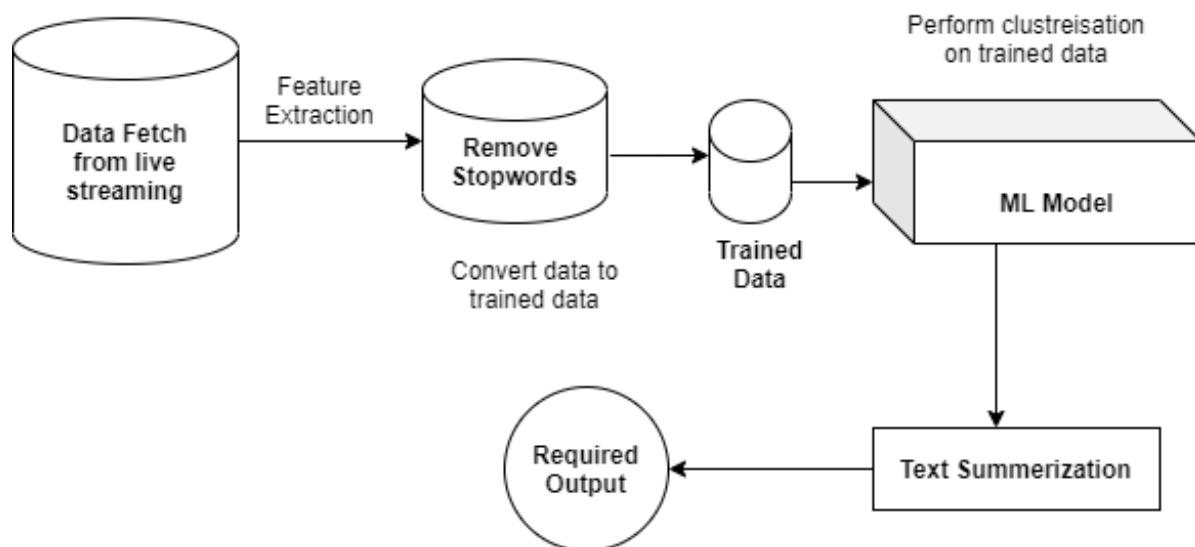
Fig. ML Model Overview

3.2.1.4 Accuracy

The data passed to machine learning model are chats after removing stop words. It gives clusters of similar data which are most closely related to each other. The accuracy achieved by the machine learning model is 82.72%.

3.2.1.5 Saving ML Model:

As once done with one-time training of the ML model and getting proper accuracy from it then it requires saving the ML model to use it.

3.2.2. Pytchat:

Pytchat library is used for retrieving data from live you tube stream.

3.2.3. Clusterization:

For clustering we have K-mean clustering algorithm. The k-means clustering algorithm is an unsupervised clustering algorithm which determines the optimal number of clusters using the elbow method. It works iteratively by selecting a random coordinate of the cluster center and assign the data points to a cluster.

3.2.4. Text Summarization:

After performing clustarization on live chat text summarization is used to summarize the organized data.

Dependencies and Requirements

4.1 Libraries used:

4.1.1 flask 1.1.2:

Flask is a lightweight WGSI (Web Server Gateway Interface) web application framework. It began as a simple wrapper around werkzeug and jinja and has become one of the most popular Python web application frameworks
License: BSD-3-Clause

4.1.2 numpy 1.18.1:

A fundamental package for array computing in python
License: OSI Approved (BSD)

4.1.3 scikit-learn 0.22.1:

Scikit-learn is a free software machine learning library for python programming. License: BSD 3-Clause

4.1.4 NLTK (Natural Language ToolKit):

NLTK is a leading platform for building Python programs to work with human language data.

4.1.5 requests 2.22.0:

Python HTTP for Humans.
License: Apache 2.0

4.1.6 TensorFlow 1.13.0:

TensorFlow is an open-source machine learning framework for everyone.
License: Apache 2.0

4.2 Tools:
- **IDE:**
  - Google Colaboratory.
  - Spider for the ML model
  - Visual Studio Code and Atom for plugin
- **Analysis:**
  - K-means algorithm for clustering.
- **Testing:**
  - Insomnia for testing Flask API.

4.3 System Environment:

- **Processor:**
  - 2.5 gigahertz (GHz) or faster processor.
- **RAM:**
  - 8 GB or more
- **Hard drive space:**
  - 48 GB for 64-bit OS or Higher
- **Operating Systems:**
  - Linux 18.04 or Higher
  - Windows 10
- **GPU:**
  - NVIDIA GTX 1050(4 GB) Compute Capability 3.5 or higher.
- **Language:**
  - Python 3.6.
- **Tool:**
  - Anaconda 3-5.2.0-Linux.
  - Anaconda3-5.2.0-Windows-x86_64.
  - Firebase connection.
- **Internet Connection:**
  - Internet connectivity is necessary to download some Libraries. Internet connection required during the training of the ML model.

Instructions for Deployment

Step 1: Download and install Visual Studio Code.

Step 2: Clone project repo to a file.

Name	Date modified	Type	Size
Images	01-Jul-21 1:43 PM	File folder	
Lively_Chat	01-Jul-21 1:43 PM	File folder	
ML Model	01-Jul-21 1:43 PM	File folder	
Project Report	29-Jun-21 6:06 AM	File folder	
README	01-Jul-21 1:43 PM	Markdown Source...	3 KB
requirement	01-Jul-21 1:43 PM	Text Document	1 KB

Name	Date modified	Type	Size
css	01-Jul-21 1:43 PM	File folder	
images	01-Jul-21 1:43 PM	File folder	
homepage	01-Jul-21 1:43 PM	Microsoft Edge H...	5 KB
index	01-Jul-21 1:43 PM	Microsoft Edge H...	6 KB
signin	01-Jul-21 1:43 PM	Microsoft Edge H...	5 KB
signup	01-Jul-21 1:43 PM	Microsoft Edge H...	6 KB

Step 3: Install packages which are in requirement.txt.

```
!pip install gensim
!pip install pytchat
!pip install scipy
!pip install sklearn
!pip install numpy
!pip install nltk
import nltk
nltk.download('punkt')
nltk.download('stopwords')
!pip install pyrebase4
```

Step 4: Open the HTML file with Google Chrome
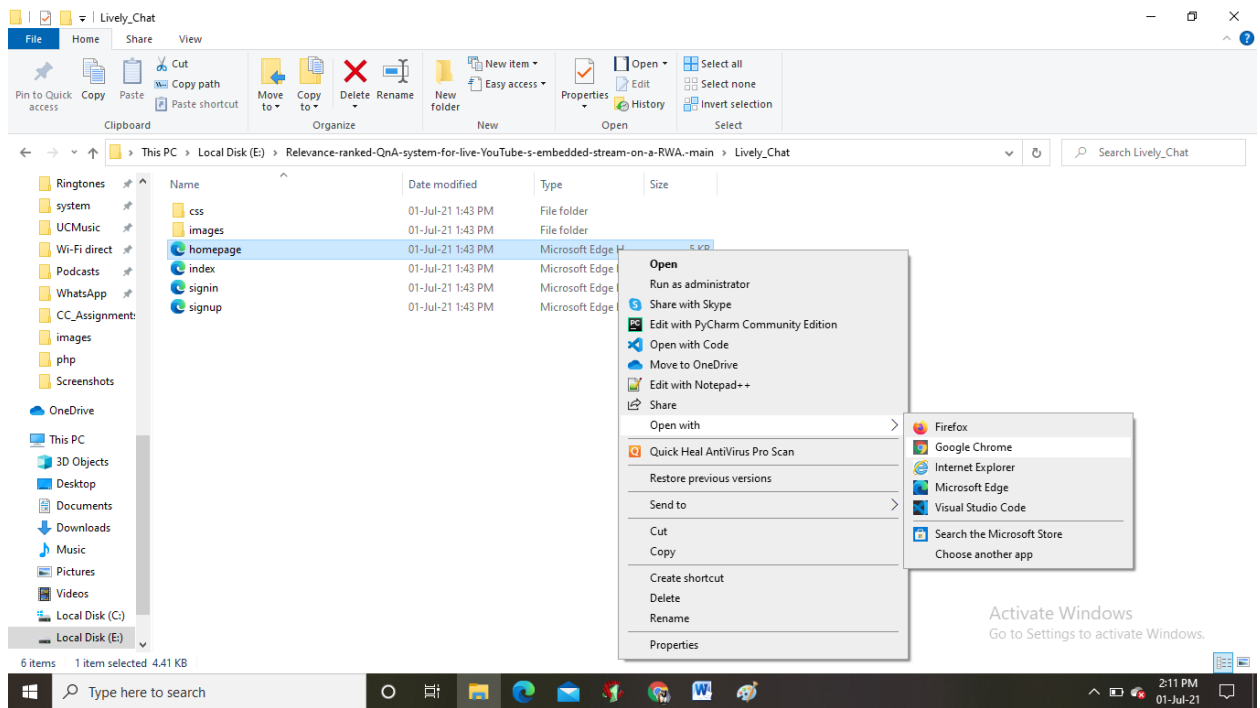
Summary

We have developed our own responsive website so that we will provide a chat window which will display all the relevance ranked questions which will make the work of the host easier for answering the questions in minimum time span and also with the chat window we are also displaying the live streaming video.

After finding many options we get pytchat library for retrieving the live chat of a live YouTube stream. The live chat is retrieved using python using pytchat library and this data does not give relevance rank so we need to put those questions with ranking.

With the help of NLTK (Natural Language Toolkit) we have removed stop words from it so that the most common words that appear in our language are removed. After that we have used Sklearn so that we can perform clustering and text summarization on that data after removal of stop words which we retrieved using pytchat and got the relevance ranked questions which are displayed in the web page which will be easier for the host to answer.

Thus, our website provides the relevance ranked questions based on the topic related to live streaming video.

Future Scope

As future work on identifying new features and related data to improve the performance of the ML model and increase the complexity of the ML model.

To extract chat with emoji and apply stemming and summarization on it.

Automatic QnA system on generated data.

To develop a mobile app version for this website.

References

1. https://developer.mozilla.org/en-S/docs/Learn/Getting_started_with_the_web/HTML_basics; HTML Basics.
2. https://www.w3schools.com/css/default.asp; CSS.
3. https://getbootstrap.com/docs/4.5/getting-started/introduction/; Introduction to Bootstrap.
4. https://www.w3schools.com/jquery/default.asp; Introduction to jQuery.
5. https://www.quora.com/How-can-I-embed-a-live-stream-with-a-live-chat-on-my-website-without-leaving-my-website; Embedding live stream on website.
6. https://stackoverflow.com/questions/21607808/convert-a-youtube-video-url-to-embed-code/21607897; Convert YouTube URL to embed code.
7. https://stackoverflow.com/questions/38541098/how-to-retrieve-data-from-firebase-database; How to retrieve data from firebase database using JavaScript.
8. https://colab.research.google.com/?utm_source=scs-index; Google Colaboratory for running ML Model.
9. https://www.python.org/; Python Documentation
10. https://flask.palletsprojects.com/en/1.1.x/tutorial/; Python Flask
11. https://github.com/taizan-hokuto/pytchat/wiki/LiveChat; Pytchat library.
12. https://www.geeksforgeeks.org/removing-stop-words-nltk-python/; Removing Stop words.
13. https://blog.eduonix.com/artificial-intelligence/clustering-similar-sentences-together-using-machine-learning/; Clusterisation
14. https://towardsdatascience.com/making-sense-of-text-clustering-ca649c190b20; Basics of clustering
15. https://colorwhistle.com/live-stream-video-on-website/; Displaying live steam on web page.
16. https://www.python.org/ftp/python/3.6.8/python-3.6.8-amd64.exe; Python3.6 Download link
17. www.gstatic.com/firebasejs/8.6.1/firebase.js; Firebase database.
18. https://becominghuman.ai/text-summarization-in-5-steps-using-nltk-65b21e352b65; Text summarisation
19. https://scikit-learn.org/stable/modules/tree.html; Scikit-Learn
20. https://www.w3schools.com/php/php_intro.asp; Introduction to PHP

Plagiarism check

Abstract:

Scan Properties

Number of Words : 222
Results Found : 0

To or From

Binary Translator

To or From

PDF Converter

0% Plagiarism   100% Unique

Start New Search

To check plagiarism in photos click here

Reverse Image Search

Nowadays, all over the planet embedding live streaming youtube video in our own responsive web site. Embedding videos is simply like making backlinks to your website. Like in SEO (Search Engine Optimization), embedding your videos in a very website behaves specifically sort of a back link and so facultative your videos to urge placed in program results and find a lot of views. The a lot of no of views our videos receive, our video quality and complete image increase too. Not solely video quality, your product and website quality too increase so increasing your sales and profits.

Video embedding is the method of adding a video player to our web site victimisation of an internet video platform. There are several web sites that are building their own social media platforms, so it is as easy as

Introduction:

Scan Properties

Number of Words : 425
Results Found : 0

To or From

Binary Translator

To or From

PDF Converter

0% Plagiarism   100% Unique

Start New Search

To check plagiarism in photos click here

Reverse Image Search

In this real time all over the world live streaming video is embedded into their own responsive website.Since recent years have brought an increase in the popularity of video-sharing across hundreds of different platforms. This means that a number of people are sharing live and on demand videos regularly. Knowing how to broadcast live and embed live streaming video on your website is also becoming increasingly important for all kinds of broadcasters.People can not only watch live streaming video and live chat, but they can also leave blog post comments which mainly remain beyond the broadcast. People can also post supplemental information and links related to our livestream for viewers. If we stream using multiple platforms our blog is the place where fans can always find our latest broadcast.

As we know that many people make their own live streaming video.And the viewers who are watching their streaming they put or

Front End:

Scan Properties

Number of Words : 189
Results Found : 1

To or From
Binary Translator

To or From
PDF Converter

9%   91%
Plagiarism   Unique

Make it Unique   **Start New Search**

To check plagiarism in photos click here
Reverse Image Search

2.1.1 HTML :

HTML is a standard markup language for creating web pages and web applications. It can be assisted by Cascading Style Sheets (CSS) and scripting languages like JavaScript. The front end of a project is built using HTML and CSS. The browser does not display the HTML tags but is utilized to build content of the page also used in plug-in development.
2.1.2 CSS:
CSS is used for styling the content of web pages including colors, layout, fonts. It will allow adapting the presentation of different CSS styles of documents written in a markup language like HTML.

Similarity 10%
Frontend Web Development Guide on the App Store

· jQuery is a fast and concise JS library. jQuery simplifies HTML document traversing, event handling, and Ajax interactions for Rapid Web Development.

https://apps.apple.com/us/app/frontend-web-development-guide/id1564207517

Back End:

Scan Properties

Number of Words : 286
Results Found : 0

To or From
Binary Translator

To or From
PDF Converter

0%   100%
Plagiarism   Unique

Start New Search

To check plagiarism in photos click here
Reverse Image Search

2.2.1 Flask
Flask may be a fashionable light-weight python internet application framework and baseband on the WSGI toolkit. Flask provides an easy example to create internet applications. Flask may be accustomed save time building internet applications once being foreign into python. it's no info, abstraction layer, or kind validation or the other parts however flask supports extensions. Flask is employed to create the remainder API of the project with Oauth1. Oauth1 is Authentication level one used for authentication functions.
Features of Flask:

Methodology:

System Architecture Overview (High-level Architecture):

Scan Properties

Number of Words : 252
Results Found : 0

To or From
Binary Translator

To or From
PDF Converter

0%
Plagiarism

100%
Unique

Start New Search

To check plagiarism in photos click here
Reverse Image Search

4.1.1 Data Collection:
Data should be collected from YouTube's live chat for preprocessing.
The messages are collected from Youtube in systematic manner
4.1.2 Data Preprocessing:
Real-world data is often incomplete, inconsistent, and lacking in
certain behavior trends and likely contains many errors. Data
processing includes transforming raw data into an understandable
format.
4.1.3 Prepared Data:
After data collection and data pre-processing the prepared data is
saved in the required format (like CSV).
4.1.4 Splitting Data:
Partition of data into training data and test data. Splitting training data
in

Low-Level Architecture:

Scan Properties

Number of Words : 309
Results Found : 1

To or From
Binary Translator

To or From
PDF Converter

6%
Plagiarism

94%
Unique

Make it Unique   Start New Search

To check plagiarism in photos click here
Reverse Image Search

3.2 Low-Level Architecture

3.2.1 Machine Learning Model
3.2.1.1 Data Exploration:
As the data required to train machine learning module is live chat

Similarity 7%
Clustering Similar Sentences Together Using Machine Learning

https://blog.eduonix.com/artificial-intelligence/clustering-similar-sentences-together
-using-machine-learning/#:~:text=The%20k%2Dmeans%20clustering%20algorithm,dat
a%20points%20to%20a%20cluster.

Summary:

Scan Properties

Number of Words : 189
Results Found : 0

To or From | To or From
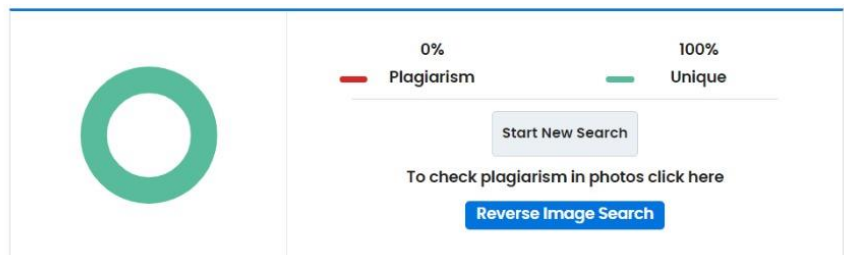
Binary Translator | PDF Converter

0%   Plagiarism
100%   Unique
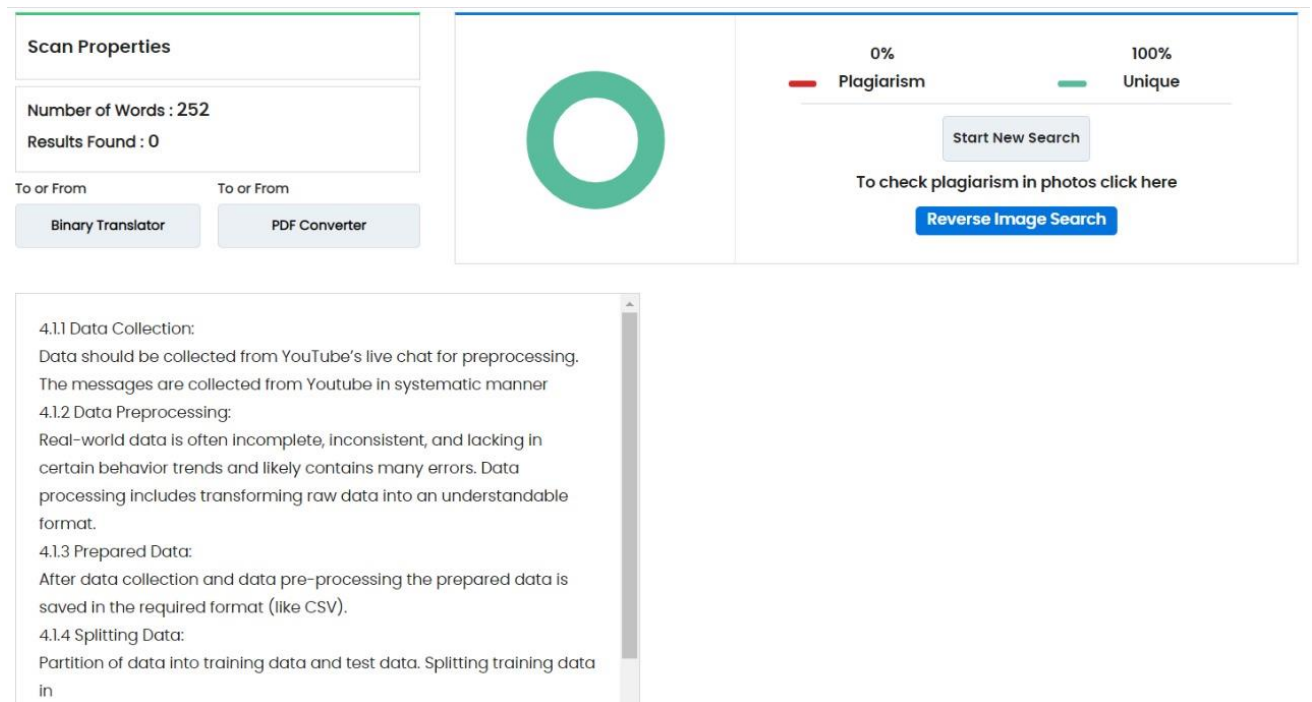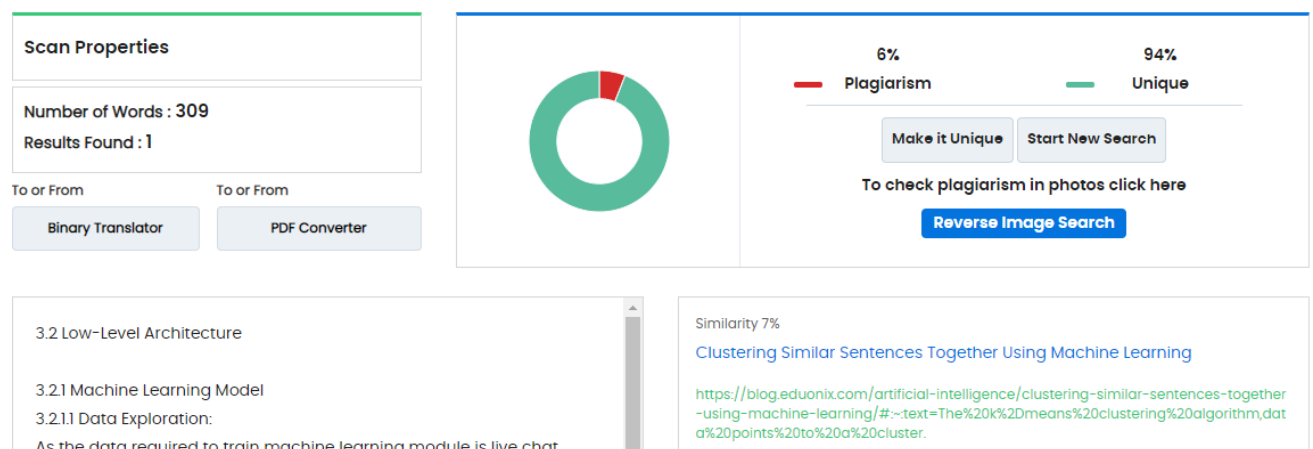
Start New Search

To check plagiarism in photos click here

Reverse Image Search

We have developed our own responsive website so that we will provide a chat window which will display all the relevance ranked questions which will make the work of the host easier for answering the questions in minimum time span and also with the chat window we are also displaying the live streaming video.

After finding many options we get pytchat library for retrieving the live chat of a live youtube stream.The live chat is retrieved using python using pytchat library and this data does not give relevance rank so we need to put those questions with ranking.

With the help of nltk(Natural Language ToolKit) we have removed

Student Details:

Name	Roll No.	Seat No.	Email ID	Contact No.
Anjali Chougule	40		anjalichougule23@gmail.com	7083233686
Janvi Pampattiwar	41		janvipampatiiwar123@gmail.com	8408954389
Pragati Phand	42		pragatiphand13@gmail.com	9067922831
DurreAfshan Shaikh	43		afshan8856@gmail.com	9518907974

Group Photo:

Anjali Chougule
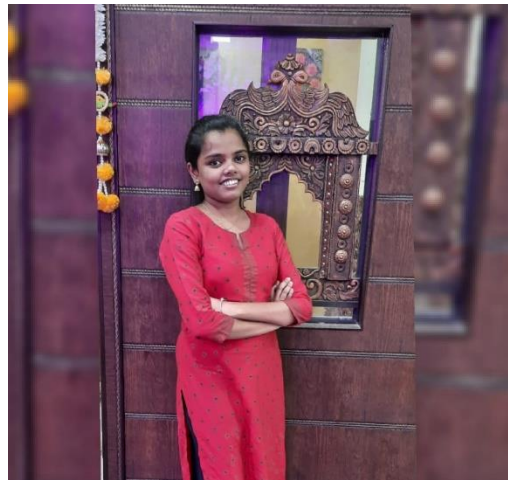
Janvi Pampattiwar

Pragati Phand

DurreAfshan Shaikh

GitHub Link: