TEXT AND IMAGE PLAGIARISM DETECTION

**1.INTRODUCTION:**

**\*INTRODUCTION ABOUT PROBLEM YOU ARE TRYING TO SOLVE:**

Plagiarism in itself cannot be considered as a crime but as copyright violation. In the academics and other industries that are sensitive to copyright infringement, plagiarism is grave misconduct in integrity. The law cannot and usually will not punish plagiarism, but it is up to the institution on how to handle it once it happens Plagiarism detection is usually split into two which is text-based plagiarism detection and image-based plagiarism detection. For text-based plagiarism detection there are currently five techniques that is used most often in different fields. These techniques are Fingerprinting, String Matching, Bag of Words, Citation Analysis and Stylometry. String Matching is mostly used in computer science where it compares the documents words for words. Bag of words represents the documents in one or two vectors for comparison. Citation analysis is mainly used in scientific texts because it only compares the citation and reference of the documents. Stylometry check the author’s unique writing style for detection author’s ownership . For image-based plagiarism detection, there is no commonly used techniques like the text-based plagiarism detection, but they usually share the same processes and steps. When we say plagiarism checking or detection we usually mean checking only the text in the file or document for plagiarism. Most of the times when you check your documents or files for plagiarism through a plagiarism checker software they will check for images and then discard themPlagiarism basically means the wrongful stealing of an author’s work, thoughts, ideas, etc. and claiming it as your own original work. Plagiarism is considered as deceit and a breach of ethics. In academics, students that are caught with plagiarism are exposed to various level of penalties and punishment and may even lead to expulsion and we have to try to solve .

\* **INTRODUCTION ABOUT EXISTING SYSTEM:**

Plagiarism is the practice of copying someone else’s work or ideas, and passing them off as one’s own original work. Not only images but, architecture, flow diagram, UML diagrams, even snapshots of test results can be plagiarized. If the author has not mentioned the credit for the original author from where he/she copied the image then it is said to be plagiarized.

\* **INTRODUCTION ABOUT PROPOSED SYSTEM:**

The proposed work mainly focus on finding the similarity between two images. Sample image is given as the reference and it is compared with the other image which is taken from any journal and comparison is done through histogram. Histogram is the best way to visualize the largest intensities of an image. It is used to find the problems which originate during image acquisition such as exposure, contrast etc. even a minute difference with the pixel is noticed by histogram

BRIEF INTRODUCTION ABOUT PLATFROM AND TECHNOLOGY USED AND WHY:

**What is Python :**

Below are some facts about Python.

Python is currently the most widely used multi-purpose, high-level programming language.

Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.

Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.

Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard library which can be used for the following –

* + [Machine Learning](https://www.geeksforgeeks.org/machine-learning/)
  + GUI Applications (like Kivy, Tkinter, pyqt etc. )
  + Web frameworks like Django (used by youtube, Instagram, Dropbox)
  + Image processing (like Opencv, Pillow)
  + Web scraping (like Scrapy, beautifulsoup, Selenium)
  + Test frameworks
  + Multimedia

**Advantages of Python :-**

Let’s see how Python dominates over other languages.

## 1. Extensive Libraries

Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don’t have to write the complete code for that manually.

## 2. Extensible

As we have seen earlier, Python can be**extended to other languages**. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

## 3. Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add **scripting capabilities**to our code in the other language.

## 4. Improved Productivity

The language’s simplicity and extensive libraries render programmers**more productive** than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

## 5. IOT Opportunities

Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet Of Things. This is a way to connect the language with the real world.

When working with Java, you may have to create a class to print **‘Hello World’**. But in Python, just a print statement will do. It is also quite **easy to learn, understand,** and**code.** This is why when people pick up Python, they have a hard time adjusting to other more verbose languages like Java.

## 7. Readable

Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and **indentation is mandatory.** This further aids the readability of the code.

## 8. Object-Oriented

This language supports both the **procedural and object-oriented**programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the **encapsulation of data** and functions into one.

## 9. Free and Open-Source

Like we said earlier, Python is **freely available.** But not only can you[**download Python**](https://data-flair.training/blogs/install-python-windows/) for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

#### 10. Portable

When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn’t the same with Python. Here, you need to**code only once**, and you can run it anywhere. This is called **Write Once Run Anywhere (WORA)**. However, you need to be careful enough not to include any system-dependent features.

## 11. Interpreted

Lastly, we will say that it is an interpreted language. Since statements are executed one by one, **debugging is easier** than in compiled languages.

Any doubts till now in the advantages of Python? Mention in the comment section.

# ADVANTAGES OF PYTHON OVER OTHER LANGUAGES :

## 1. Less Coding

Almost all of the tasks done in Python requires less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don’t have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python to beginners.

## 2. Affordable

Python is free therefore individuals, small companies or big organizations can leverage the free available resources to build applications. Python is popular and widely used so it gives you better community support.

**The 2019 Github annual survey showed us that Python has overtaken Java in the most popular programming language category.**

## 3. Python is for Everyone

Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and [**machine learning**](https://data-flair.training/blogs/machine-learning-tutorials-home/), automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

# DISADVANTAGES OF PYTHON :

So far, we’ve seen why Python is a great choice for your project. But if you choose it, you should be aware of its consequences as well. Let’s now see the downsides of choosing Python over another language.

#### 1. Speed Limitations

We have seen that Python code is executed line by line. But since [Python](https://www.python.org/) is interpreted, it often results in **slow execution**. This, however, isn’t a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

#### 2. Weak in Mobile Computing and Browsers

While it serves as an excellent server-side language, Python is much rarely seen on the **client-side**. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is called **Carbonnelle**.

The reason it is not so famous despite the existence of Brython is that it isn’t that secure.

#### 3. Design Restrictions

As you know, Python is **dynamically-typed**. This means that you don’t need to declare the type of variable while writing the code. It uses **duck-typing**. But wait, what’s that? Well, it just means that if it looks like a duck, it must be a duck. While this is easy on the programmers during coding, it can**raise run-time errors**.

#### 4. Underdeveloped Database Access Layers

Compared to more widely used technologies like **JDBC (Java database Connectivity)** and **ODBC (Open database Connectivity)**, Python’s database access layers are a bit underdeveloped. Consequently, it is less often applied in huge enterprises.

#### 5. Simple

No, we’re not kidding. Python’s simplicity can indeed be a problem. Take my example. I don’t do Java, I’m more of a Python person. To me, its syntax is so simple that the verbosity of Java code seems unnecessary.

This was all about the Advantages and Disadvantages of Python Programming Language.

# HISTORY OF PYTHON : -

What do the alphabet and the programming language Python have in common? Right, both start with ABC. If we are talking about ABC in the Python context, it's clear that the programming language ABC is meant. ABC is a general-purpose programming language and programming environment, which had been developed in the Netherlands, Amsterdam, at the CWI (Centrum Wiskunde &Informatica). The greatest achievement of ABC was to influence the design of Python.Python was conceptualized in the late 1980s. Guido van Rossum worked that time in a project at the CWI, called Amoeba, a distributed operating system. In an interview with Bill Venners1, Guido van Rossum said: "In the early 1980s, I worked as an implementer on a team building a language called ABC at Centrum voor Wiskunde en Informatica (CWI). I don't know how well people know ABC's influence on Python. I try to mention ABC's influence because I'm indebted to everything I learned during that project and to the people who worked on it."Later on in the same Interview, Guido van Rossum continued: "I remembered all my experience and some of my frustration with ABC. I decided to try to design a simple scripting language that possessed some of ABC's better properties, but without its problems. So I started typing. I created a simple virtual machine, a simple parser, and a simple runtime. I made my own version of the various ABC parts that I liked. I created a basic syntax, used indentation for statement grouping instead of curly braces or begin-end blocks, and developed a small number of powerful data types: a hash table (or dictionary, as we call it), a list, strings, and numbers."

# WHAT IS MACHINE LEARNING : -

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data.

Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models tunable parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain.Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

## Categories Of Machine Leaning :-

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning.

Supervised learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classification tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section.

Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

## Need for Machine Learning

Human beings, at this moment, are the most intelligent and advanced species on earth because they can think, evaluate and solve complex problems. On the other side, AI is still in its initial stage and haven’t surpassed human intelligence in many aspects. Then the question is that what is the need to make machine learn? The most suitable reason for doing this is, “to make decisions, based on data, with efficiency and scale”.

Lately, organizations are investing heavily in newer technologies like Artificial Intelligence, Machine Learning and Deep Learning to get the key information from data to perform several real-world tasks and solve problems. We can call it data-driven decisions taken by machines, particularly to automate the process. These data-driven decisions can be used, instead of using programing logic, in the problems that cannot be programmed inherently. The fact is that we can’t do without human intelligence, but other aspect is that we all need to solve real-world problems with efficiency at a huge scale. That is why the need for machine learning arises.

## Challenges in Machines Learning :-

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are −

**Quality of data** − Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.

**Time-Consuming task** − Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.

**Lack of specialist persons** − As ML technology is still in its infancy stage, availability of expert resources is a tough job.

**No clear objective for formulating business problems** − Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.

**Issue of overfitting & underfitting** − If the model is overfitting or underfitting, it cannot be represented well for the problem.

**Curse of dimensionality** − Another challenge ML model faces is too many features of data points. This can be a real hindrance.

**Difficulty in deployment** − Complexity of the ML model makes it quite difficult to be deployed in real life.

## Applications of Machines Learning :-

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications of ML −

* Emotion analysis
* Sentiment analysis
* Error detection and prevention
* Weather forecasting and prediction
* Stock market analysis and forecasting
* Speech synthesis
* Speech recognition
* Customer segmentation
* Object recognition
* Fraud detection
* Fraud prevention
* Recommendation of products to customer in online shopping

# How to Start Learning Machine Learning?

Arthur Samuel coined the term **“Machine Learning”** in 1959 and defined it as a **“Field of study that gives computers the capability to learn without being explicitly programmed”.**

And that was the beginning of Machine Learning! In modern times, Machine Learning is one of the most popular (if not the most!) career choices. According to [Indeed](http://blog.indeed.com/2019/03/14/best-jobs-2019/), Machine Learning Engineer Is The Best Job of 2019 with a 344% growth and an average base salary of **$146,085** per year.

But there is still a lot of doubt about what exactly is Machine Learning and how to start learning it? So this article deals with the Basics of Machine Learning and also the path you can follow to eventually become a full-fledged Machine Learning Engineer. Now let’s get started!!!

### How to start learning ML?

This is a rough roadmap you can follow on your way to becoming an insanely talented Machine Learning Engineer. Of course, you can always modify the steps according to your needs to reach your desired end-goal!

### Step 1 – Understand the Prerequisites

In case you are a genius, you could start ML directly but normally, there are some prerequisites that you need to know which include Linear Algebra, Multivariate Calculus, Statistics, and Python. And if you don’t know these, never fear! You don’t need a Ph.D. degree in these topics to get started but you do need a basic understanding.

#### (a) Learn Linear Algebra and Multivariate Calculus

Both Linear Algebra and Multivariate Calculus are important in Machine Learning. However, the extent to which you need them depends on your role as a data scientist. If you are more focused on application heavy machine learning, then you will not be that heavily focused on maths as there are many common libraries available. But if you want to focus on R&D in Machine Learning, then mastery of Linear Algebra and Multivariate Calculus is very important as you will have to implement many ML algorithms from scratch.

#### (b) Learn Statistics

Data plays a huge role in Machine Learning. In fact, around 80% of your time as an ML expert will be spent collecting and cleaning data. And statistics is a field that handles the collection, analysis, and presentation of data. So it is no surprise that you need to learn it!!!  
Some of the key concepts in statistics that are important are Statistical Significance, Probability Distributions, Hypothesis Testing, Regression, etc. Also, Bayesian Thinking is also a very important part of ML which deals with various concepts like Conditional Probability, Priors, and Posteriors, Maximum Likelihood, etc.

#### (c) Learn Python

Some people prefer to skip Linear Algebra, Multivariate Calculus and Statistics and learn them as they go along with trial and error. But the one thing that you absolutely cannot skip is [Python](https://www.geeksforgeeks.org/python-programming-language/)! While there are other languages you can use for Machine Learning like R, Scala, etc. Python is currently the most popular language for ML. In fact, there are many Python libraries that are specifically useful for Artificial Intelligence and Machine Learning such as [Keras](https://keras.io/), [TensorFlow](https://www.tensorflow.org/), [Scikit-learn](https://scikit-learn.org/stable/), etc.

So if you want to learn ML, it’s best if you learn Python! You can do that using various online resources and courses such as [**Fork Python**](https://practice.geeksforgeeks.org/courses/fork-python) available Free on GeeksforGeeks.

### Step 2 – Learn Various ML Concepts

Now that you are done with the prerequisites, you can move on to actually learning ML (Which is the fun part!!!) It’s best to start with the basics and then move on to the more complicated stuff. Some of the basic concepts in ML are:

#### (a) Terminologies of Machine Learning

* **Model –**A model is a specific representation learned from data by applying some machine learning algorithm. A model is also called a hypothesis.
* **Feature –**A feature is an individual measurable property of the data. A set of numeric features can be conveniently described by a feature vector. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like color, smell, taste, etc.
* **Target (Label) –**A target variable or label is the value to be predicted by our model. For the fruit example discussed in the feature section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.
* **Training –**The idea is to give a set of inputs(features) and it’s expected outputs(labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.
* **Prediction –**Once our model is ready, it can be fed a set of inputs to which it will provide a predicted output(label).

#### (b) Types of Machine Learning

* **Supervised Learning –**This involves learning from a training dataset with labeled data using classification and regression models. This learning process continues until the required level of performance is achieved.
* **Unsupervised Learning –**This involves using unlabelled data and then finding the underlying structure in the data in order to learn more and more about the data itself using factor and cluster analysis models.
* **Semi-supervised Learning –**This involves using unlabelled data like Unsupervised Learning with a small amount of labeled data. Using labeled data vastly increases the learning accuracy and is also more cost-effective than Supervised Learning.
* **Reinforcement Learning –**This involves learning optimal actions through trial and error. So the next action is decided by learning behaviors that are based on the current state and that will maximize the reward in the future.

### Advantages of Machine learning :-

#### 1. Easily identifies trends and patterns -

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

#### 2. No human intervention needed (automation)

With ML, you don’t need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwares; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

#### 3. Continuous Improvement

As [**ML algorithms**](https://data-flair.training/blogs/machine-learning-algorithms/) gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

#### 4. Handling multi-dimensional and multi-variety data

Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

#### 5. Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

### Disadvantages of Machine Learning :-

#### 1. Data Acquisition

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

#### 2. Time and Resources

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

#### 3. Interpretation of Results

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

#### 4. High error-susceptibility

[Machine Learning](https://en.wikipedia.org/wiki/Machine_learning) is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

**Python Development Steps : -**

Guido Van Rossum published the first version of Python code (version 0.9.0) at alt.sources in February 1991. This release included already exception handling, functions, and the core data types of list, dict, str and others. It was also object oriented and had a module system.  
Python version 1.0 was released in January 1994. The major new features included in this release were the functional programming tools lambda, map, filter and reduce, which Guido Van Rossum never liked.Six and a half years later in October 2000, Python 2.0 was introduced. This release included list comprehensions, a full garbage collector and it was supporting unicode.Python flourished for another 8 years in the versions 2.x before the next major release as Python 3.0 (also known as "Python 3000" and "Py3K") was released. Python 3 is not backwards compatible with Python 2.x. The emphasis in Python 3 had been on the removal of duplicate programming constructs and modules, thus fulfilling or coming close to fulfilling the 13th law of the Zen of Python: "There should be one -- and preferably only one -- obvious way to do it."Some changes in Python 7.3:

* Print is now a function
* Views and iterators instead of lists
* The rules for ordering comparisons have been simplified. E.g. a heterogeneous list cannot be sorted, because all the elements of a list must be comparable to each other.
* There is only one integer type left, i.e. int. long is int as well.
* The division of two integers returns a float instead of an integer. "//" can be used to have the "old" behaviour.
* Text Vs. Data Instead Of Unicode Vs. 8-bit

**Purpose :-**

We demonstrated that our approach enables successful segmentation of intra-retinal layers—even with low-quality images containing speckle noise, low contrast, and different intensity ranges throughout—with the assistance of the ANIS feature.

**Python**

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

* Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

**Modules Used in Project :-**

**Tensorflow**

TensorFlow is a [free](https://en.wikipedia.org/wiki/Free_software) and [open-source](https://en.wikipedia.org/wiki/Open-source_software) [software library for dataflow and differentiable programming](https://en.wikipedia.org/wiki/Library_(computing)) across a range of tasks. It is a symbolic math library, and is also used for [machine learning](https://en.wikipedia.org/wiki/Machine_learning) applications such as [neural networks](https://en.wikipedia.org/wiki/Neural_networks). It is used for both research and production at [Google](https://en.wikipedia.org/wiki/Google).‍

TensorFlow was developed by the [Google Brain](https://en.wikipedia.org/wiki/Google_Brain) team for internal Google use. It was released under the [Apache 2.0](https://en.wikipedia.org/wiki/Apache_License) [open-source license](https://en.wikipedia.org/wiki/Open-source_license) on November 9, 2015.

**Numpy**

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

* A powerful N-dimensional array object
* Sophisticated (broadcasting) functions
* Tools for integrating C/C++ and Fortran code
* Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

**Pandas**

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

**Matplotlib**

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and [IPython](http://ipython.org/) shells, the [Jupyter](http://jupyter.org/) Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the [sample plots](https://matplotlib.org/tutorials/introductory/sample_plots.html) and [thumbnail gallery](https://matplotlib.org/gallery/index.html).

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

**Scikit – learn**

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. **Python**

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace.

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

* Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

**Install Python Step-by-Step in Windows and Mac :**

Python a versatile programming language doesn’t come pre-installed on your computer devices. Python was first released in the year 1991 and until today it is a very popular high-level programming language. Its style philosophy emphasizes code readability with its notable use of great whitespace.

The object-oriented approach and language construct provided by Python enables programmers to write both clear and logical code for projects. This software does not come pre-packaged with Windows.

## How to Install Python on Windows and Mac :

There have been several updates in the Python version over the years. The question is how to install Python? It might be confusing for the beginner who is willing to start learning Python but this tutorial will solve your query. The latest or the newest version of Python is version 3.7.4 or in other words, it is Python 3.

**Note:** The python version 3.7.4 cannot be used on Windows XP or earlier devices.

Before you start with the installation process of Python. First, you need to know about your **System Requirements**. Based on your system type i.e. operating system and based processor, you must download the python version. My system type is a **Windows 64-bit operating system**. So the steps below are to install python version 3.7.4 on Windows 7 device or to install Python 3. [Download the Python Cheatsheet here.](https://myelearninghub.com/python-cheat-sheet/)The steps on how to install Python on Windows 10, 8 and 7 are **divided into 4 parts** to help understand better.

**PROPOSE OF WORK :**

The algorithm describes about the comparison between the images. In the initial stage the two images are loaded into the work space. Later the images are re-sized in-order to get the same size for both the images so that the result will be accurate when compared. The image is also compressed in order to improve the accuracy while comparing. Image is automatically saved in a folder after compression. To over- come the weakness of having images with different colours, the image is converted into grayscale image. In the next step the similar features of the images are detected and the comparison is done using image subtraction method where each pixel value of an image is compared with pixel value of another image and finally the result will be displayed

**SCOPE OF WORK :**

In recent times, the use of internet has widely increased which is leading to easy opportunity of plagiarism, the proposed system will help in detecting the same. So the plagiarism detection will be very helpful in the future. Our system can also be used as ‘search by image

**2.LITERATURE SURVEY:**

**\* SURVEY AND STUDY OF PUBLISHED LITERTURE ON THE ASSIGNED TOPIC:**

**Plagiarism-a survey**

**AUTHORS :** **H. A. Maurer, F. Kappe, and B. Zaka**

**ABSTRACT:**

Plagiarism in the sense of "theft of intellectual property" has been around for as long as humans have produced work of art and research. However, easy access to the Web, large databases, and telecommunication in general, has turned plagiarism into a serious problem for publishers, researchers and educational institutions. In this paper, we concentrate on textual plagiarism (as opposed to plagiarism in music, paintings, pictures, maps, technical drawings, etc.). We first discuss the complex general setting, then report on some results of plagiarism detection software and finally draw attention to the fact that any serious investigation in plagiarism turns up rather unexpected side-effects. We believe that this paper is of value to all researchers, educators and students and should be considered as seminal work that hopefully will encourage many still deeper investigations.

**Introduction to qualitative research methods: A guidebook and resource**

**AUTHORS :** **S. J. Taylor, R. Bogdan, and M. DeVault,**

**ABSTRACT:**

presentation of qualitative data. Unique in the market, this book describes the entire research process — from design through writing — illustrated by examples of real, complete qualitative work that clearly demonstrates how methods are used in actual practice. This updated fourth edition includes all new case studies, with additional coverage of mixed methods, non-sociological settings, funding, and a sample interview guide. The studies profiled are accompanied by observation field notes, and the text includes additional readings for both students and instructors. More than just theory, this guide is designed to give you a real-world practitioner's view of how qualitative research is handled every step of the way. Many different disciplines rely on qualitative research as a method of inquiry, to gain an in-depth understanding of human behavior and the governing forces behind it. Qualitative research asks "why" and "how," and the data is frequently complex and difficult to measure. This book shows you how to effectively handle qualitative work, regardless of where it's being applied.

**5.PREMILINARY ANALYSIS:**

Plagiarism can be defined as an unfair appropriation, theft and publication of language, thoughts, ideas or statements of other authors, and redefine they as their main work and in their own name, or stated a thought so that those ideas stay as vague and non-transparent or illegal. This phenomenon, often symmetric in today research and in the creation of information technology and have many influenced in the intensification of consciously or unconsciously plagiarism [1]. Because of its importance, from 2009, Annual international Practical tournament of plagiarism detection evaluation methods, identify the author and related works be held at a conference entitled PAN / CLEF. Since 2012, PAN has developed a web service as TIRA and urged participants, instead of sending the output of its implementation, submitted running software [2]. Plagiarism detection by identifying the source is based on identify the text in massive amounts of documents. In plagiarism Identify without a source, the main focus is on the writers written pattern and structure of the text itself [

**5.1ABOUT INPUT DATA:**

**LIST OF RESARCH GAPS:**

[1] H. A. Maurer, F. Kappe, and B. Zaka. "Plagiarism-a survey." J. UCS 12.8 (2006): 1050-1084.

[2] S. J. Taylor, R. Bogdan, and M. DeVault, "Introduction to qualitative research methods: A guidebook and resource," John Wiley & Sons, 2015.

[3] M. A. M. Noh, A. L. A. Nudin, N. Narimah, and W. N. K. Shamsudin "Plagiarism in Graphic Design." Proceedings of the Art and Design International Conference (AnDIC 2016). Springer, Singapore, 2018.

[4] N. Meuschke, C. Gondek, D. Seebacher, C. Breitinger, D. Keim, and B. Gipp "An adaptive image-based plagiarism detection approach." Proc. ACM/IEEE Joint Conf. on Digital Libraries (JCDL). 2018.

[5] X. Bo, W. Junwen, L. Guangjie, and D. Yuewei, "Image copy-move forgery detection based on SURF." Multimedia information networking and security (MINES), 2010 international conference on. IEEE, 2010.

[6] A. Alahmadi, Mu. Hussain, H. Aboalsamh, G. Muhammad, G. Bebis, and H. Mathkour, "Passive detection of image forgery using DCT and local binary pattern." Signal, Image and Video Processing 11.1 (2017): 81-88.

[7] B. Peng, W. Wang, J. Dong, and T. Tan, "Optimized 3D lighting environment estimation for image forgery detection." IEEE Transactions on Information Forensics and Security 12.2 (2017): 479-494.

[8] R. E. J. Granty, T. S. Aditya, and S. S. Madhu, "Survey on passive methods of image tampering detection", Proc. of the International Conference on Communication and Computational Intelligence, Page(s): 431 ̽ 436, 2010.

[9] D. G. Lowe, Distinctive image features from scale-invariant keypoints. Internationaljournal of computer vision, 60(20): 91-110, 2004.

[10] M. Sharma, , G. N. Purohit, and S. Mukherjee. "Information Retrieves from Brain MRI Images for Tumor Detection Using Hybrid Technique K-means and Artificial Neural Network (KMANN)." Networking Communication and Data Knowledge Engineering. Springer, Singapore, p.p. 145-157, 2018.

**3 PROBLEM STATEMENT :**

**\*EXISTING SYSTEM:**

The existing methodology maybe sufficient for detecting plagiarism of images when the source and suspected image have not been rotated by a large margin, but in case of rotational changes the existing methodology will fail. The proposed methodology will ensure that even if the image is rotated plagiarism is detected if it has occurred or if an attack of rotational change has been made. Also the existing system is not efficient to detect plagiarism properly for different types of images. The proposed system will ensure that by using adaptive threshold values. The algorithm makes sure that the matching time of the images is less by reducing the search field by a significant factor each time the refinement is done

**\*PROBLEM DEFINATION :**

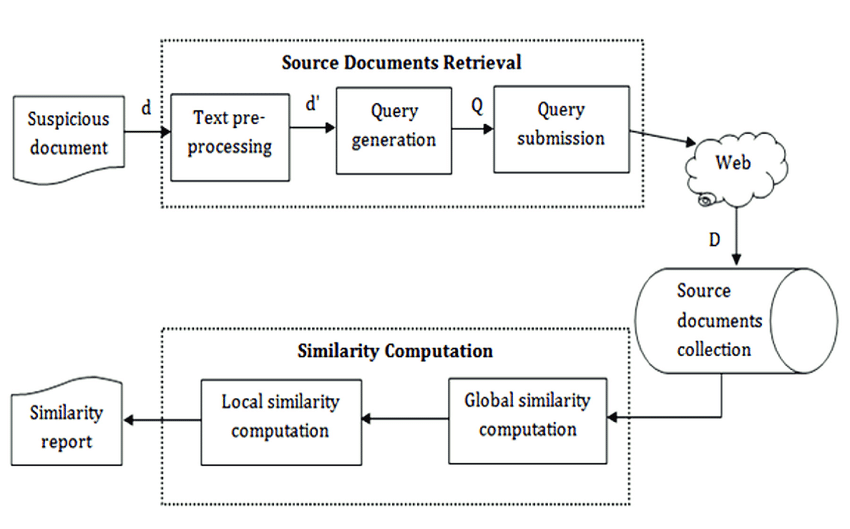
A research problem is defined as an area of concern that requires a meaningful understanding of a specific topic, a condition, a contradiction, or a difficulty. So what is research problem? A research problem means finding answers to questions or strengthening existing findings to bridge the knowledge gap to solve problems

**PROBLEM STATEMENT :**

The Indeed Editorial Team comprises a diverse and talented team of writers, researchers and subject matter experts equipped with Indeed's data and insights to deliver useful tips to help guide your career journey. A problem statement addresses issues in a timely and efficient manner. They help professionals break down complex situations into tangible goals that they can then communicate throughout an organisation. In every workplace, problems are inevitable. Thus, a problem statement is an effective tool to put into practice so that employees recognise issues before they disrupt multiple functions of the business. In this article, we discuss what a problem statement is, why they are important, how to write one and provide a comprehensive template and example for your reference**.**

**4. SYSTEM ARCHITECTURE :**

**\*NEAT BLOCK DIAGRAM:**



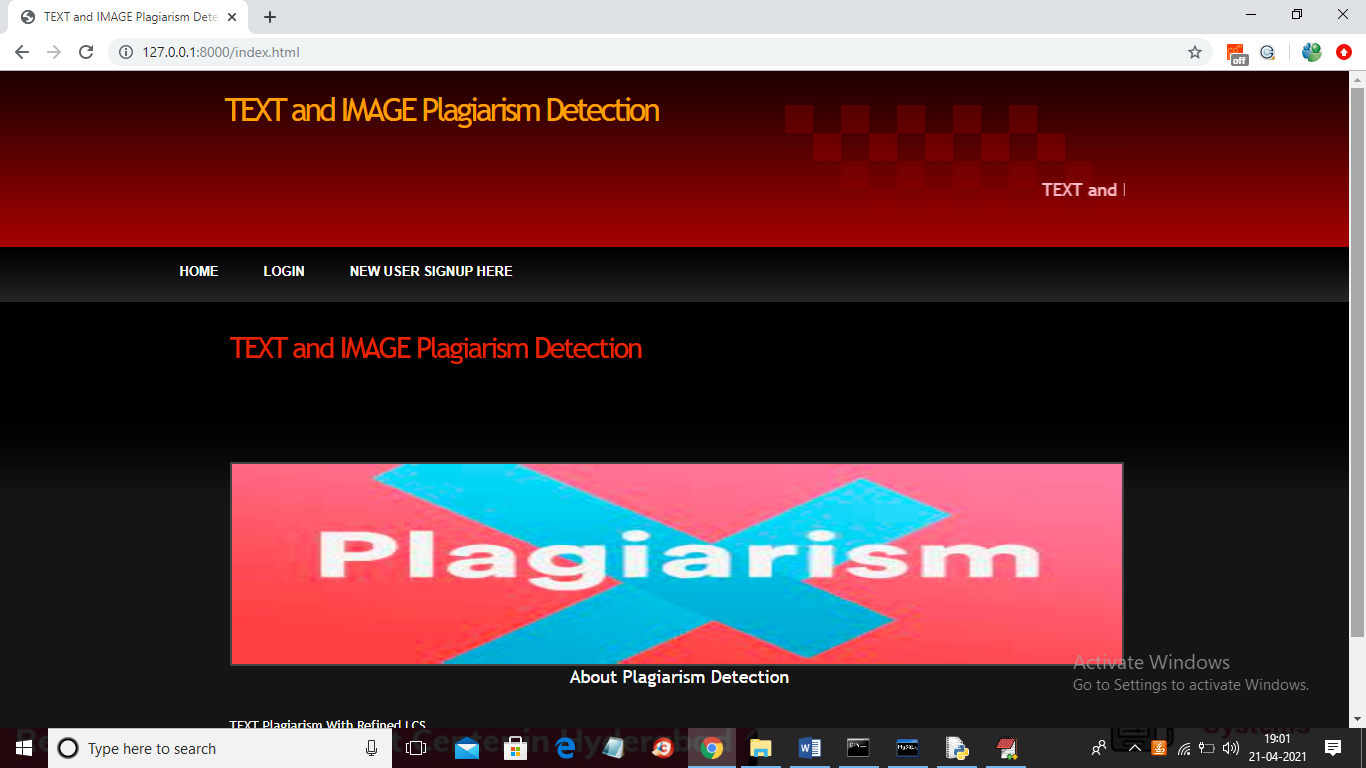
**\*MODEULES EXPLANATION :**

* **1.New user Signup**
* Firstly user will register in to Application.It helpful to login into Application with username and password.
* 2.LOGIN :
* User will login into Application through username and password**.**
* 3.UPLOAD SOURCE FILE :
* Folder is created into Upload Source Files’ link to load all files from corpus folder
* 4.UPLOAD SUSPICIOUS FILE :
* To load suspicious file and get result.user will upload file to Upload Suspicious files the result is execute. LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result.
* 5.UPLOAD SOURCE IMAGE :
* In this module from all database images histogram will be calculated and store in array and whenever we upload new test image then both histogram will get matched.
* 6.UPLOAD SUSPICOUS IMAGE :
* we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected. histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarised and now upload image from “images” folder and see result. we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now get below result. histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result

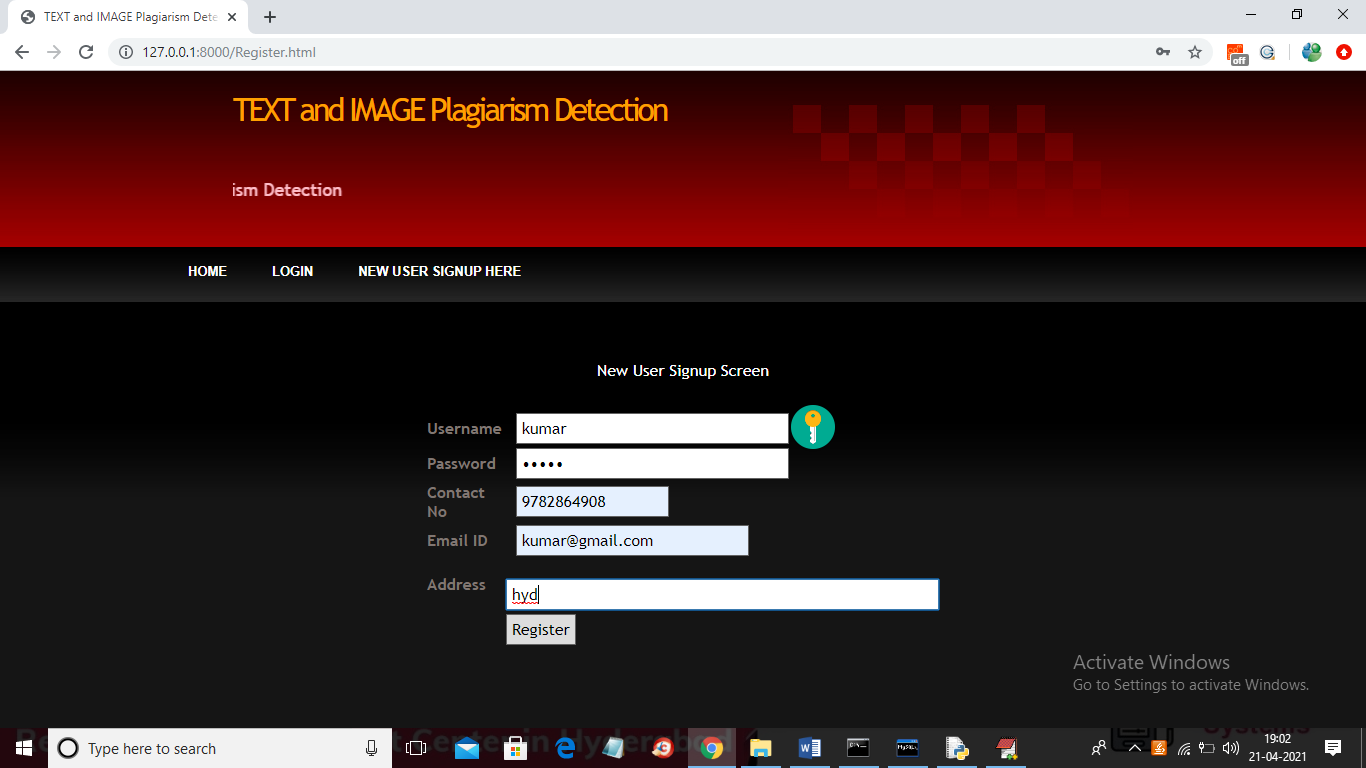
**5 .PRELIMINARY ANALYSIS :**

**\*BREIF ABOUT INPUT DATA:**

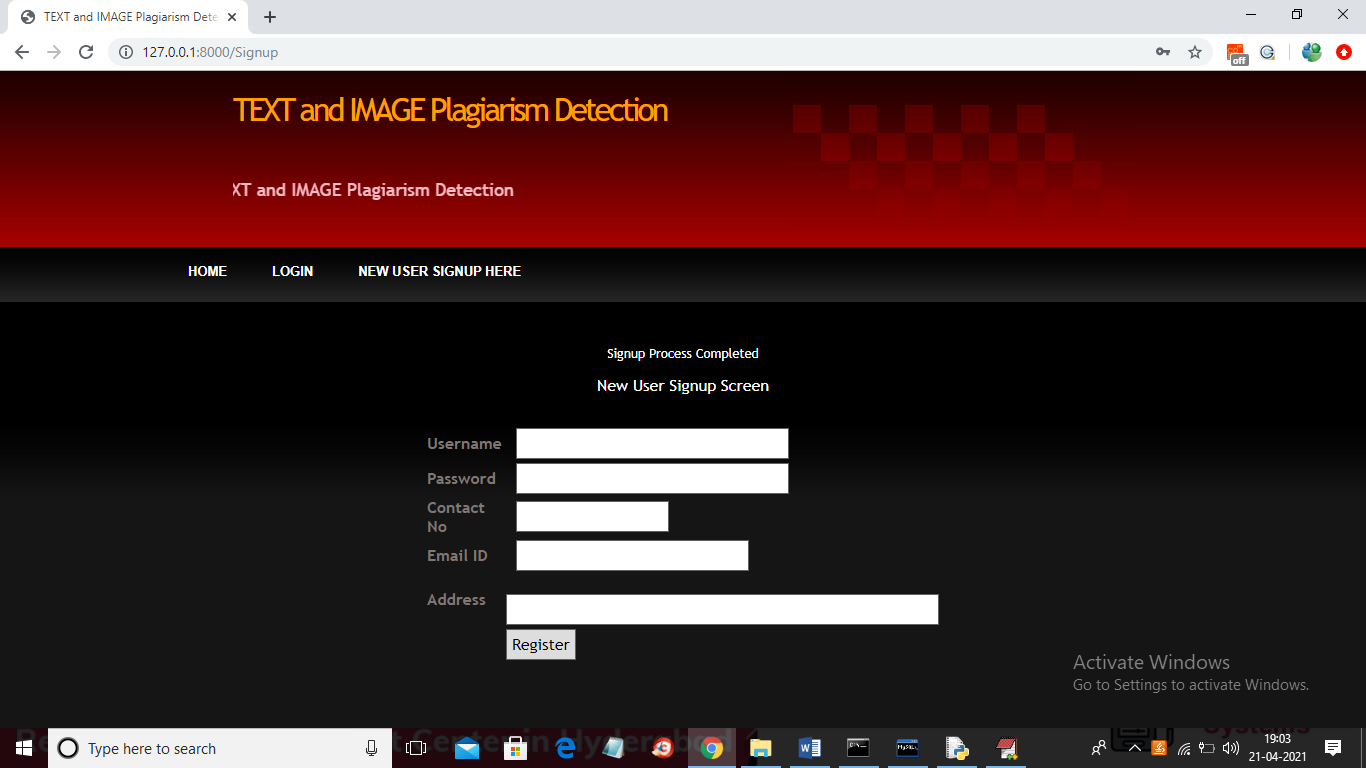
To run project install python 3.7 and then install DJANGO server and deploy code on that server and run from browser to get below screen



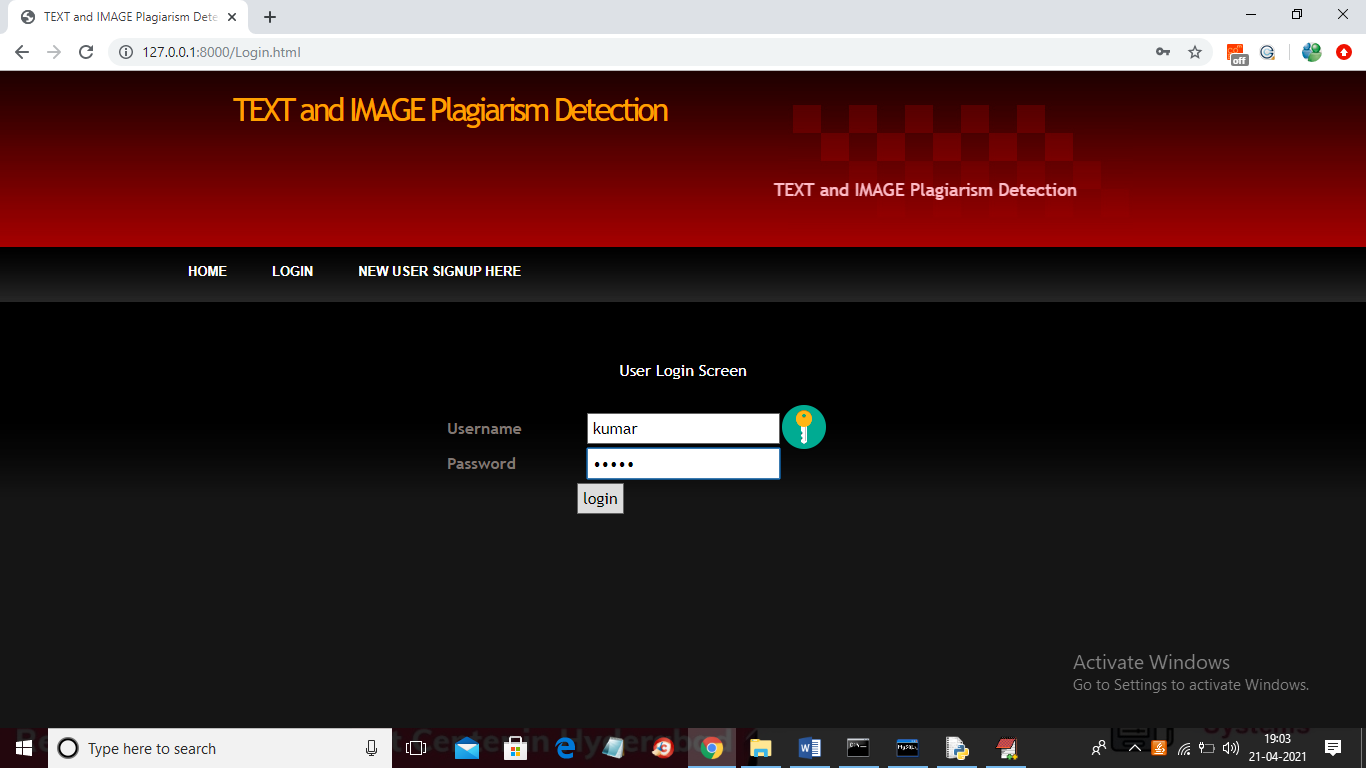
In above screen click on ‘New User Signup Here’ link to get below screen



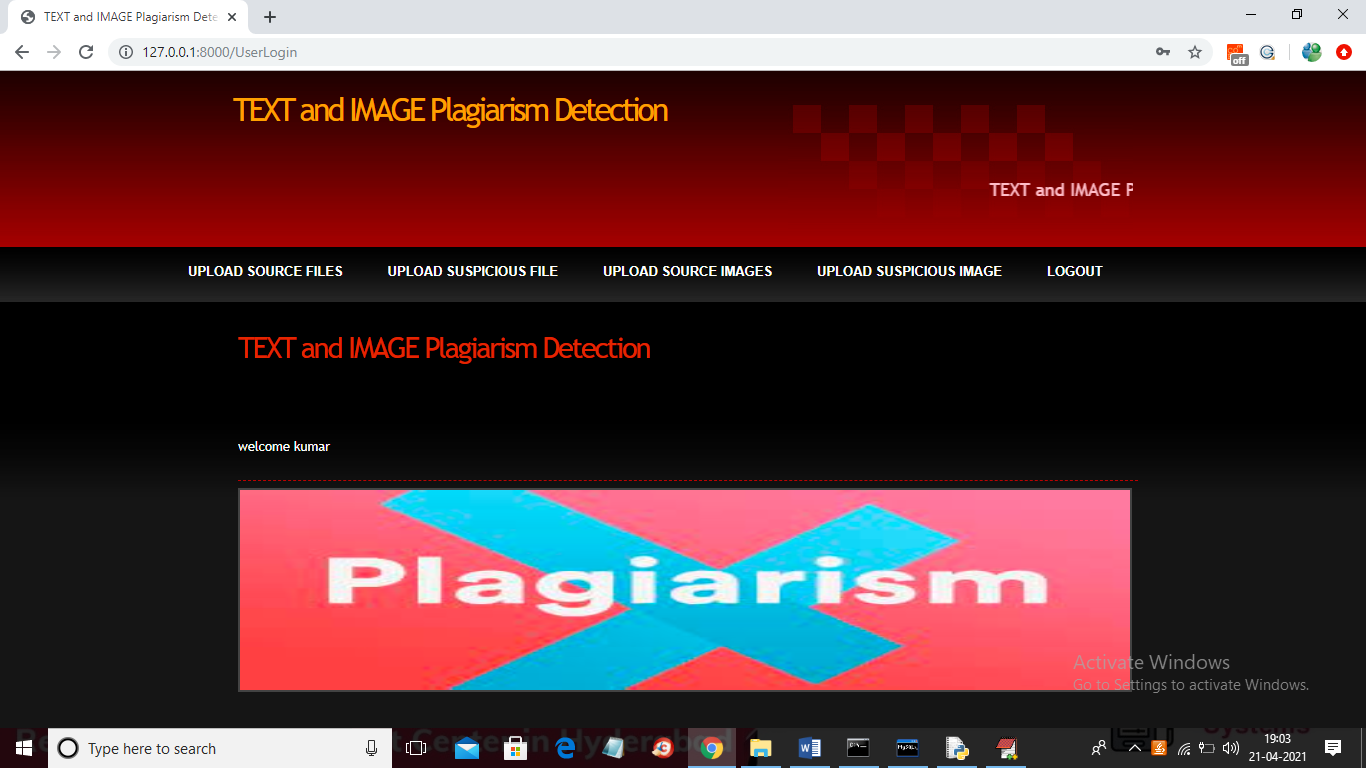
In above screen user signup details entered and then click on ‘Register’ button to get below screen



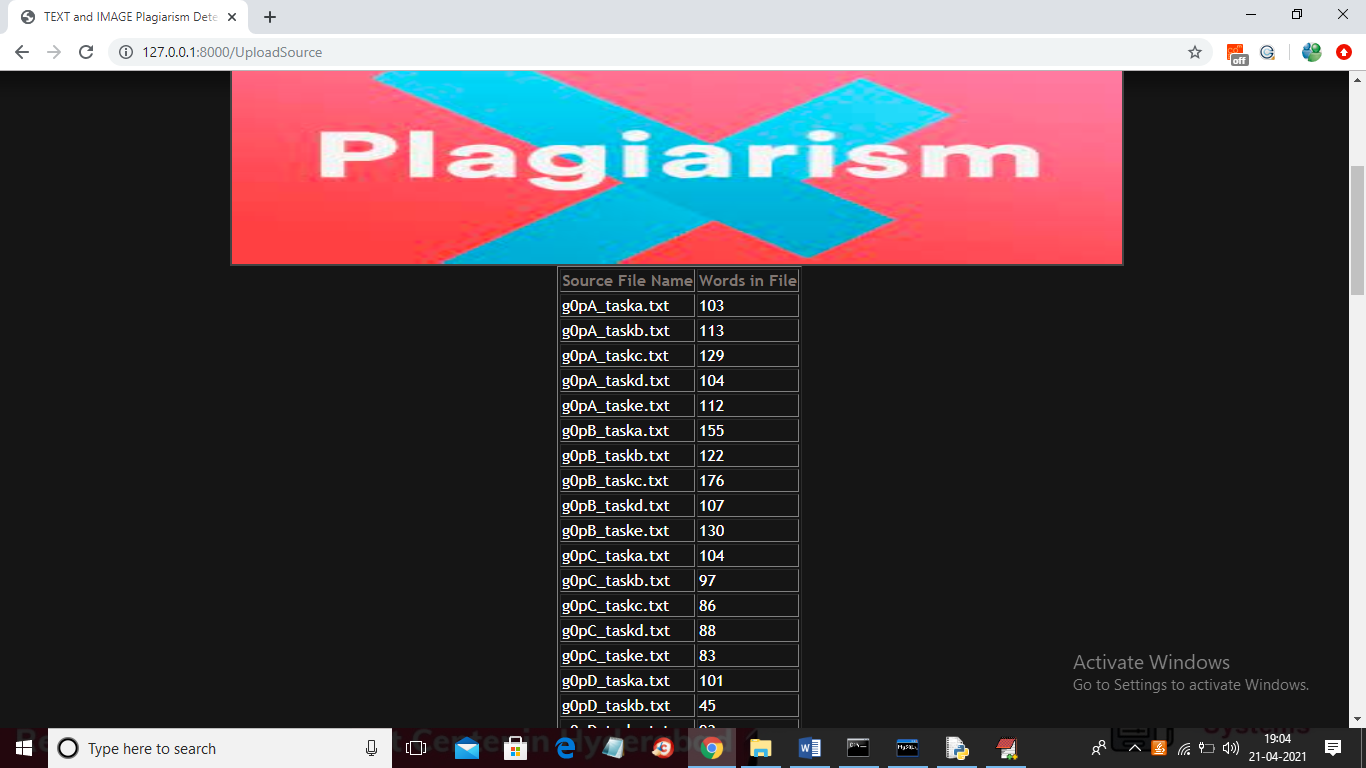
In above screen user signup process completed and now click on ‘Login’ link to get below screen



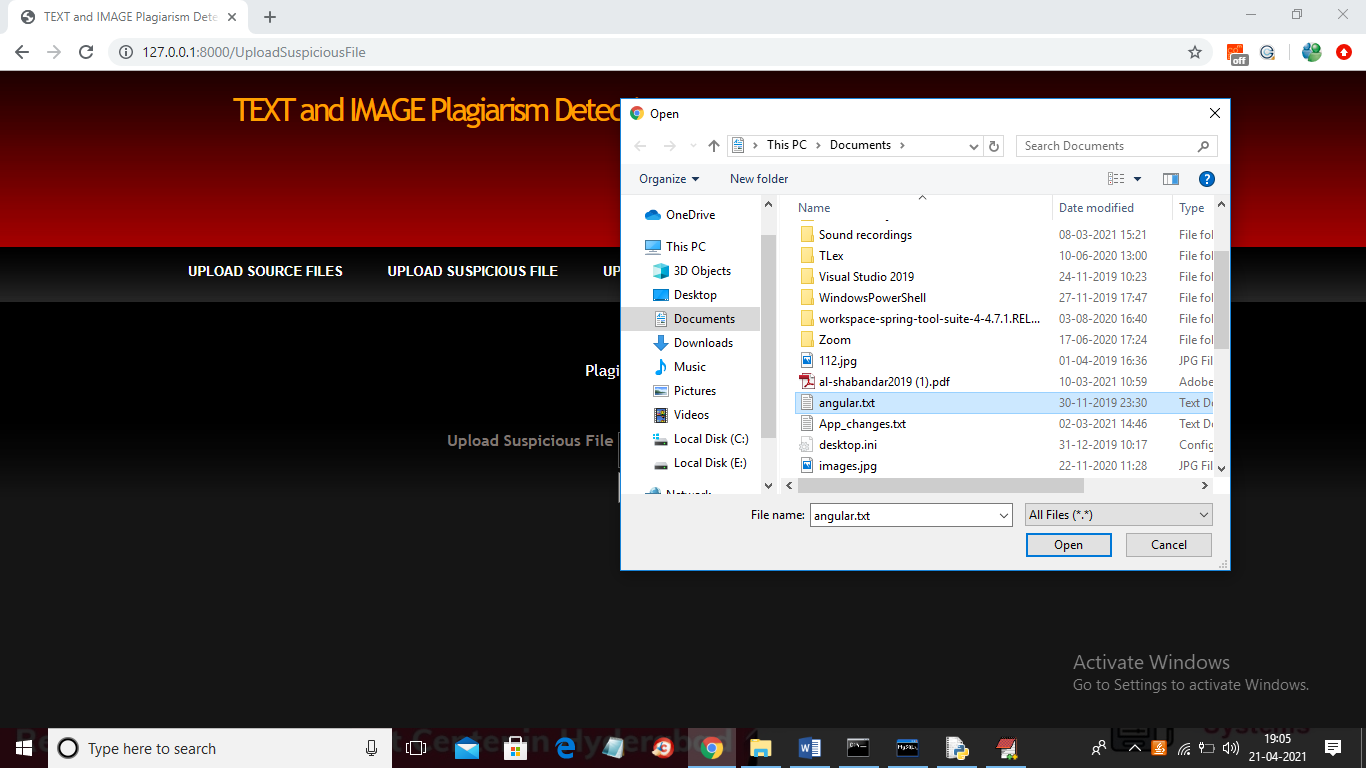
In above screen user is login and then click on button to get below screen



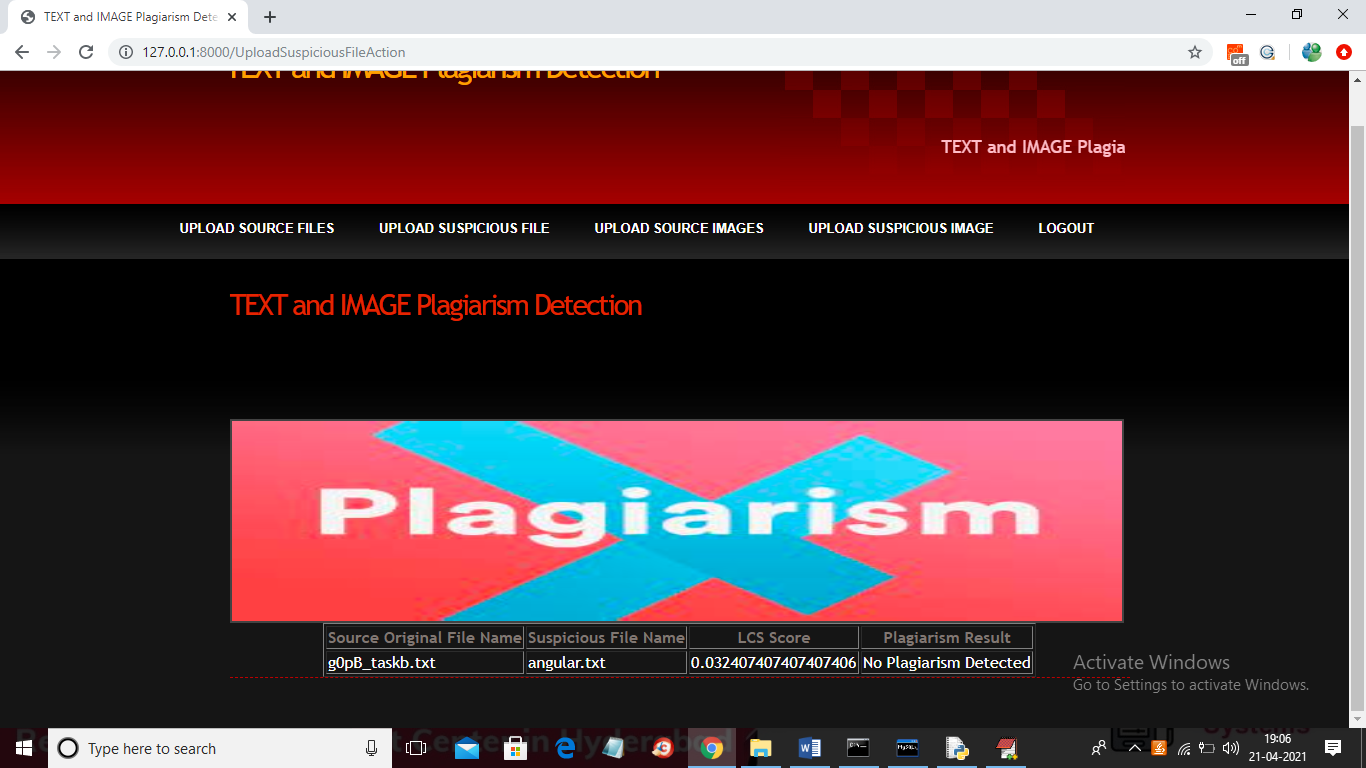
In above screen click on ‘Upload Source Files’ link to load all files from corpus folder



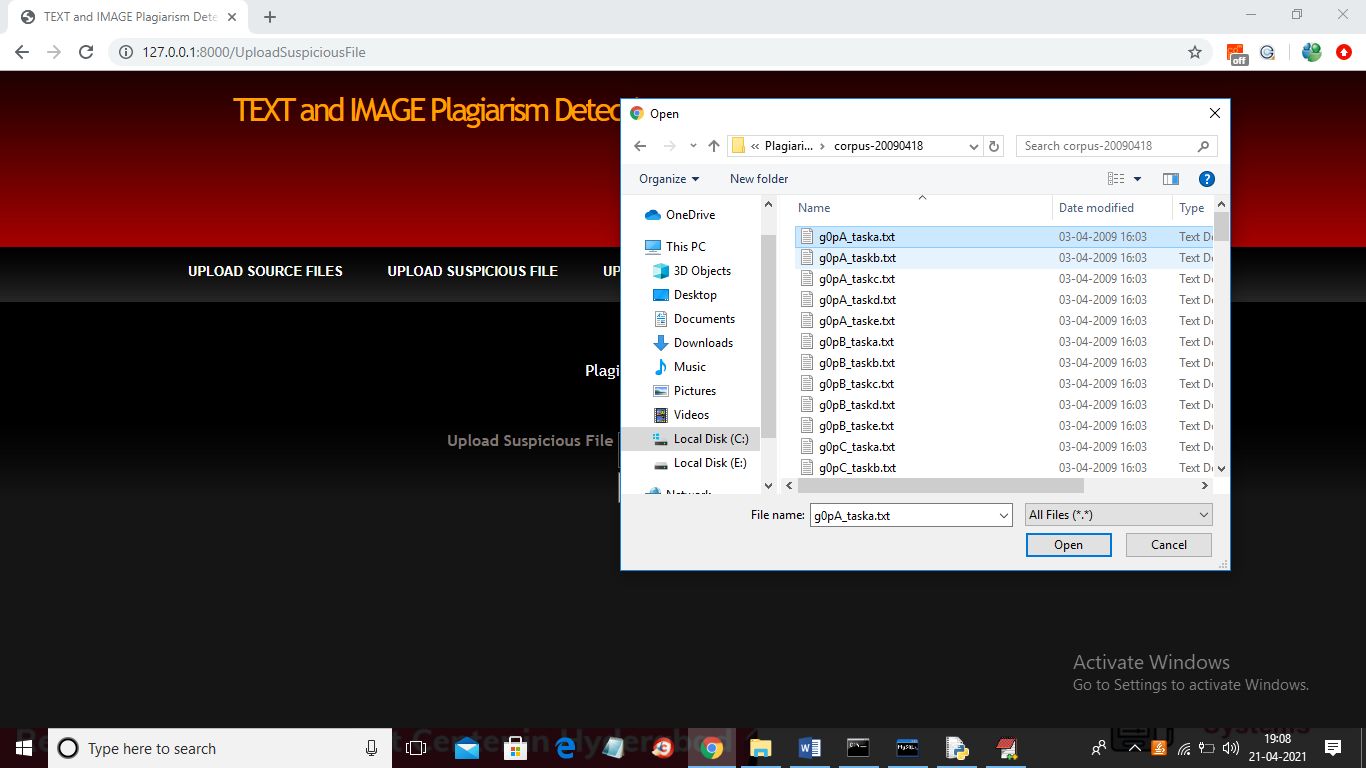
In above screen all files are loaded now click on ‘Upload Suspicious File’ button to load suspicious file and get result



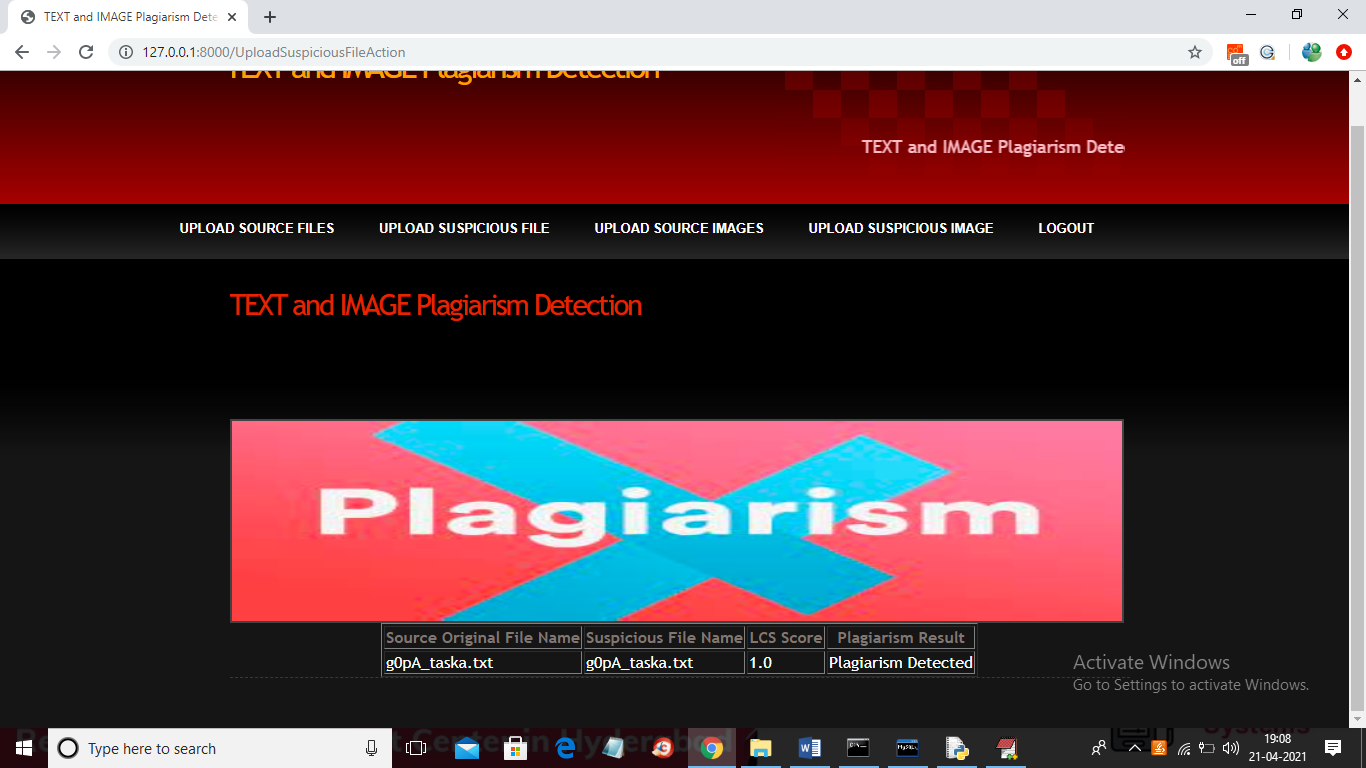
In above screen I am selecting and uploading ‘angular.txt’ file and then click on ‘Open’ button to get below result and then click on ‘Check Plagiarism’ button to get result



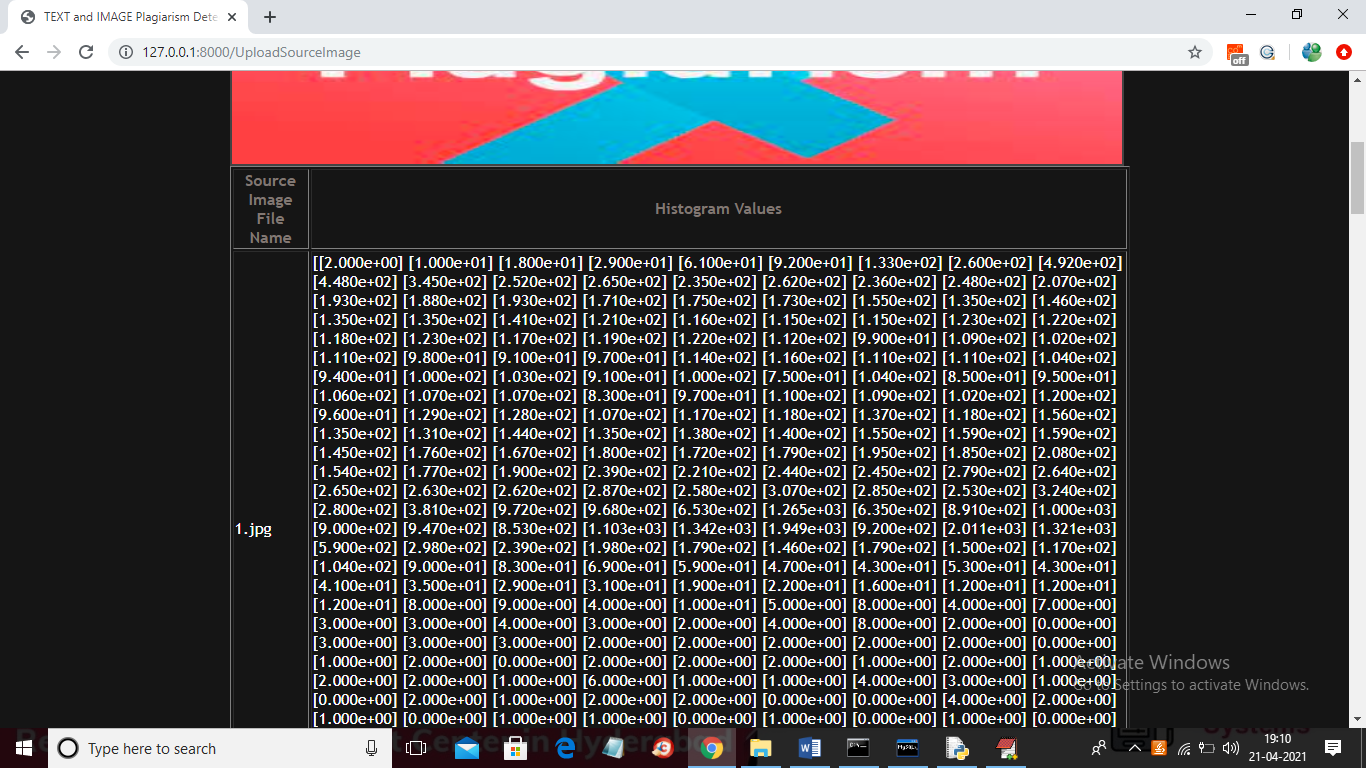
In above screen angular.txt file matched very little with g)pB\_taskb.txt corpus file and we got similarity score as 0.03 so no plagiarism detected and now upload any file from corpus and see result



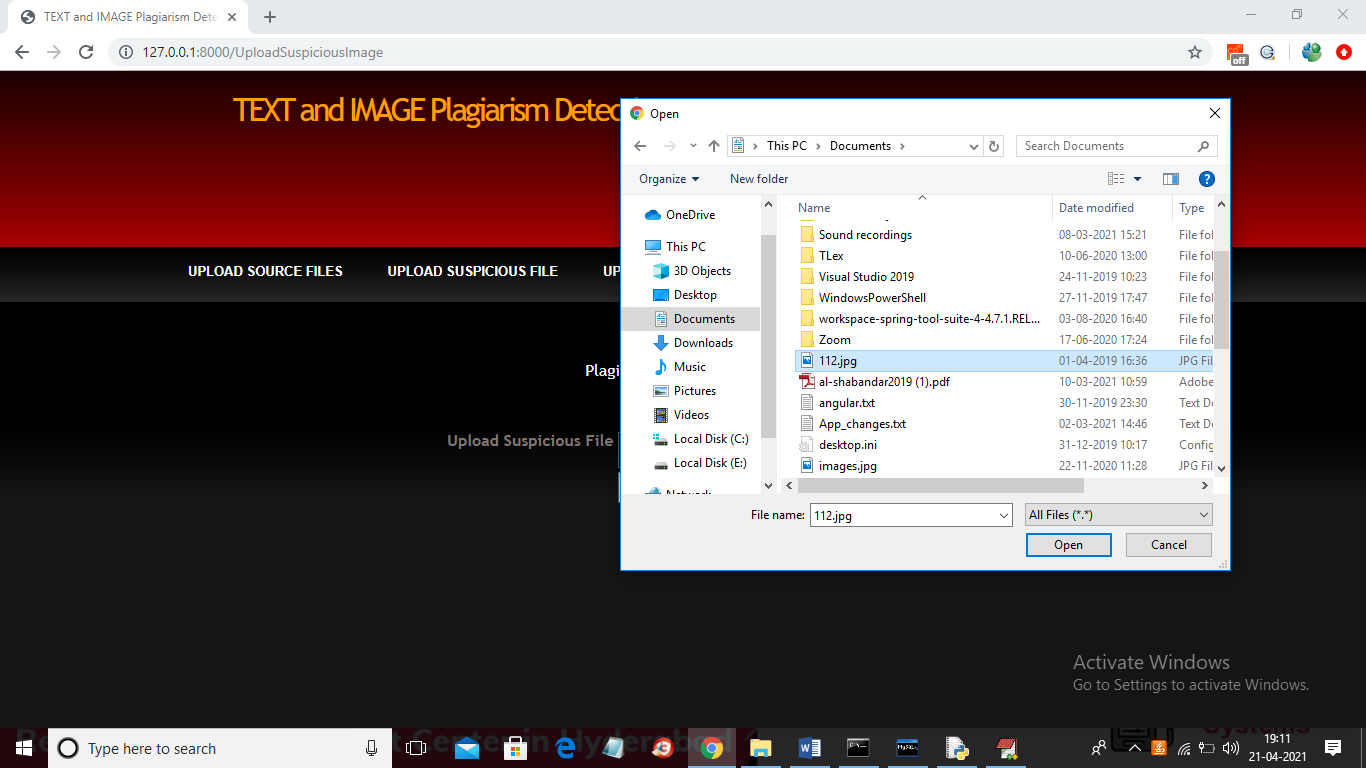
In above screen I am selecting and uploading first file and then click on button to get below result



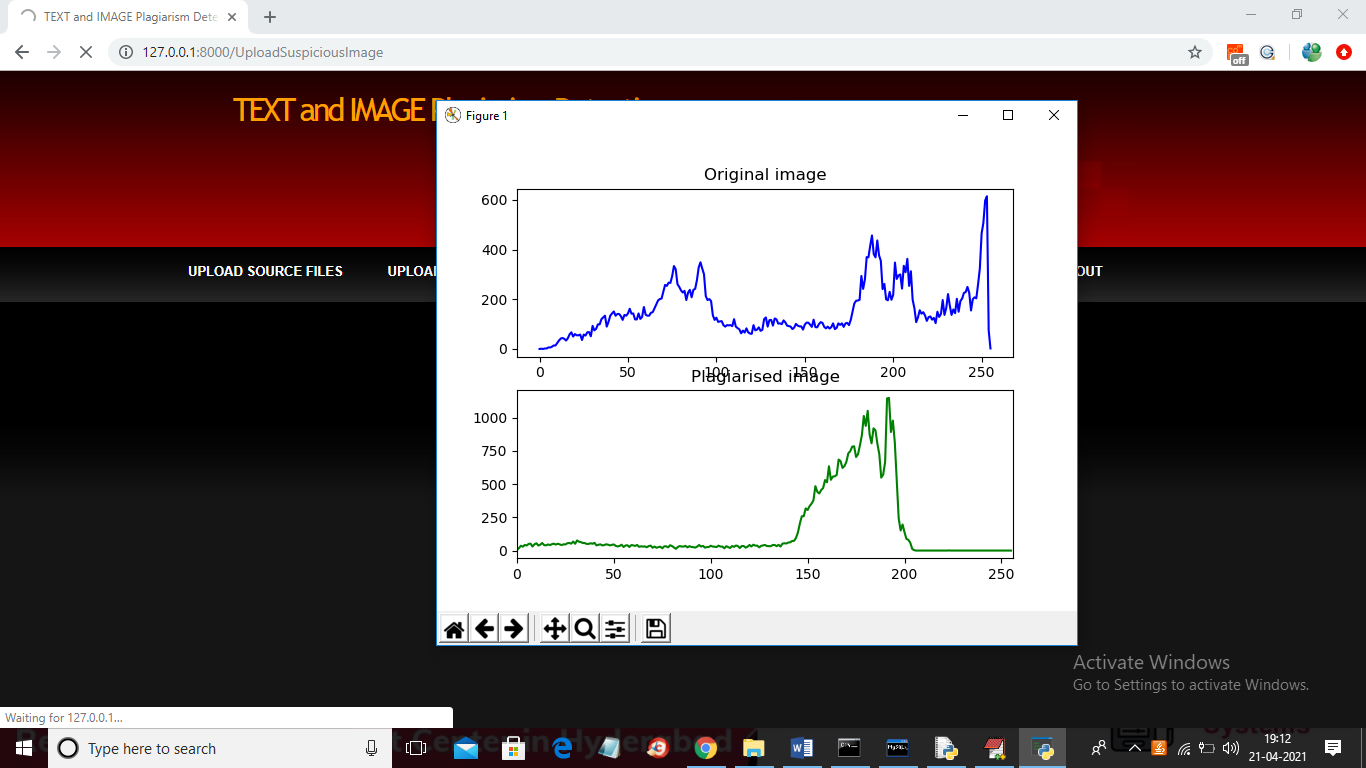
In above screen LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result. Now click on ‘Upload Source Images’ link to upload all images from ‘images’ folder



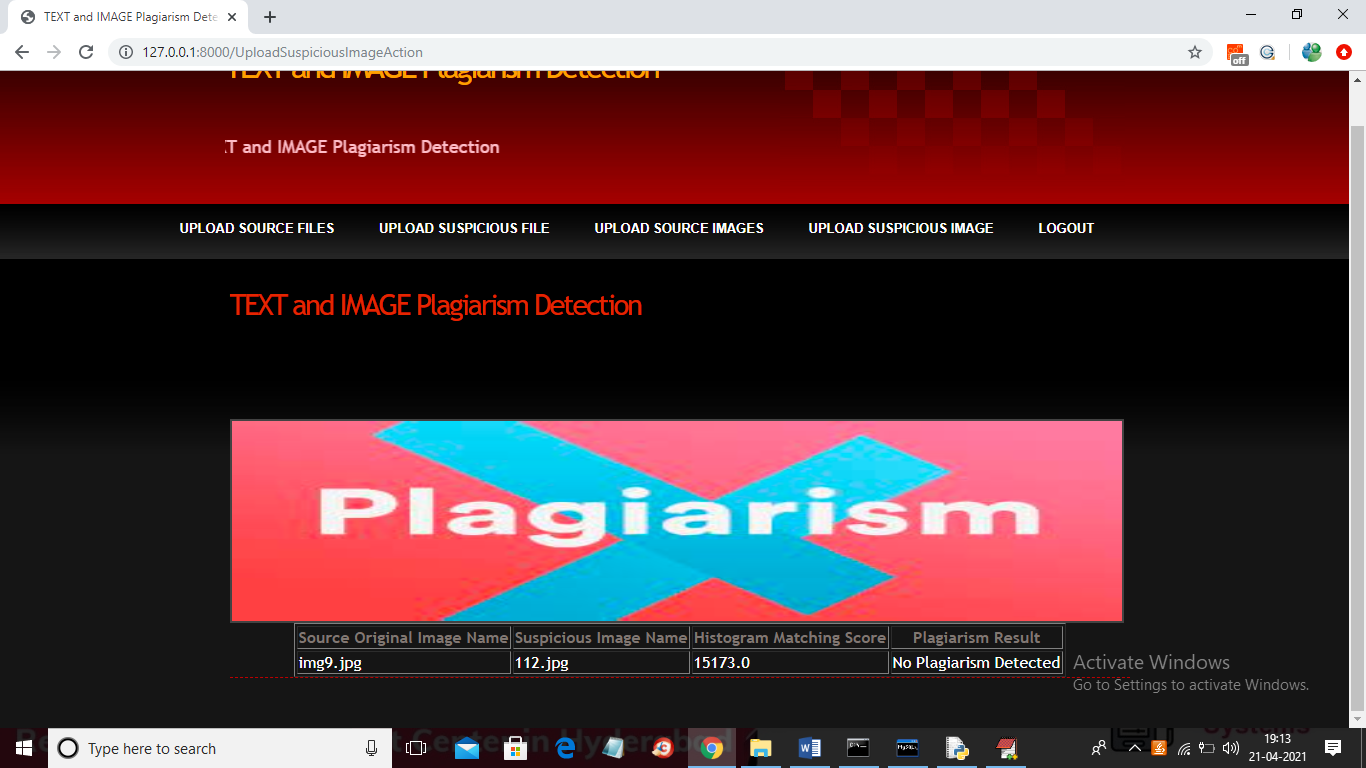
In above screen from all database images histogram will be calculated and store in array and whenever we upload new test image then both histogram will get matched and now click on ‘Upload Suspicious Image’ link to upload some image



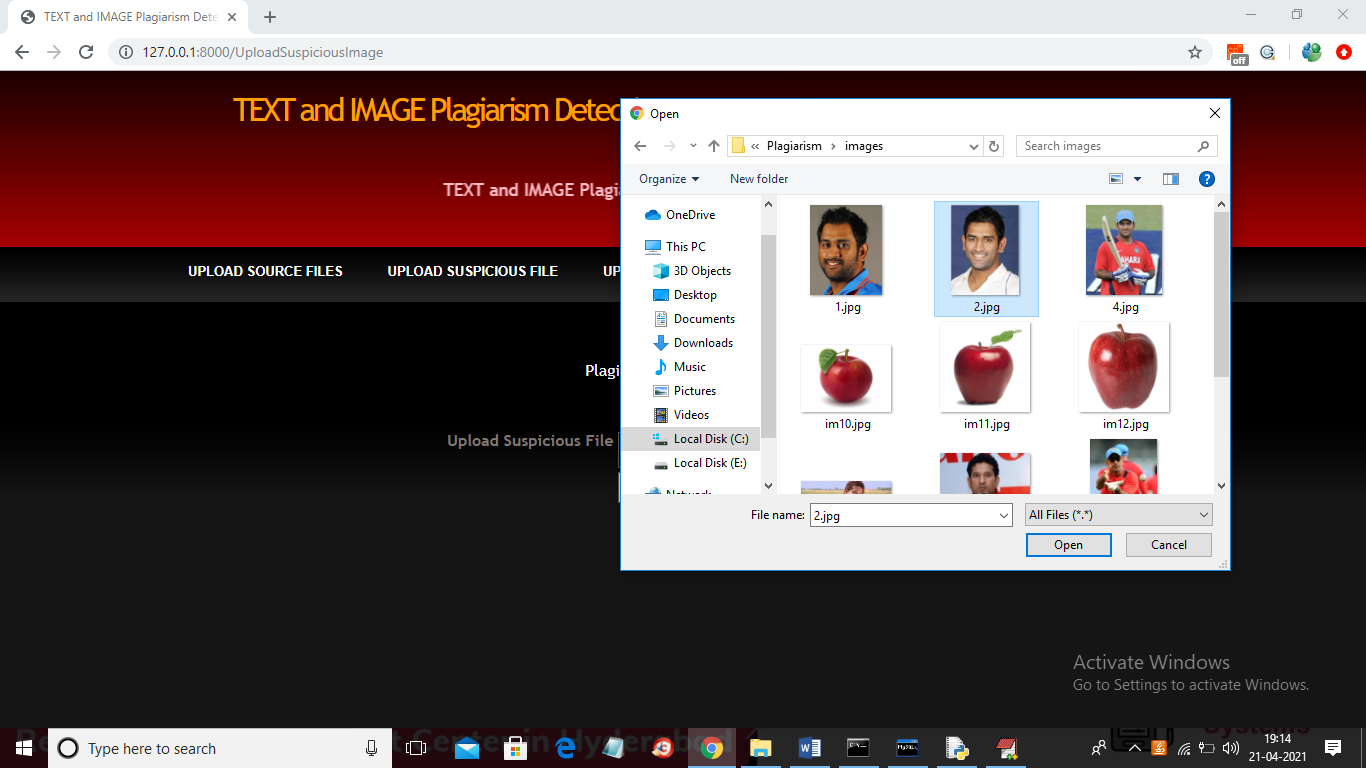
In above screen I am selecting and uploading ‘112.jpg’ file and then click on ‘Open’ button to get below result



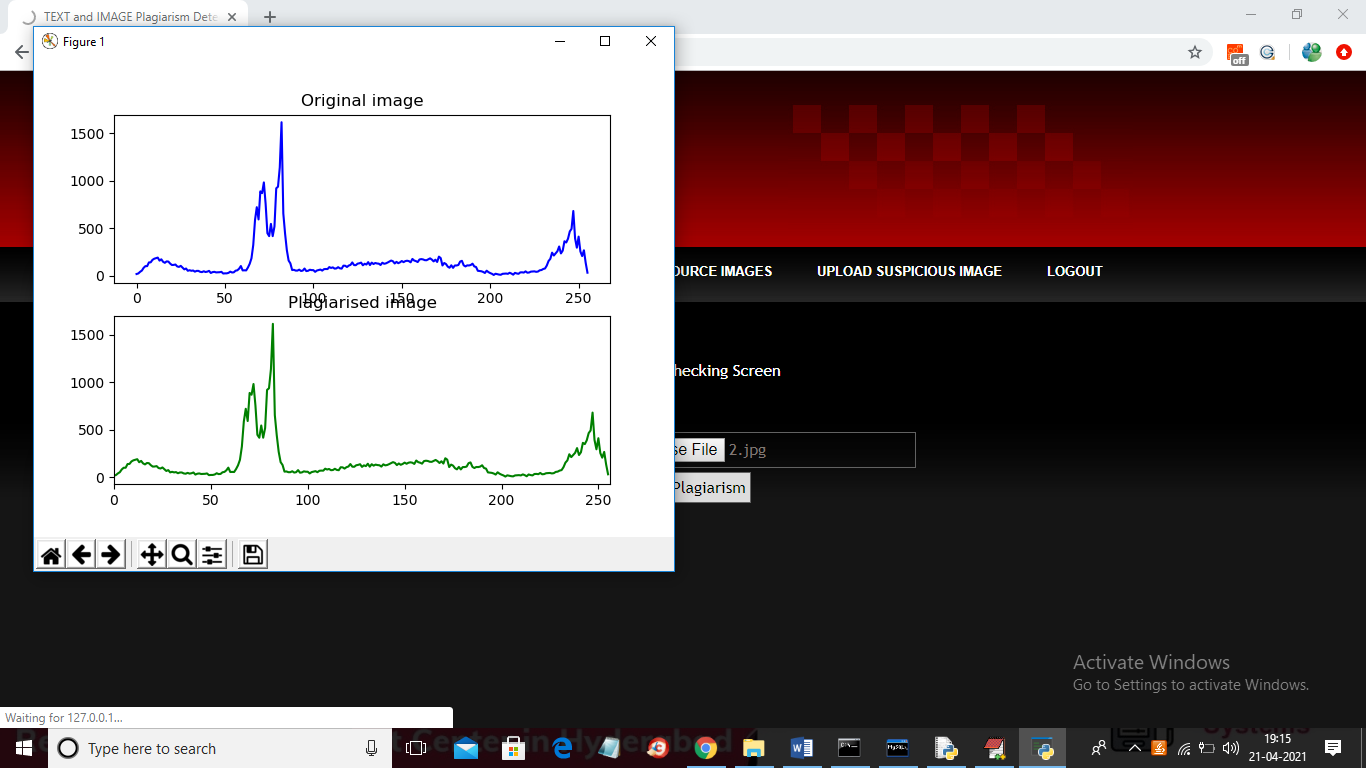
In above screen we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected and now close above graph to get below result



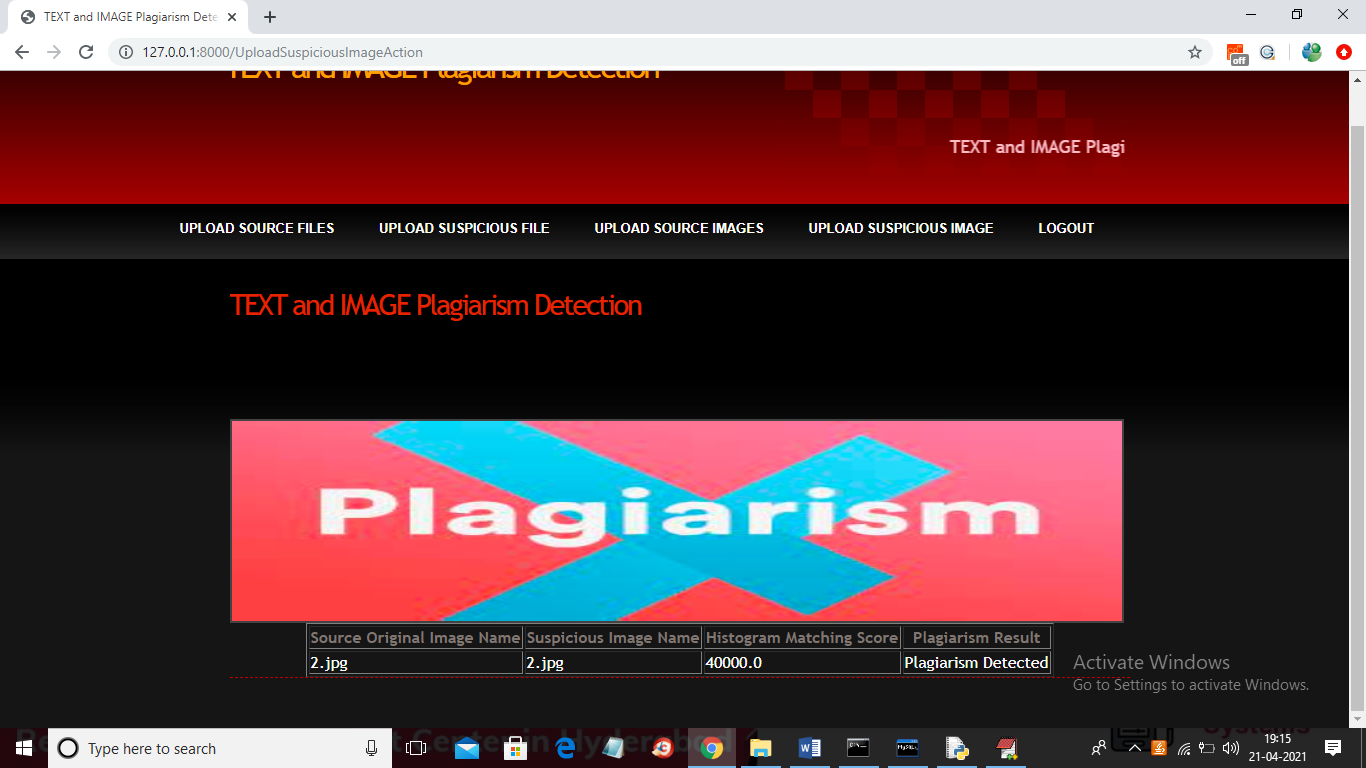
In above screen histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarised and now upload image from “images” folder and see result



In above screen I am selecting and uploading ‘2.jpg’ file from “images” database folder and below is the result



In above screen we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now close above graph to get below result



In above screen histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result.

Similarly u can upl

**\*TYPE OF ANALYSIS DOING ON DATA :**

1. [Types of data analysis: Descriptive](https://careerfoundry.com/en/blog/data-analytics/different-types-of-data-analysis/#descriptive-analytics-what-happened)
2. [Types of data analysis: Diagnostic](https://careerfoundry.com/en/blog/data-analytics/different-types-of-data-analysis/#diagnostic-analytics-why-did-it-happen)
3. [Types of data analysis: Predictive](https://careerfoundry.com/en/blog/data-analytics/different-types-of-data-analysis/#predictive-analytics-what-is-likely-to-happen-in-the-future)
4. [Types of data analysis: Prescriptive](https://careerfoundry.com/en/blog/data-analytics/different-types-of-data-analysis/#prescriptive-analytics-whats-the-best-course-of-action)
5. [Key takeaways and further reading](https://careerfoundry.com/en/blog/data-analytics/different-types-of-data-analysis/#key-takeaways-and-further-reading)

**6. FEASIBILITY STUDY:**

**WORKING OUT A PRELIMINARY APPROACH TO THE PROBLEM RELATION TO THE ASSIGNED TOPICS**

* + **TECHNICAL FEASIBILTY**

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. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

* + **OPERATION FEASIBILITY**

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

* + **ECONOMICAL FEASIBILITY**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased

1. **SUMMARY OF PROJECTS:**

Io use this plagiarism checker, please copy and paste your text in the box below, then click on the “Check Plagiarism” button, sit back and watch as your article undergoes a free plagiarism scame

**ARCHITECTURE:**

