IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

RETRIEVING CONTENT BASED DOCUMENTS BY USING IMAGE PROCESSING AND NLP THROUGH QUERY

Prof. Jitendra Musale, Shon Nikam, Akash Deshmukh, Shantanu Gavane, Swapnil Matkar, Rahul Nimbalkar

Department of Computer Engineering, Anantrao Pawar College Of Engineering and Reasearch , Parvati, Pune

Abstract: In th early of all technology's the traditional way is used to search or retrive the documents by manuallysuch as using the data attributes like name, topic, keyword, date, time, which would be easy for computer to understand it and perform retrieval process. So to avoid this traditional methodology to retrieve or search the document we used text summarisation, image processing and natural language processing this lead to provide required data files which is stored in the database. By using text summarization and NLP the textual annotations will actually match the textual information which is present kin data files. Objective of our system is to retrive a particular document from the database by passing keyword as a query to it and comparing the passed keyword with the actual content of the document it should be retrieved and aim is to extract particular document from the database which is saved by comparing its content with keywords which passed as query. The traditional way is to storing the documents but instead of storing the digital copy of the document is the effiecient to retrive the documents. Storing paper-based documents converted into digital image format iamge o effiective solution to preserve content in document. Searching the document images stored in the data set is a repository by using time as a quries in one of importance tasks of document retrival.

***Index Terms* - Text Summarization, Image Processing ,Natural Language Programming, Optical Charachter Recognition, Text Recognition, Image Processing, .**

• Introduction

A couple of yeans ago the documents were searched or retrieved on the basis of the textual annotation which were given to it manually such as name, topic, keyword, date, time which would be easy for computer to understand it and perform retrieval process over annotations.So to overcome such scenario , image processing & NLP can play an crucial role so that the required information can be retrieved from the database itself rather using the textual annotations and then with help of NLP we can actually match the textual information.Aim of the system is to retrieve a particular document from database by passing a keyword as a query to it and by comparing the keyword with the actual content of the document it should be retrieved.Objective is to extract particular document from database by comparing its content with keywords which passed as query.So to make system working we are used image processing and one of the important part from AI i.e Natural Language Processing. Our system aims to find and retrieve content based document from the database by image processing & NLP through query.The system will provide document as per queries provided by user. User have to store the image files in the database. User have to give the priamry and unique key or attribute for the required documnet. The attribute which have unique identity which leads specifies the document that is required to the user. And as the primary unique attribute provided by the user then system will give required document.

Image Processing : Image processing is method to perform actions on an image i.e checking for presence ,object detection and localization, measurement and most important is identification and verification. Image Processing is can be used for get an enhanced image or to extract some required part of information from it. Image processing enhanced the image according to the specific tast. Its type of signal processing in which an image is inputed and output may be a required information or the image. Most of the Colleges, University, Schools are dealing with the students scanned documents. In this system, document image retrieval is very interesting and efficient way to retrive infromation or the image documents. The system overcomes the problem of the plenty of time required to search the documents by usig image processing. The traditional method of searching documents required many details of that image or document and also required time to search but using image processing, process of searching document images with the help of queries will results accurate output with image file. This can be occurs with the help of image processing.

Natural Language Processing : Natural Language Processing(NLP) is one of the most important part of AI(Airtificial Intelligence). In our system NLP plays very importatnt role , the language which is used by user naturally is going to process by the system with the help of AI. In NLP, Tokenization is a way of separating a piece of text into smaller units and these units are also known as Token. tokens can be words, characters, subwords or numbers. In this Tokenization, it tokenize each word of the extracted text data by the Optical Character Recognisation(OCR) to separate each word by column. In NLP it also removes the Stopwords of the language like as, the, are, they, we,etc which helps to filter out the textual data and main content of the data file is filtered. As the data gets filtered the query pass by the user is searched in the textual data files and provides required document. This helps to reduce the time complexity of system. The Text Normalization converts all the data into lower case. Due to text normalization all words will be converted into lower case. This will helps to find accurate keyword or query passed by the user. Text Normalization makes system efficient and also provide required output. Using NLP , we can remove the punctuation marks. In this we remove all the punctuation marks which are included in text data (. " ; ? ! :). This removal of punctuation marks makes the extracted data more refined tit means it will required less time to search query passed by the user. As we use all this factors of NLP we will get accurate output with less time which leads to better system.

• Literature Work

Following are the research papers we studied to understand the retrieving document using image processing:

1. The Simple Image Processing Scheme for Document Retrieval Using Date of Issue as Query.
Authors : Panuwat Ketwong, Piyabhorn Hongsa-arparsat, Ekkharin Srilaphat, Wilailuck Kaprasit
Description : System returning the desired documents as per the queries provided by users
Journal : IEEE international conference[2017]

2. Information Processing and Retrieval from CSV File by Natural Language
Authors : Charmpol Tapsai
Description : Non-technician users easily retrieve information without the need to learn any additional computer languages or programs
Journal : IEEE international conference[2019]

3. Image processing technology for text recognition
Authors : Yen-Min Su,Hsing-Wei Peng,Ko-Wei Huang,Chu-Sing Yang
Description : This study demonstrates how image-processing technologies can be used in combination with optical character recognition to improve recognition accuracy, efficiency of extracting text from images
Journal : IEEE international conference [2019]

• Existing System

It presents a simple scheme of image processing to retrieve the documents, which contain the desired date of issue printed in Thai alphabets , from a repository. The procedure of this retrieval scheme consists of 4 stages: image acquisition, pre-processing, zone identification, and pattern recognition.In this system their actual focus was to retrieve particular document only on the basis of date of issue rather then any other factor. So the keywords which were passed as a query are dates which are in Thai language and later on they are matched with English language by using template matching technique.

• Gaps In Existing System

Existing scheme was the retrieval system dealing with query based only on date of issue machine-printed documents.As the documents are been retrieved on the basis of date of issue so there can be multiple documents which are been issued on same date.If there are multiple documents retrieved by using date of issue then we may face one more problem that again we have to find our desired document among documents retrieved.

• Proposed Work

The main aim of the system is to find and retrieve content based image document from the storage by using image processing and natural language processing (NLP) by passing query as an inputto the system. We will be using python as our backend language, html, CSS are for designing web pages and Google Firebase storage is used for storing is used for storing the images. As we are developing web app we are using Python's FLASK framework for server side programming.

Initially we have stored images on Google firebase storage and links of these images we have given as input to the main programming logic. As now we are getting the images as an input to the programming logic first we have to configure Google firebase storage with python. And to do so import a package called PyreBase using which we are configuring firebase to our progeam.

The links of the images are stored individually as element in a python data stucture called list[url_list].

As now we are going to traverse each element i.e link of images in this list for processing those images by which we will be able to extraction textual data from the images. For processing images and extracting textual data from it we using technology called OCR.

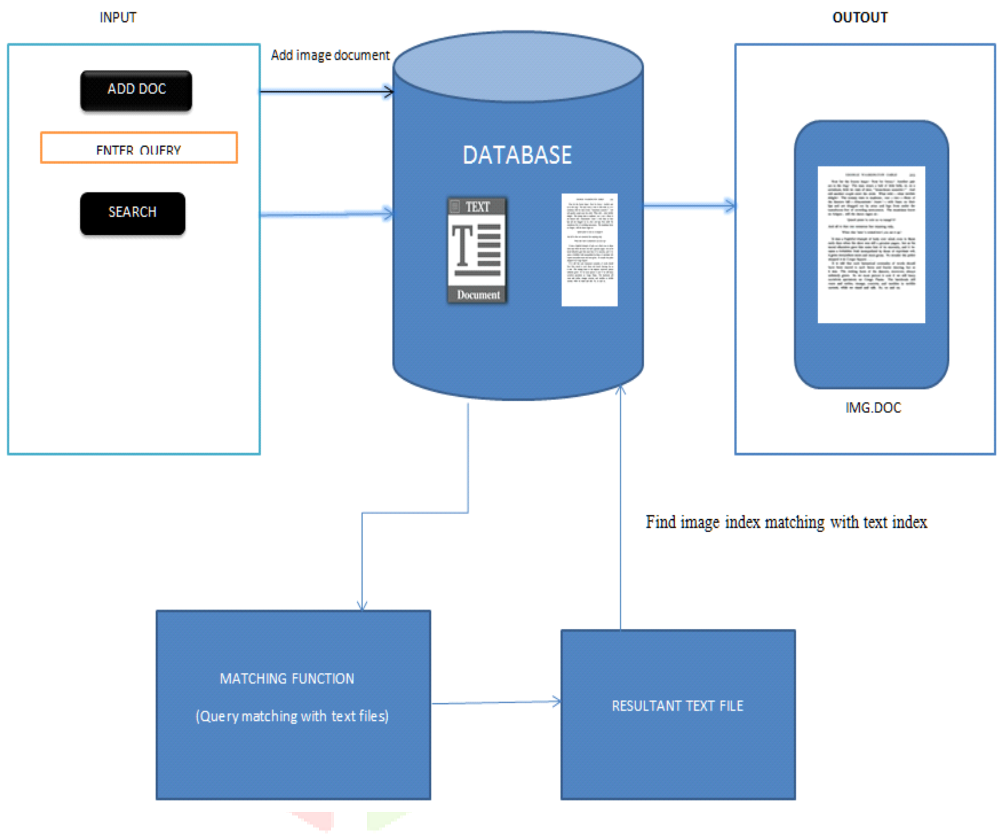
Optical Character Recognisation or OCR is technology that enables you to convert different types of documents such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data for this purpose we are using

Tessaract which is a OCR engine. But installing only this is not enough , once we install tesseract we need to connect with the Python. Pytessaract library is used to bind tesseract python, so we can call tesseract from python code.

Pytesseract i.e Python-tesseract is an optical character recognition(OCR) tool for python that will recognize and "read" the text embedded in images. Once all the images are processed and the textual data is extracted from them then that textual data is stored separately as an element in a list data structure. And here install package of NLTK and apply NLP's text summarization techniques of the textual data is stored in list[tex_list]. NLTK(Natural Language Toolkit) has strong corpora and lexical resourses such walnet. NLTK is open-source tool created to make NLP processes in Python.

Now when we pass keyword or multiple keywords from client or user side to server side , the matching of that query whether it is present in any textual data which is processed by NLP or whatever it matches with it or not .This process takes place at servers side and as a response sever sends the links of images of whos textual data contains the keywords passed as a query and if the keywords are not matched then server passes the response as an web page which contains the message that is match is not found.

- **Architecture Diagram**



- **Modules :**

Home Page - In this module there is Add button by which user can add the new documents or image data files in database. In Search bar user have to enter the primary unique key and Search button is to search the keyword in the files which are stored in the database

Database - It is used to store image documents and text documents added the user . We can search documents in the database with the help of keyword which is pass as query by user.

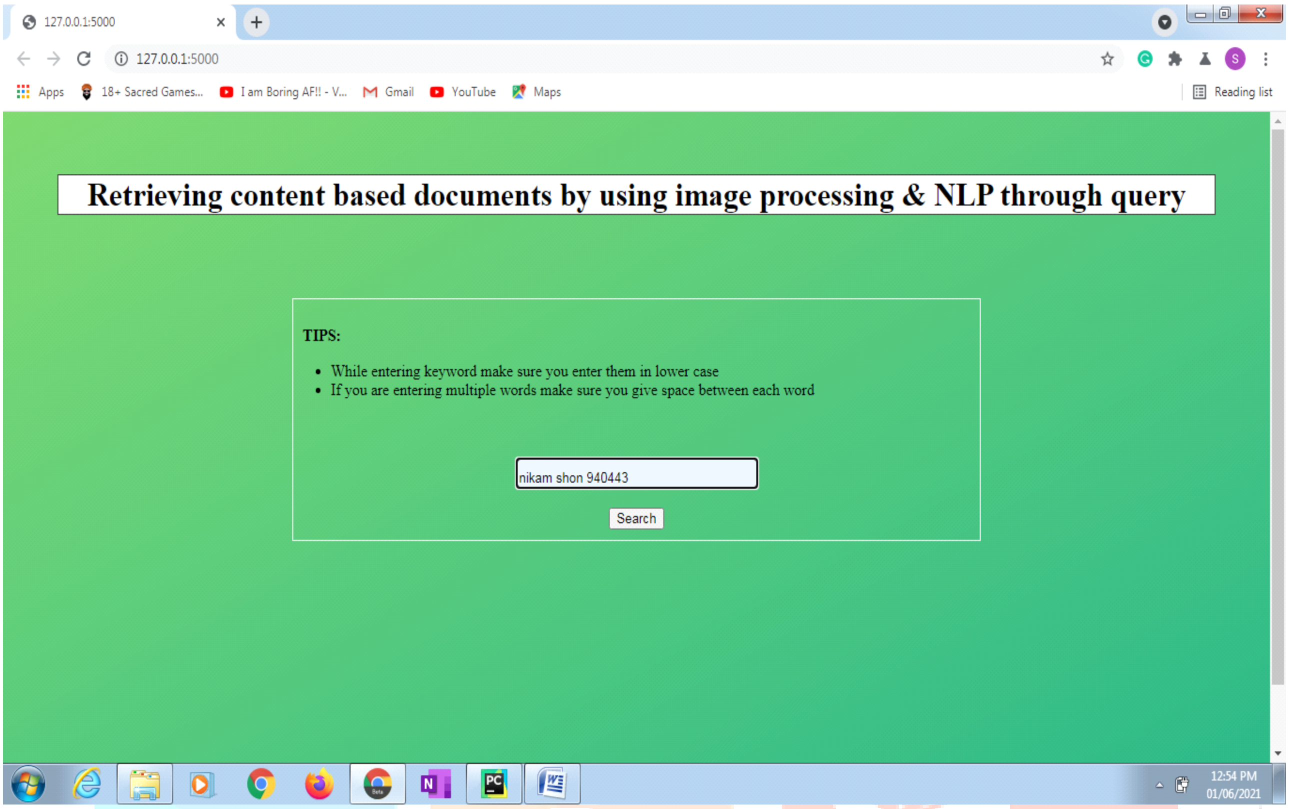
Matching Function - In this module keyword is pass as query has been searched in text document generated after image processing technique OCR . The query pass by the user is searched in the Database if yes future processes will takes place.

Resultant Text File - In previous module that is in matching fucntion module ,if the keyword is found in any text document or in image file, then that resultant text document is been passed to database to find its matching indexed image document.

Output - If the text document received from matching fuction module finds image document which is required by user document with same index then that image document is displayed as an output.

- **System's Design :**

1. Home Page :



Retrieving content based documents by using image processing & NLP through query

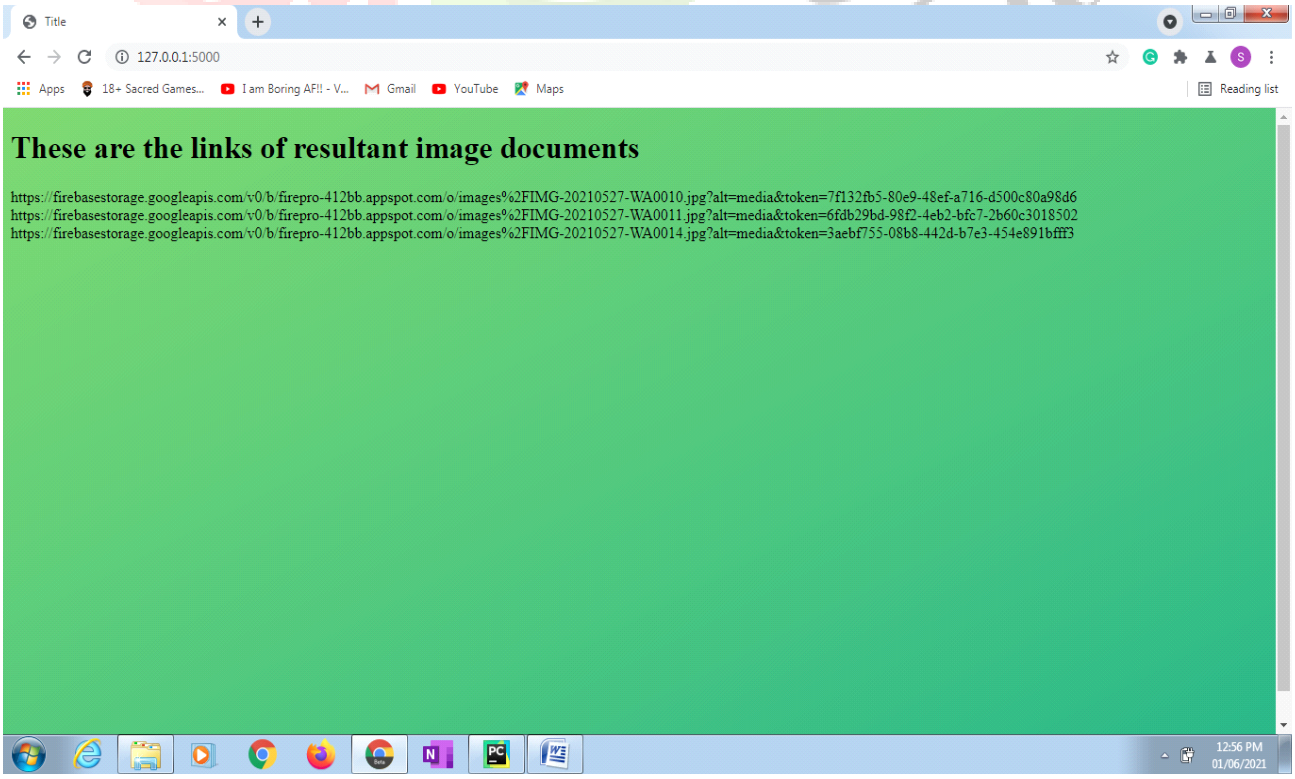
TIPS:

- While entering keyword make sure you enter them in lower case
- If you are entering multiple words make sure you give space between each word

nikam shon 940443

Above mentioned daigram is of Home page of system , In this we simply add keyword in search bar. After entering the keywords in search bar then user will click on Search button and exeution will be start.

2.Links of entered keywords :

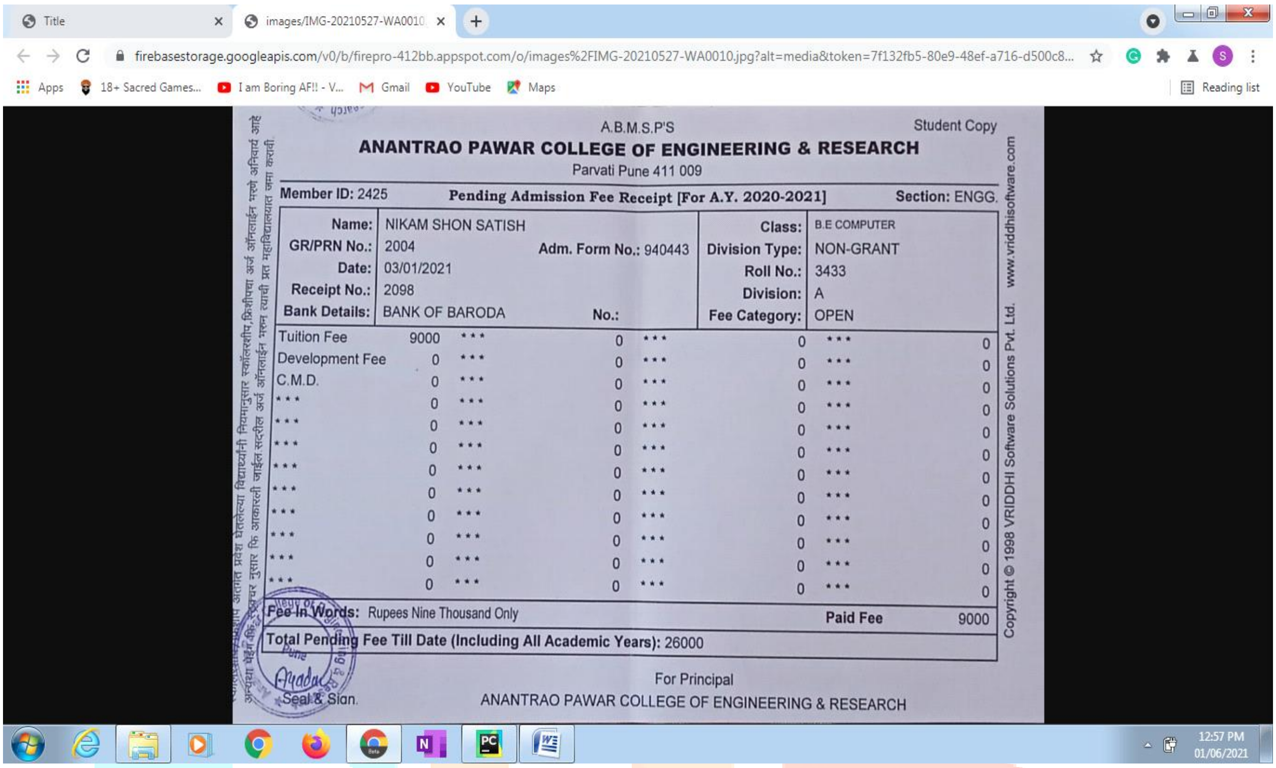


These are the links of resultant image documents

```
https://firebasestorage.googleapis.com/v0/b/firepro-412bb.appspot.com/o/images%2FIMG-20210527-WA0010.jpg?alt=media&token=7f132fb5-80e9-48ef-a716-d500c80a98d6
https://firebasestorage.googleapis.com/v0/b/firepro-412bb.appspot.com/o/images%2FIMG-20210527-WA0011.jpg?alt=media&token=6fdb29bd-98f2-4eb2-bfc7-2b60c3018502
https://firebasestorage.googleapis.com/v0/b/firepro-412bb.appspot.com/o/images%2FIMG-20210527-WA0014.jpg?alt=media&token=3aebf755-08b8-442d-b7e3-454e891bfff3
```

In this module, as the user entered keyword as query , then system gives response with the links in which query passed by the user is occured and user get options whether which document is required.

3. Result :


Title
images/IMG-20210527-WA0010
firebasestorage.googleapis.com/v0/b/firepro-412bb.appspot.com/o/images%2FIMG-20210527-WA0010.jpg?alt=media&token=7f132fb5-80e9-48ef-a716-d500c8...
Apps
18+ Sacred Games...
I am Boring AF!! - V...
Gmail
YouTube
Maps
Reading list
12:57 PM
01/06/2021

A.B.M.S.P'S

Student Copy

ANANTRAO PAWAR COLLEGE OF ENGINEERING & RESEARCH

Parvati Pune 411 009

Member ID: 2425 **Pending Admission Fee Receipt [For A.Y. 2020-2021]** **Section:** ENGG.

Name:	NIKAM SHON SATISH	**Class:**	B.E COMPUTER
GR/PRN No.:	2004 **Adm. Form No.:** 940443	**Division Type:**	NON-GRANT
Date:	03/01/2021	**Roll No.:**	3433
Receipt No.:	2098	**Division:**	A
Bank Details:	BANK OF BARODA **No.:**	**Fee Category:**	OPEN

Tuition Fee	9000	***	0	***	0	***	0
Development Fee	0	***	0	***	0	***	0
C.M.D.	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0
***	0	***	0	***	0	***	0

Fee In Words: Rupees Nine Thousand Only **Paid Fee** 9000

Total Pending Fee Till Date (Including All Academic Years): 26000

For Principal

ANANTRAO PAWAR COLLEGE OF ENGINEERING & RESEARCH

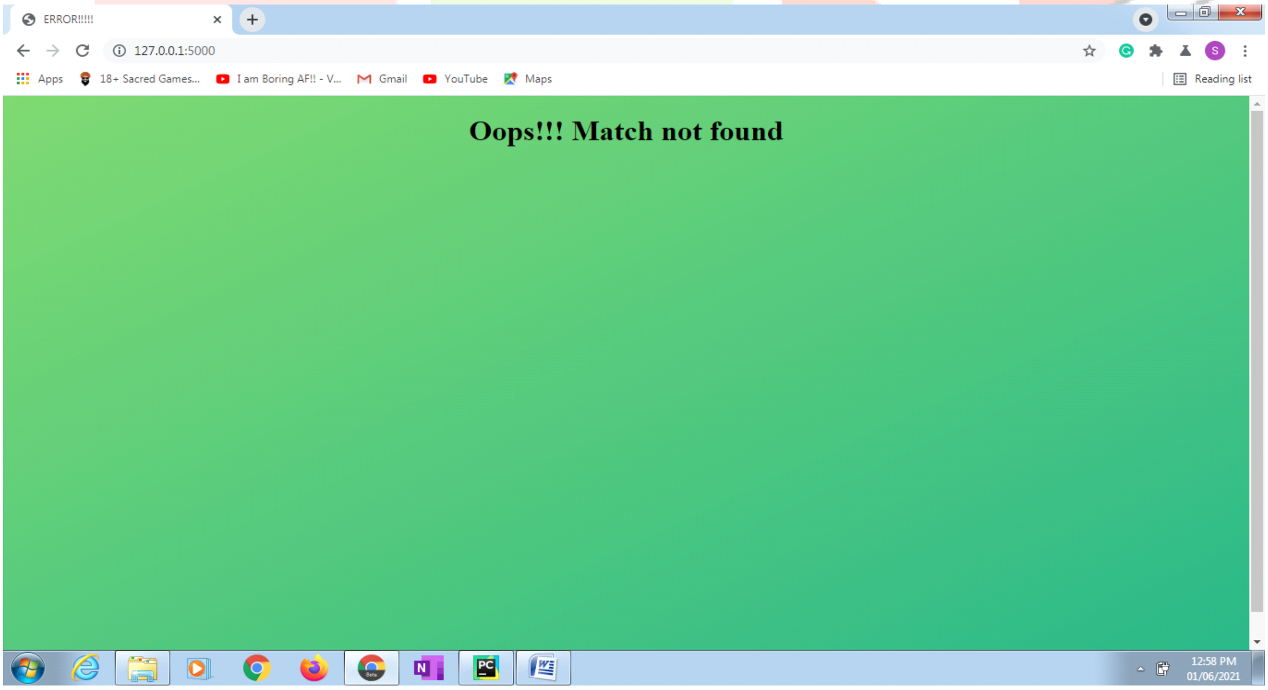
Seal & Sign.

अंतर्गत प्रवेश घेतलेल्या विद्यार्थ्यांनी नियमानुसार स्कॉलरशीप, फ्रिशीपचा अर्ज ऑनलाईन भरणे अनिवार्य आहे अन्यथा [illegible] नुसार फि आकारली जाईल सदरील अर्ज ऑनलाईन भरून त्याची प्रत महाविद्यालयात जमा करावी.

Copyright © 1998 VRIDDHI Software Solutions Pvt. Ltd. www.vriddhisoftware.com

As the user select the link and system will response with the photo.

4. Query or Keyword is not matches with Data files:



Oops!!! Match not found

As user entered wrong query in the search bar then system will be show this message "Oop!!! Match not found".

- **Technologies :**

OCR - When we will store image document in the database,text file will also be created for the same by using optical character recognition(OCR)**.** The system takes image as input and OCR converts actual image content to textual format and it is stored in database.

Open CV- OpenCV is a cross-platform library using which we can develop real-time computer vision applications. It mainly focuses on image processing, video capture and analysis including features like face detection and object detection.OpenCV-Python is a library of Python bindings designed to solve computer vision problems. OpenCV-Python makes use of Numpy, which is a highly optimized library for numerical operations.

Numpy- NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python.

Tesseract, Pytesseract – Tesseract is OCR engine. But installing only this is not enough, once we install tesseract we need to connect with python and opencv. Pytesseract library is used to bind tesseract with python, so we can call tesseract from python code.

Python-tesseract is an optical character recognition (OCR) tool for python. That will recognize and “read” the text embedded in images. Python-tesseract is wrapper for Google tesseract OCR engine.

- **Applications :**

1. School , Colleges , University , Classes
2. Govenment oiices
3. Banking sctors
4. Manufactoring company

- **Conclusion :**

In this topic we have seen how the document is been retrieved using image processing and NLP by passing keywords within content of document as query which are machine typed.

- **References :**

1. Panuwat Ketwong, Piyabhorn Hongsa-arparsat, Ekkharin Srilaphat, Wilailuck Kaprasit,”The Simple Image Processing Scheme for Document Retrieval Using Date of Issue as Query”,IEEE International conference,2017.
2. Charmpol Tapsai, “Information Processing and Retrieval from CSV File by Natural Language”,IEEE International conference,2018.
3. Yen-Min Su,Hsing-Wei Peng,Ko-Wei Huang,Chu-Sing Yang,”Image processing technology for text recognition”,IEEE International conference,2019.