BUILDING SEARCH ENGINE USING MACHINE LEARNING TECHNIQUE

A Project report submitted in partial

fulfilment for the award of the degree of

MASTER OF COMPUTER APPLICATIONS

(2022-2024)

Submitted by

Y. VARSHITHA

(22G21F0117)

Under the esteemed guidance of

Mr. A. HEMANTHA KUMAR

Associate Professor

Department of Master of Computer Applications

AUDISANKARA COLLEGE OF ENGINEERING& TECHNOLOGY

(AUTONOMOUS)

(Accredited by NAAC A+)

Approved by AICTE, Affiliated to JNTUA,

Ananthapuramu, Tirupati (DT), Andhra Pradesh,

NH-5, Bypass Road, Gudur, Tirupati (Dt.)

WWW.audisankara.ac.in

2022-2024

AUDISANKARA COLLEGE OF ENGINEERING & TECHNOLOGY

(AUTONOMOUS)

(Accredited by NAAC A+)

Approved by AICTE, Affiliated to JNTUA,

Ananthapuramu, Tirupati (DT), Andhra Pradesh

CERTIFICATE

This is to certify that the project report entitled" **BUILDING SEARCH ENGINE USING MACHINE LEARNING TECHNIQUE"** is the bonafide work done by me **Y. VARSHITHA, REGD.NO-22G21F0117** in partial fulfillment of the requirements for the award of the degree of **Master of computer Applications**, from Jawaharlal Nehru Technological University Anantapur, Ananthapuramu, during the year 2022-2024.

Project Guide

Prof.A. HEMANTHA KUMAR

Associate professor

Department of master of computer Applications
AUDISANKARA COLLEGE OF ENGG. & TECH.
GUDUR- TIRUPATI DISTRICT

Head of the Department

Prof. V. CHANDRASEKHAR

Associate professor

Department of master of computer Applications
AUDISANKARA COLLEGE OF ENGG. & TECH.
GUDUR- TIRUPATI DISTRICT

Submitted for the viva- voice examination held on________________________

Internal Examiner **External Examiner**

AUDISANKARA COLLEGE OF ENGINEERING & TECHNOLOGY

(AUTONOMOUS)

(Accredited by NAAC A+)

Approved by AICTE, Affiliated to JNTUA
Ananthapuramu, Tirupati (DT), Andhra Pradesh

DECLARATION

I, **Ms. Y. VARSHITHA, Regd.No-22G21F0117,** here by declare that the project work entitled **"BUILDING SEARCH ENGINE USING MACHINE LEARNING TECHNIQUE"** done by us under the esteemed guidance of Assistant Professor **Mr. A. HEMANTHA KUMAR,** and is submitted in partial fulfillment of the requirements for the award of the **Master's degree in computer applications.**

Date:

Place:

Signature of the candidate

Y. VARSHITHA
(22G21F0117)

ACKNOWLEDGEMENT

The satisfaction and elation that accompany the successful completion of any task would be incomplete without the mention of the people who have made it a possibility. It ismy great privilege to express my gratitude and respect to all those who have guided me andinspired me during the course of this project work.

First and for most, I express my sincere gratitude to our honorable chairman **Dr. VANKI PENCHALAIAH, M.A., M.L., Ph.D**. who provided all facilities and necessary encouragement during the course of study.

I extend my gratitude and sincere thanks to our beloved Director **Dr. A. MOHANBABU** principal **Dr. J. RAJA MURUGADOSS** for motivating and providing necessary infrastructure and permitting me to complete the project.

I would like to express the sense of gratitude towards our Head of the Department **Prof. V. CHANDRASEKHAR**, my internal guide **Prof. A. HEMANTHA KUMAR,** Associate Professor and for their support and encouragement for completion of this project.

Finally, I express my sincere thanks to all the teaching and non-teaching staff that guided and helped me to complete the project work successfully.

ABSTRACT

Building Search Engine using Machine Learning Technique The web is the huge and most extravagant wellspring of data. To recover the information from the World Wide Web, Search Engines are commonly utilized. Search engines provide a simple interface for searching for user query and displaying results in the form of the web address of the relevant web page but using traditional search engines has become very challenging to obtain suitable information. This project proposed a search engine using Machine Learning technique that will give more relevant web pages at top for user queries.

INDEX

S.NO	CONTENTS	PAGE NO
	ABSTRACT	
1.	INTRODUCTION	1
2.	SYSTEM STUDY	2
2.1	EXISTING SYSTEM	2
2.2	PROPOSED SYSTEM	2
3.	SYSTEM REQUIREMENTS	3
3.1	SOFTWARE REQUIREMENTS	3
3.2	HARDWARE REQUIRMENTS	3
3.3	FUNCTIONAL REQUIREMENTS	4
3.4	NON-FUNCTIONAL REQUIREMENTS	4
4.	LITERATURE SERVEY	6
5.	SYSTEM DESIGN	7
5.1	UML DIAGRAMS	7
6.	SOFTWARE ENVIRONMENT	11
7.	SYSTEM TESTING	30
7.1	TESTING STRATEGIES	30
7.2	TEST CASES	31
8.	SCREENSHOTS	33
9.	CONCLUSION	53
10.	REFERENCES	56

1. INTRODUCTION

World Wide Web is actually a web of individual systems and servers which are connected with different technology and methods. Every site comprises the heaps of site pages that are being made and sent on the server. So, if a user needs something, then he or she needs to type a keyword. Keyword is a set of words extracted from user search input. Search input given by a user may be syntactically incorrect. Here comes the actual need for search engines. Search engines provide you a simple interface to search user queries and display the results. Web crawlers help in collecting data about a website and the links related to them. We are only using web crawlers for collecting data and information from WWW and storing it in our database. Indexer which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository. Query Engine is mainly used to reply to the user's keyword and show the effective outcome for their keyword. In the query engine, the Page ranking algorithm ranks the URL by using different algorithms in the query engine. This project utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of the Page Rank algorithm is given as input to the machine learning algorithm.

2. SYSTEM STUDY

2.1 EXISTING SYSTEM

- Information retrieval is to retrieve the information resources that we are interested in or extract whatever information we need.
- Information Retrieval (IR) may deal with the organization, storage, retrieval and evaluation of information from documents, particularly textual information. But we cannot give the ranks to those documents.

DISADVANTAGES OF EXISTING SYSTEM

- Information retrieval will be very difficult in large numbers of texts in a document.
- Difficult to identify the important concepts or topic in a collection of documents.
- The explicit rankings are always difficult to obtain or even not available in many documents.

2.2 PROPOSED SYSTEM

- In this project author is using machine learning algorithms called SVM and XGBOOST to predict search result of given query and building search engine with machine learning algorithms.
- To train this algorithm author is using website data and then this data will be converted to numeric vector called TFIDF (term frequency inverse document frequency). TFIDF vector contains average frequency of each word.
- The proposed search engine is very useful for finding out more relevant URLs for given keywords.

ADVANTAGES OF PROPOSED SYSTEM

- We will build a search engine which gives the web address of the most relevant web page at the top of the search result, according to user queries.
- The main focus of our system is to build a search engine to discover the utmost suitable web address for the given keyword by using machine learning techniques for increasing accuracy compared to available search engines.

3. SYSTEM REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation. The appropriation of requirements and implementation constraints gives the general overview of the project in regards to what the areas of strength and deficit are and how to tackle them.

Python idle 3.7 version (or)
Anaconda 3.7 (or)
Jupiter (or)
Google Collab

3.2 HARDWARE REQUIREMENTS

Minimum hardware requirements are very dependent on the particular software being developed by a given Enthought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

Operating system	**:**	**Windows, Linux**
Processor	**:**	**Minimum intel i3**
Ram	**:**	**Minimum 4GB**
Hard disk	**:**	**Minimum 250GB**

3.3 FUNCTIONAL REQUIREMENTS

- Data Collection
- Data Preprocessing
- Training And Testing
- Prediction

3.4 NON-FUNCTIONAL REQUIREMENT

Non- functional Requirements allows you to impose constraints or restrictions on the design of the system across the various agile backlogs. Example, the site should load in 3 seconds when the number of simultaneous users is > 10000. Description of non-functional requirements is just as critical as a functional requirement.

- Usability requirement
- Serviceability requirement
- Manageability requirement
- Recoverability requirement
- Security requirement
- Data Integrity requirement
- Capacity requirement
- Availability requirement
- Scalability requirement
- Interoperability requirement
- Reliability requirement
- Maintainability requirement
- Regulatory requirement
- Environmental requirement

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

- Social Feasibility
- Economical Feasibility
- Technical Feasibility

SOCIAL FEASIBILITY

The aspect of project 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 use 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.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. 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.

4. LITERATURE SURVEY

1. S. Su, Y. Sun, X. Gao, J. Qiu* and Z. Tian*. A Correlation-change based Feature Selection Method for IoT Equipment Anomaly Detection. Applied Sciences.

In the era of the fourth industrial revolution, there is a growing trend to deploy sensors on industrial equipment, and analyses the industrial equipment's running status according to the sensor data. Thanks to the rapid development of IoT technologies [1], sensor data could be easily fetched from industrial equipment, and analyses to produce further value for industrial control at the edge of the network or at data centers. Due to the considerable development of deep learning in recent years, a common practice of such analysis is to conduct deep learning [2,3,4]. Such methods select a subset of all fetched sensor data stream as the input features, and generate equipment predictions. As a result, the performance of the learning model was seriously impacted by the features selected, thus feature selection plays a critical role for such methods.

2. X. Yu, Z. Tian, J. Qiu, F. Jiang. A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for Smart Mobile Devices. wireless communication and Mobile Computing, https://doi.org/10.1155/2018/5823439

With the development of Internet and information technology, smart mobile devices appear in our daily lives, and the problem of information leakage on smart mobile devices will follow which has become more and more serious [1, 2]. All kinds of private or sensitive information, such as intellectual property and financial data, might be distributed to unauthorized entity intentionally or accidentally. And that it is impossible to prevent from spreading once the confidential information has leaked.

3. Y. Sun, M. Li, S. Su, Z. Tian, W. Shi, M. Han. Secure Data Sharing Framework via Hierarchical Greedy Embedding in Darknets. ACM/Springer Mobile Networks an

Geometric routing, which combines greedy embedding and greedy forwarding, is a promising approach for efficient data sharing in darknets. However, the security of data sharing using geometric routing in darknets is still an issue that has not been fully studied. In this paper, we propose a Secure Data Sharing framework (SEDS) for darknets via hierarchical greedy embedding. SEDS adopts a hierarchical topology and uses a set of secure nodes to protect the whole topology. To support geometric routing in the hierarchical topology, a two-level bit-string prefix embedding approach (Prefix-T) is first proposed, and then a greedy forwarding strategy and a data mapping approach are combined with Prefix-T for data sharing.

5. SYSTEM DESIGN

5.1 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group The goal is for UML to become a common language for0 creating models of object-oriented computer software. In its current form UML is comprised of two major components a Meta-model and a notation. In the future, some form of method or process may also be added to or associated with, UML the unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems. the UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

GOALS

The Primary goals in the design of the UML are as follows:

- Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
- Provide extendibility and specialization mechanisms to extend the core concepts.
- Be independent of particular programming languages and development process.
- Provide a formal basis for understanding the modeling language.
- Encourage the growth of OO tools market.
- Support higher level development concepts such as collaborations, frameworks, patterns and components.

USE CASE DIAGRAM

A use case diagram in the Unified Modelling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The purpose of a use case diagram is to show what system functions are performed which actor. Roles of the actors in the system can be depicted.

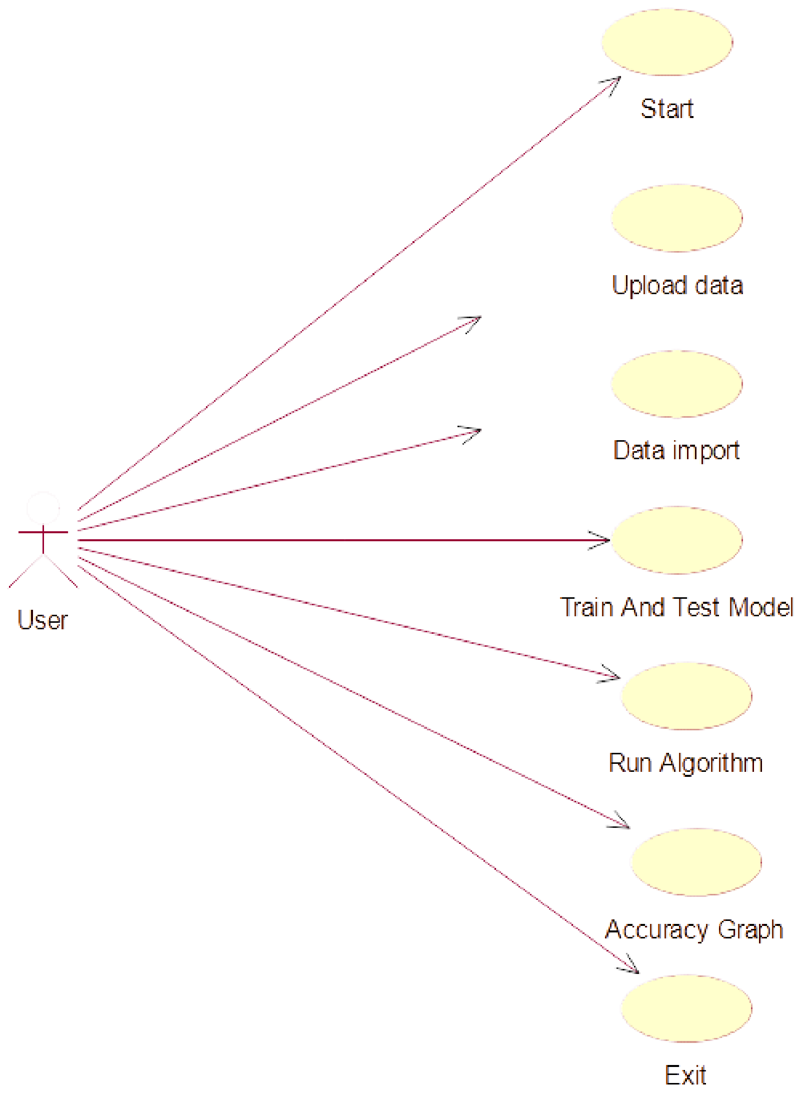


Fig: Use Case Diagram

CLASS DIAGRAM

The class diagram is used to refine the use case diagram and define a detailed design of the system. The class diagram classifies the actors defined in the use case diagram into a set of interrelated classes. The association between the classes can be either an "is-a" or "has-a" relationship.

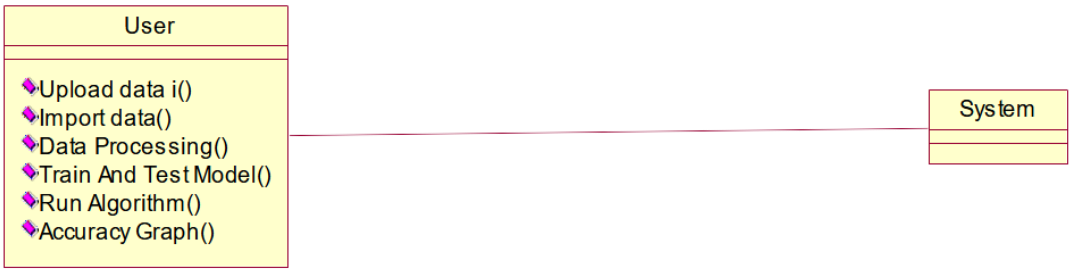


Fig: Class Diagram

SEQUENCE DIAGRAM

A sequence diagram represents the interaction between different objects in the system. The important aspect of a sequence diagram is that it is time-ordered.

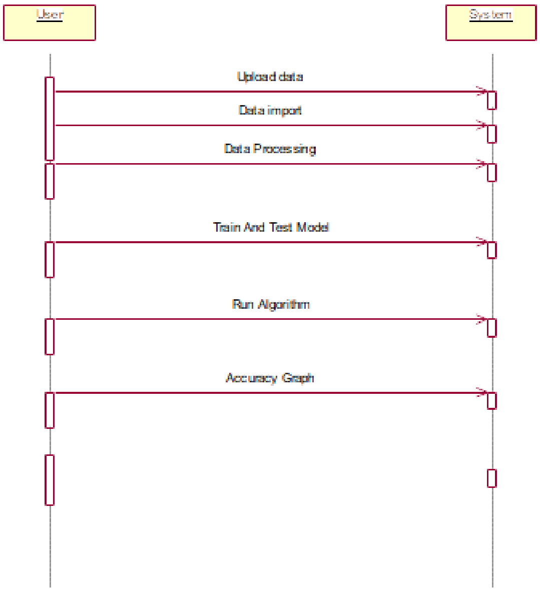


Fig: Sequence Diagram

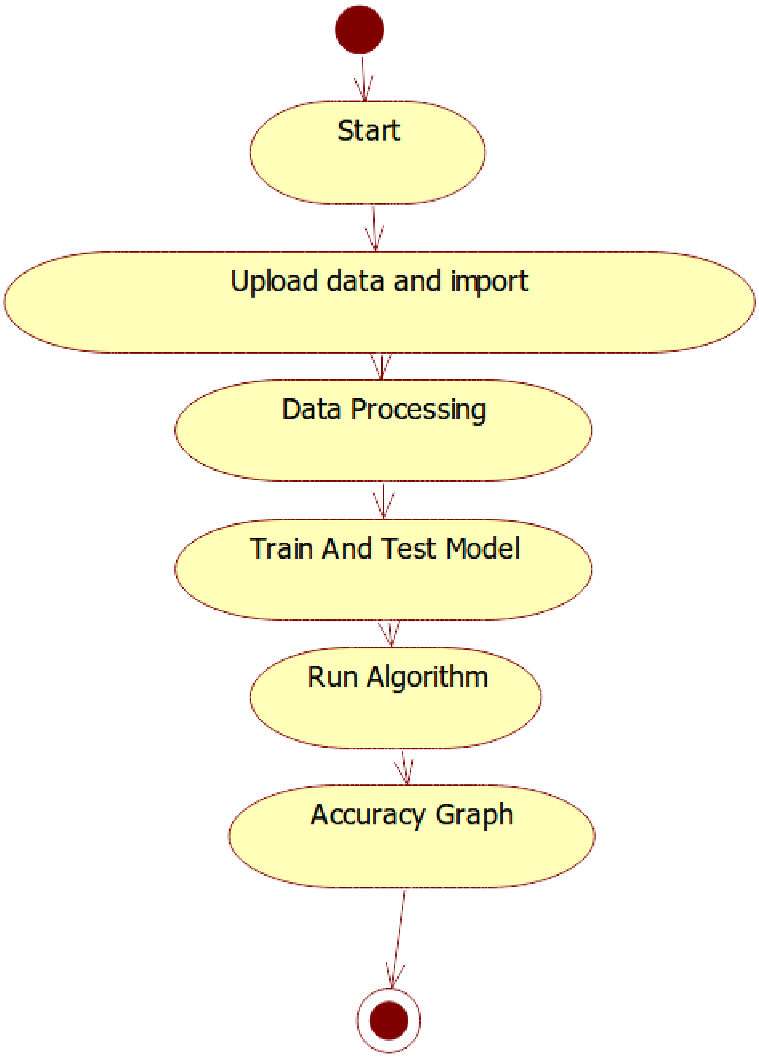


Fig: Sequence Diagram

6. SOFTWARE ENVIRONMENT

WHAT IS 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 What 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.
- GUI Applications (like KIVY, TKINTER, PYQT Etc.)
- Web frameworks like Django (used by YouTube, Instagram, Dropbox)
- Image processing (like Open CV, Pillow)
- Web scraping (like Scrapy, Beautiful Soup, Selenium)
- Test frameworks.
- Multimedia.

ADVANTAGES OF PYTHON

1.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.

2. 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.

3. 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.

4. 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.

5. SIMPLE AND EASY

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.

6. 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.

7. 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.

8. FREE AND OPEN-SOURCE

Like we said earlier, Python is freely available**.** But not only can you download Python for free, but you can also download its source code, make changes to it, even distribute it. It downloads with an extensive collection of libraries to help you with your tasks. 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.

9. PORTABLE

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.

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.

DISADVANTAGES OF PYTHON

1.SPEED LIMITATIONS

We have seen that Python code is executed line by line. But since Python 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 Carbon Nelle. The reason it is not so famous despite the existence of Bryton is that it isn’t that secure.

3. DESIGN RESTRICTIONS

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 i 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 Data Base Connectivity) and ODBC (Open Data Base Connectivity)**,** Python’s database access layers are a bit underdeveloped. Consequently, it is less often applied in huge enterprises.

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 Wickenden &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 Venners[1], Guido van Rossum said: "In the early 1980s, I worked as an implementer on a team building a language called ABC at Centrum over Wickenden an 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.

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.

CATEGORIES OF MACHINE 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."

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 programming 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. 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."

CHALLENGES IN MACHINES LEARNING

1.TIME-CONSUMING TASK

Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.

2. LACK OF SPECIALIST PERSONS

As ML technology is still in its infancy stage, availability of expert resources is a tough job.

3. 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.

4. ISSUE OF OVERFITTING & UNDERFITTING

If the model is overfitting or underfitting, it cannot be represented well for the problem.

5. CURSE OF DIMENSIONALITY

Another challenge ML model faces are too many features of data points.

6. 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,

- 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?

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, 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 PREREQUISITE

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 math’s 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.

(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.

(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! 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 Kera's, TensorFlow, Scikit-learn, 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 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,

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 colour, 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).

TYPES OF MACHINE LEARNING

SUPERVISED LEARNING – This involves learning from a training dataset with labelled data using classification and regression models. This learning process continues until the required level of performance is achieved.

UNSUPERVISED LEARNING – This involves using unlabeled 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 unlabeled data like Unsupervised Learning with a small amount of labelled data. Using labelled 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 behaviours 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 software’s; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3. CONTINUOUS IMPROVEMENT

As ML 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.

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-SUSCEPTIBLE

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 lists, dirt, str and others. 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.

- 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" behavior.
- Text Vs. Data Instead of Unicode Vs. 8-bit.

PYTHON

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 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 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. 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 behaviours. 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. 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.

MODULES USED IN PROJECT

TENSOR FLOW

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 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.

- 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. 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 I Python shells, the Jupiter 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 and thumbnail gallery.

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 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 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 behaviours. 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.

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.

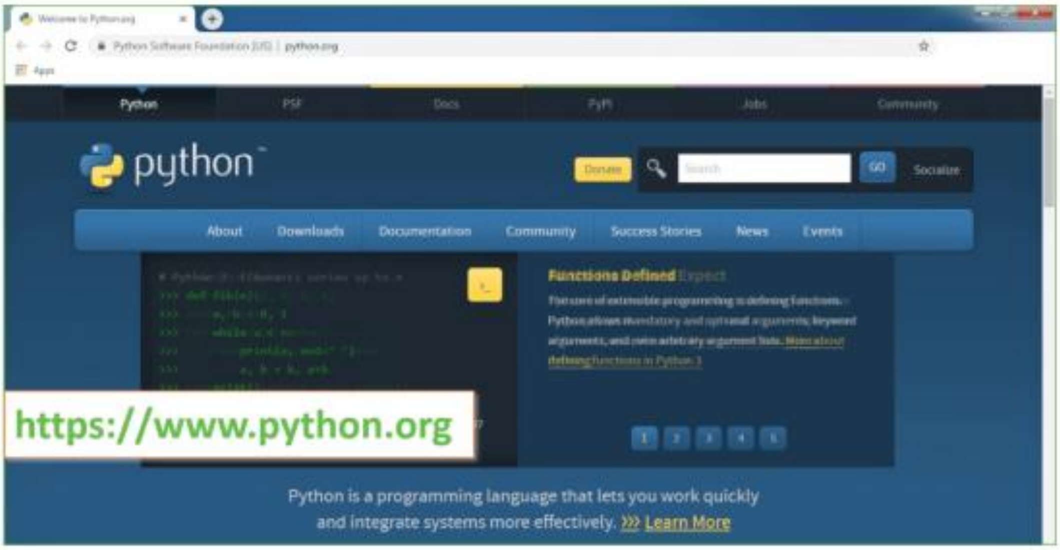
HOW TO INSTALL PYTHON ON WINDOWS AND MAC

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. 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 Cheat sheet here. The steps on how to install Python on Windows 10, 8 and 7 are divided into 4 parts to help understand better.

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

DOWNLOAD THE CORRECT VERSION INTO THE SYSTEM

STEP 1: Go to the official site to download and install python using Google Chrome or any other web browser. OR Click on the following link: **https://www.python.org**



https://www.python.org

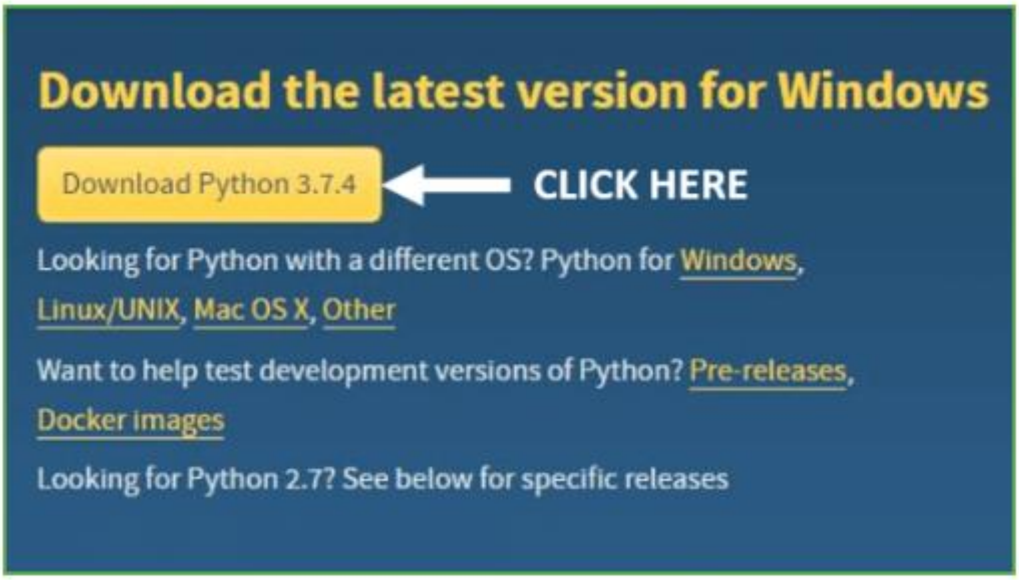
Functions Defined

The core of extensible programming is defining functions. Python allows mandatory and optional arguments, keyword arguments, and even arbitrary argument lists. More about defining functions in Python 3

Python is a programming language that lets you work quickly and integrate systems more effectively. >>> Learn More

Now, check for the latest and the correct version for your operating system.

STEP 2: Click on the Download Tab



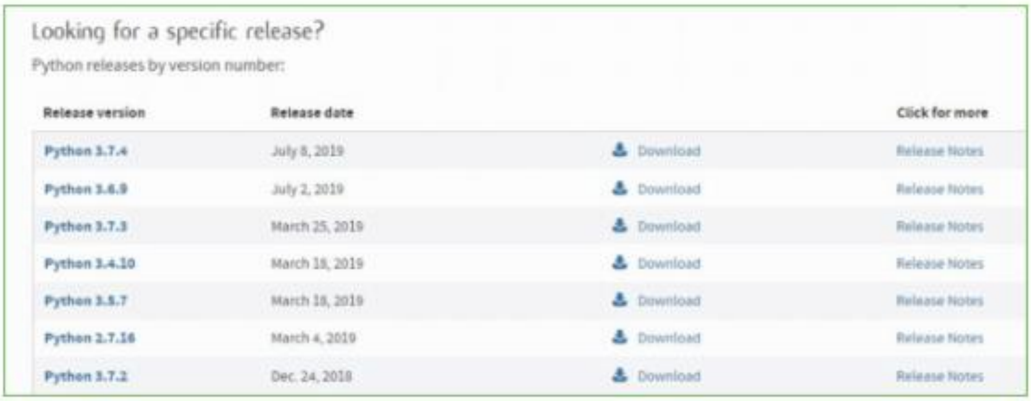
Download the latest version for Windows

Looking for Python with a different OS? Python for Windows, Linux/UNIX, Mac OS X, Other

Want to help test development versions of Python? Pre-releases, Docker images

Looking for Python 2.7? See below for specific releases

STEP 3: You can either select the Download Python for windows 3.7.4 button in Yellow Colour or you can scroll further down and click on download with respective to their version. Here, we are downloading the most recent python version for windows 3.7.4



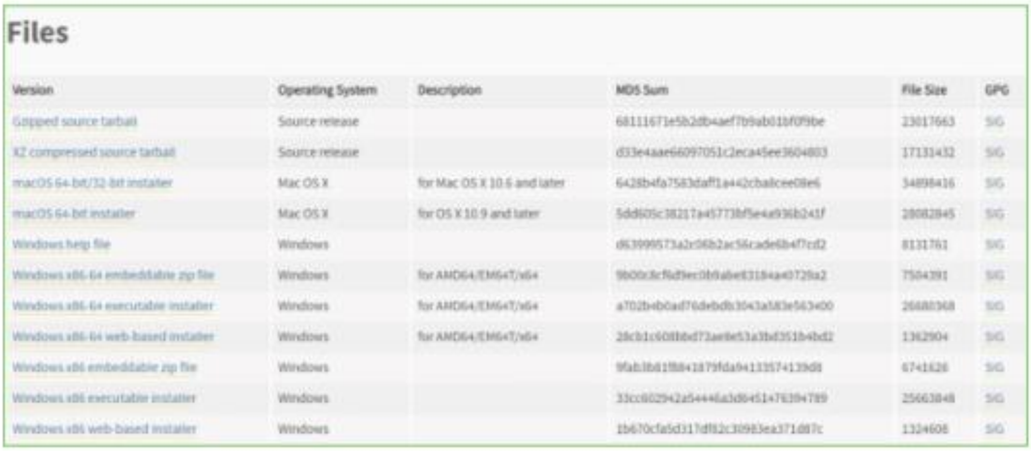
Looking for a specific release?

Python releases by version number:

Release version	Release date		Click for more
Python 3.7.4	July 8, 2019	Download	Release Notes
Python 3.6.9	July 2, 2019	Download	Release Notes
Python 3.7.3	March 25, 2019	Download	Release Notes
Python 3.4.10	March 18, 2019	Download	Release Notes
Python 3.5.7	March 18, 2019	Download	Release Notes
Python 2.7.16	March 4, 2019	Download	Release Notes
Python 3.7.2	Dec. 24, 2018	Download	Release Notes

STEP 4: Scroll down the page until you find the Files option.

STEP 5: Here you see a different version of python along with the operating **System**.



Files

Version	Operating System	Description	MD5 Sum	File Size	GPG
Gzipped source tarball	Source release		68111671e5b2db4aef7b9ab01bf0f9be	23017663	SIG
XZ compressed source tarball	Source release		d33e4aae66097051c2eca45ee3604803	17131432	SIG
macOS 64-bit/32-bit installer	Mac OS X	for Mac OS X 10.6 and later	6428b4fa7583daff1a442cba8cee08e6	34898416	SIG
macOS 64-bit installer	Mac OS X	for OS X 10.9 and later	5dd605c38217a45773bf5e4a936b241f	28082845	SIG
Windows help file	Windows		d63999573a2c06b2ac56cade6b4f7cd2	8131761	SIG
Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64	9b00c8cf6d9ec0b9abe83184a40729a2	7504391	SIG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64	a702b4b0ad76debdb3043a583e563400	26680368	SIG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64	28cb1c608bbd73ae8e53a3bd351b4bd2	1362904	SIG
Windows x86 embeddable zip file	Windows		9fab3b81f8841879fda94133574139d8	6741626	SIG
Windows x86 executable installer	Windows		33cc602942a54446a3d6451476394789	25663848	SIG
Windows x86 web-based installer	Windows		1b670cfa5d317df82c30983ea371d87c	1324608	SIG

To download Windows 32-bit python, you can select any one from the three options: Windows x86 embeddable zip file, Windows x86 executable installer or Windows x86 web-based installer.

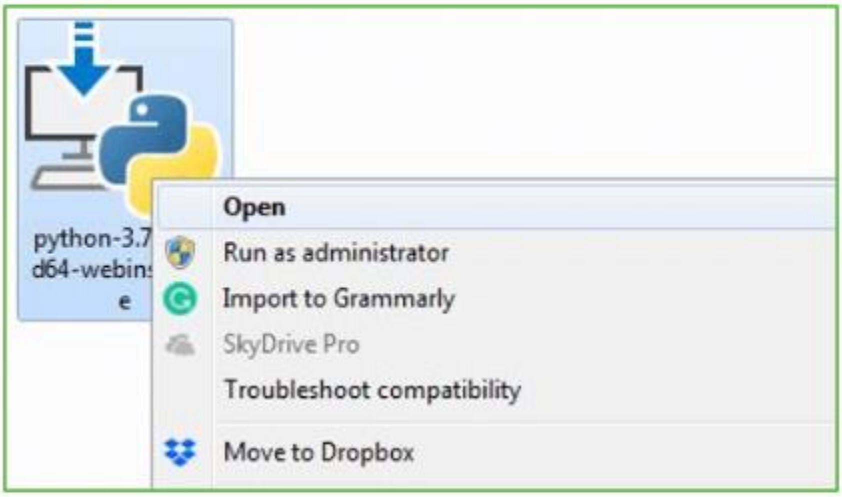
To download Windows 64-bit python, you can select any one from the three options: Windows x86-64 embeddable zip file, Windows x86-64 executable installer or Windows x86-64 web-based installer.

Here we will install Windows x86-64 web-based installer. Here your first part regarding which version of python is to be downloaded is completed. Now we move ahead with the second part in installing python i.e. Installation.

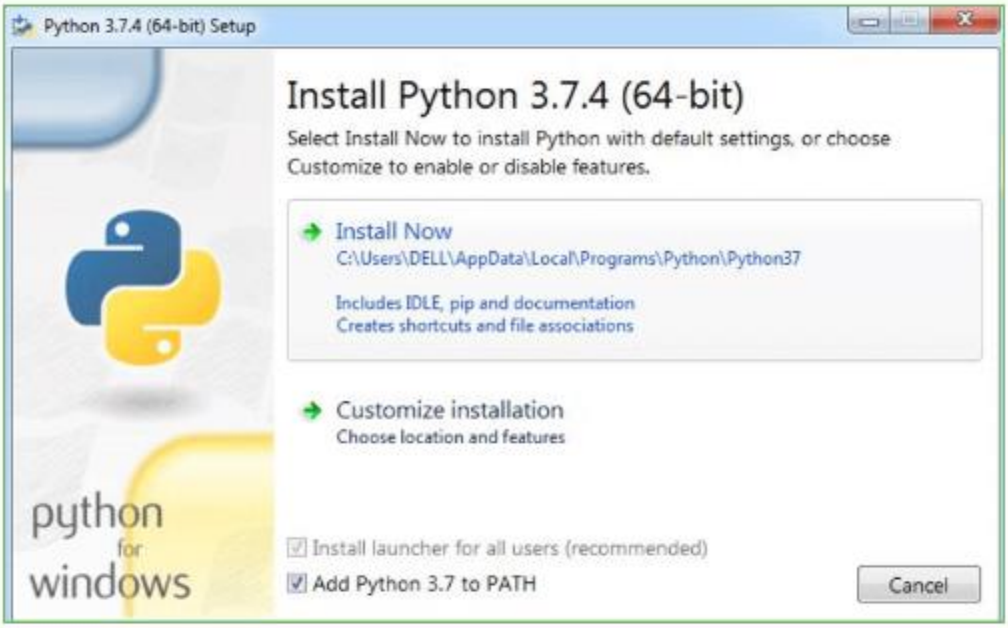
NOTE: To know the changes or updates that are made in the version you can click on the Release Note Option.

INSTALLATION OF PYTHON

STEP 1: Go to Download and Open the downloaded python version to carry out the installation process.



STEP 2: Before you click on Install Now, make sure to put a tick on Add Python 3.7 to PATH.



Python 3.7.4 (64-bit) Setup

Install Python 3.7.4 (64-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

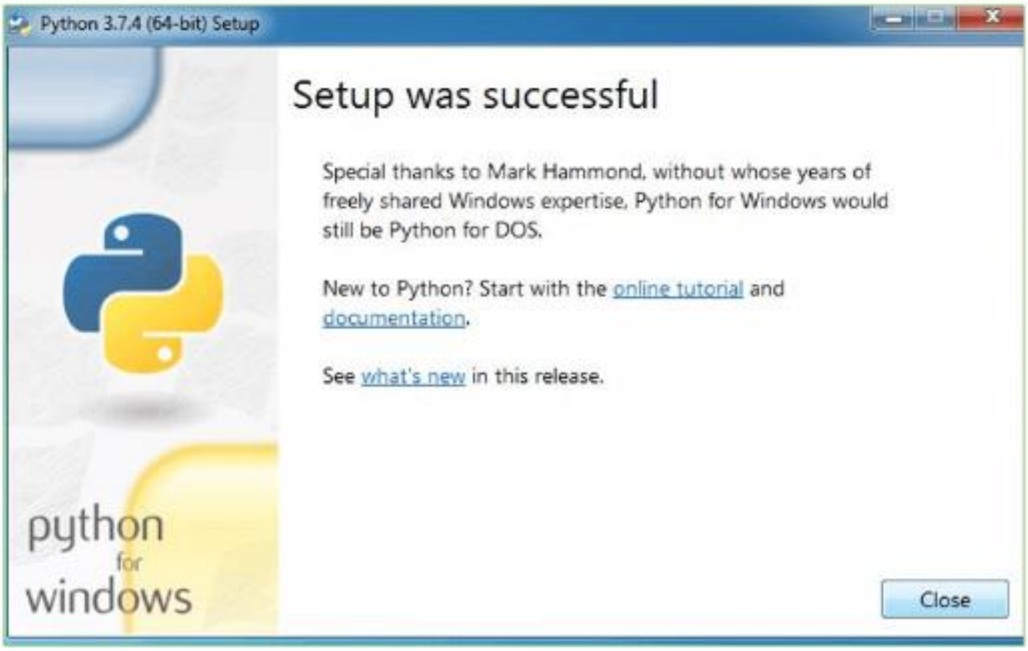
→ Install Now
C:\Users\DELL\AppData\Local\Programs\Python\Python37

Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

☑ Install launcher for all users (recommended)
☑ Add Python 3.7 to PATH

STEP 3: Click on Install NOW After the installation is successful. Click on Close.



Python 3.7.4 (64-bit) Setup

Setup was successful

Special thanks to Mark Hammond, without whose years of freely shared Windows expertise, Python for Windows would still be Python for DOS.

New to Python? Start with the online tutorial and documentation.

See what's new in this release.

Fig: With these above three steps on python installation, you have successfully and correctly installed Python. Now is the time to verify the installation.

NOTE: The installation process might take a couple of minutes.

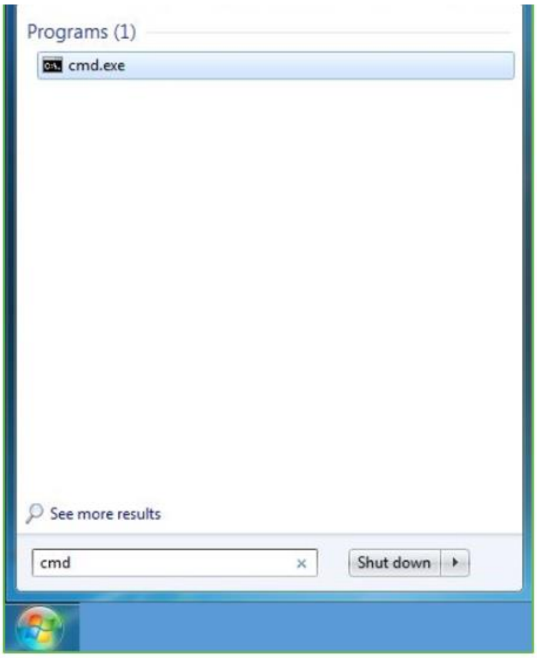
VERIFY THE PYTHON INSTALLATION

To download Windows 64-bit python, you can select any one from the three options: Windows x86-64 embeddable zip file, Windows x86-64 executable installer or Windows x86-64 web-based installer.

Here we will install Windows x86-64 web-based installer. Here your first part regarding which version of python is to be downloaded is completed.

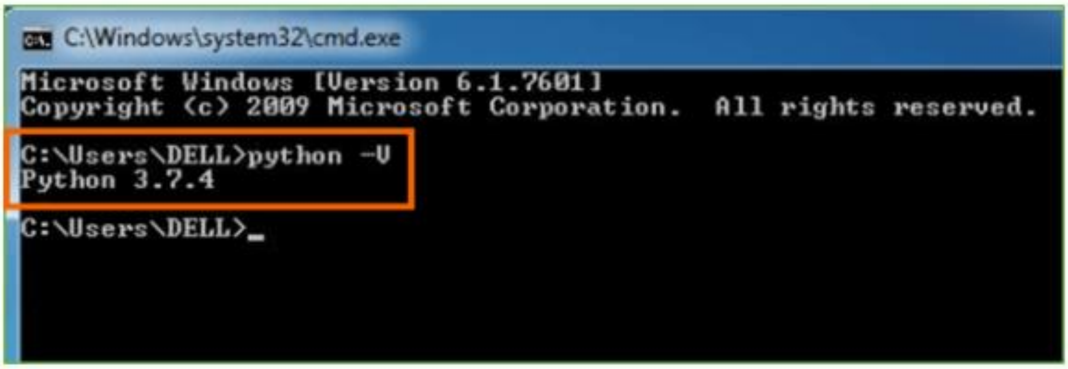
STEP 1: Click on Start

STEP 2: In the Windows Run Command, type “cmd”.



STEP 3: Open the Command prompt option.

STEP 4: Let us test whether the python is correctly installed. Type python **–V** and press Enter.



```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\DELL>python -V
Python 3.7.4

C:\Users\DELL>_
```

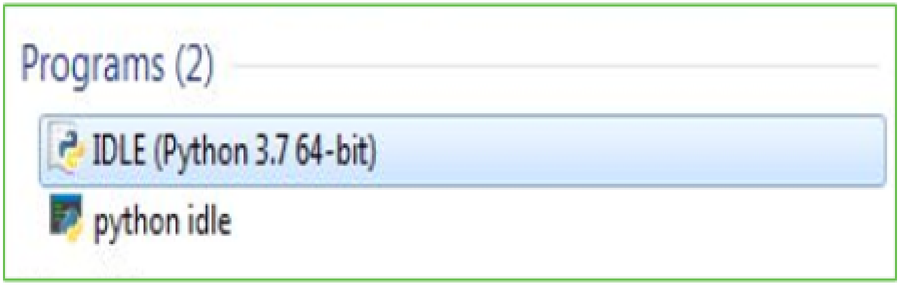
STEP 5: You will get the answer as 3.7.4

NOTE: If you have any of the earlier versions of Python already installed. You must first uninstall the earlier version and then install the new one.

CHECK HOW THE PYTHON IDLE WORKS

STEP 1: Click on Start

STEP 2: In the Windows Run command, type "python idle".



Programs (2)

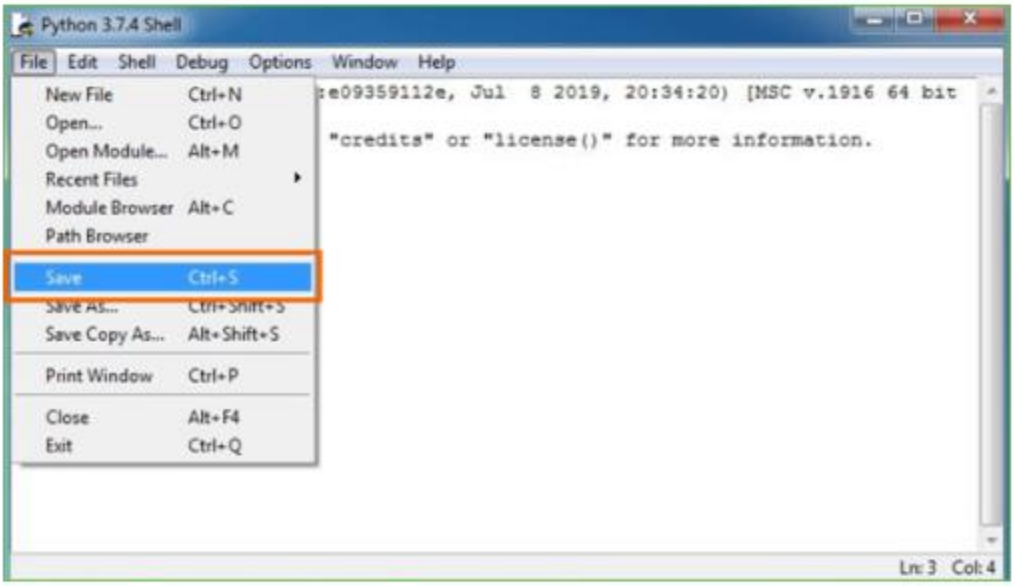
IDLE (Python 3.7 64-bit)

python idle

STEP 3: Click on IDLE (Python 3.7 64-bit) and launch the program

STEP 4: To go ahead with working in IDLE you must first save the file. **Click on**

FILE > CLICK ON SAVE



```
:e09359112e, Jul  8 2019, 20:34:20) [MSC v.1916 64 bit

 "credits" or "license()" for more information.
```

STEP 5: Name the file and save as type should be Python files. Click on SAVE. Here I have named the files as Hey World.

STEP 6: Now for e.g. **enter print.**

7. SYSTEM TESTING

7.1 TESTING STRATEGIES

7.1.1 UNIT TESTING

Unit testing, a testing technique using which individual modules are tested to determine if there are issues by the developer himself.it is concerned with functional correctness of the standalone modules. The main aim is to isolate each unit of the system to identify, analyses and fix the defects.

UNIT TESTING TECHNIQUES

BLACK BOX TESTING - Using which the user interface, input and output are tested.

WHITE BOX TESTING -Used to test each one of those functions' behaviors is tested.

7.1.2 DATA FLOW TESTING

Data flow testing is a family of testing strategies based on selecting paths through the program's control flow in order to explore sequence of events related to the status of Variables or data object. Dataflow Testing focuses on the points at which variables receive and the points at which these values are used.

7.1.3 INTEGRATION TESTING

Integration Testing done upon completion of unit testing, the units or modules are to be integrated which gives raise too integration testing. The purpose of integration testing is to verify the functional, performance, and reliability between the modules that are integrated.

7.1.4 BIG BANG INTEGRATION TESTING

Big Bang Integration Testing is an integration testing Strategy wherein all units are linked at once, resulting in a complete system. When this type of testing strategy is adopted, it is difficult to isolate any errors found, because attention is not paid to verifying the interfaces across individual units.

7.1.5 USER INTERFACE TESTING

User interface testing, a testing technique used to identify the presence of defects is a product/software under test by Graphical User interface [GUI].

7.2 TEST CASES

S.NO	INPUT	OUTPUT	RESULT
Test case 1(unit testing of dataset)	The user gives the input in the form of upload forge/real audio dataset details.	An output is to train GMM model and calculate prediction accuracy.	A result is based on GMM model generated and its prediction accuracy is 100%on test data.
Test case 2 (Unit testing of Accuracy)	The user gives the input in the form of upload forge/real audio dataset details.	An output digital audio Authentication on forensics for test data.	A result is based on GMM model generated and its prediction accuracy is 100%on test data.
Test case3 (Unit testing machine learning Algorithms)	The user gives the input in the form of upload forge/real audio dataset details.	An output digital audio Authentication on forensics for test data.	A result is based on GMM model generated and its prediction accuracy is 100%on test data.
Test case 4 (integration testing of dataset)	The user gives the input in the form of upload forge/real audio dataset details.	An output digital audio Authentication on forensics for test data.	A result is based on GMM model generated and its prediction accuracy is 100%on test data.
Testcase 5 (Big bang testing)	The user gives the input in the form of upload forge/real audio dataset details.	An output digital audio Authentication on forensics for test data.	A result is based on GMM model generated and its prediction accuracy is 100%on test data.
Test case 6 (data flow testing)	The user gives the input in the form of upload forge/real audio dataset details	An output digital audio Authentication on forensics for test data.	A result is based on GMM model generated and its prediction accuracy is 100%on test data.

S.NO	INPUT	OUTPUT	RESULT
Test case 7(user interface testing)	A result is based on GMM model generated and its prediction accuracy is 100%on test data.	An output digital audio Authentication on forensics for test data.	A result is based on GMM model generated and its prediction accuracy is 100%on test data.
Test case 8 (user interface testing-event based)	The user gives the input in the form of upload forge/real audio dataset details	An output digital audio Authentication on forensics for test data.	A result is based on GMM model generated and its prediction accuracy is 100%on test data.

8.SCREENSHOTS

BUILDING SEARCH ENGINE USING MACHINE LEARNING TECHNIQUE

In this project author is using machine learning algorithms called SVM and XGBOOST to predict search result of given query and building search engine with machine learning algorithms. To train this algorithm author is using website data and then this data will be converted to numeric vector called TFIDF (term frequency inverse document frequency). TFIDF vector contains average frequency of each word.

ADMIN MODULE: admin can login to application using username and password as admin and then accept or activate new users' registration and then train SVM and XGBOOST algorithm.

MANAGER MODULE: manager can login to application by using username and password as Manager and Manager and then upload dataset to application.

NEW USER SIGNUP: using this module new user can sign up with the application.

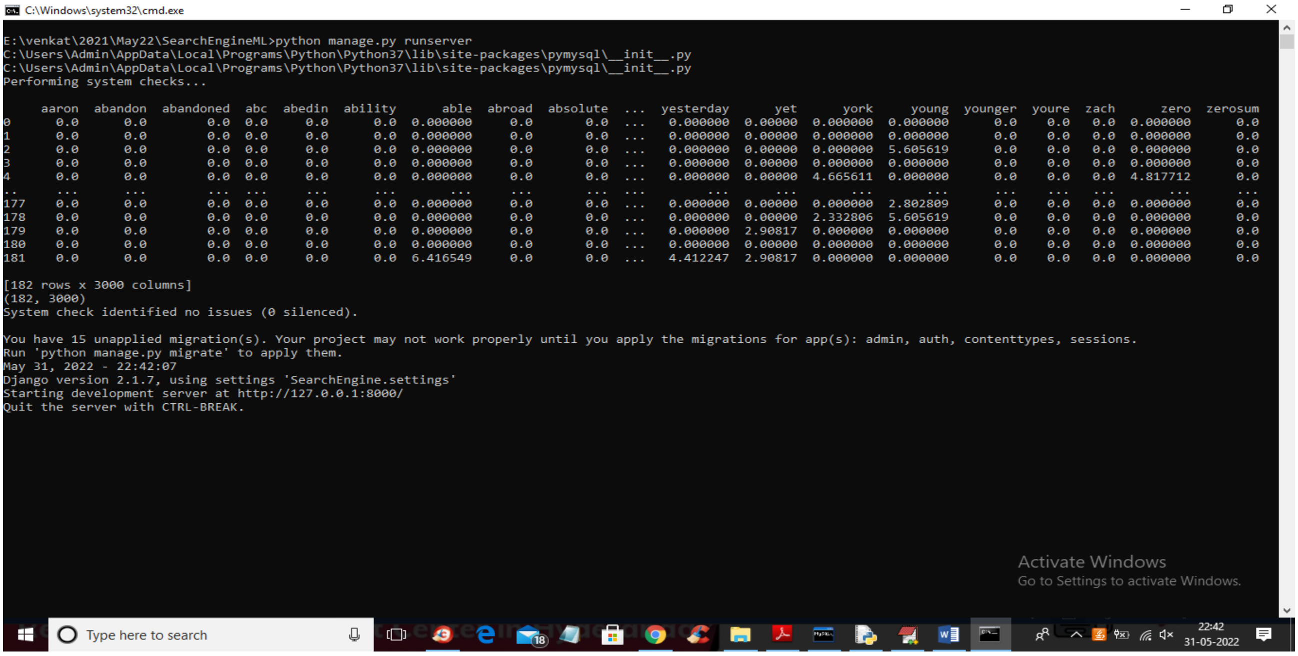
USER LOGIN: user can login to application and then perform search by giving query.

To run project, install MYSQL and python 3.7 and then copy content from DB.txt file and paste in MYSQL to create database.

Now double click on 'run.bat' file to start python DJANGO server and get below screen.

In below screen server started and build a vector from dataset where first row showing word and remaining rows contains TFIDF word frequency. Now open browser and enter URL as http://127.0.0.1:8000/index.html and press enter key to get below page.

BUILD A VECTOR FROM DATASET



```
E:\venkat\2021\May22\SearchEngineML>python manage.py runserver
C:\Users\Admin\AppData\Local\Programs\Python\Python37\lib\site-packages\pymysql\__init__.py
C:\Users\Admin\AppData\Local\Programs\Python\Python37\lib\site-packages\pymysql\__init__.py
Performing system checks...

     aaron  abandon  abandoned  abc  abedin  ability      able  abroad  absolute  ...  yesterday      yet      york     young  younger  youre  zach      zero  zerosum
0      0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  0.00000  0.000000  0.000000      0.0    0.0   0.0  0.000000      0.0
1      0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  0.00000  0.000000  0.000000      0.0    0.0   0.0  0.000000      0.0
2      0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  0.00000  0.000000  5.605619      0.0    0.0   0.0  0.000000      0.0
3      0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  0.00000  0.000000  0.000000      0.0    0.0   0.0  0.000000      0.0
4      0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  0.00000  4.665611  0.000000      0.0    0.0   0.0  4.817712      0.0
..     ...      ...        ...  ...     ...      ...       ...     ...       ...  ...        ...      ...       ...       ...      ...    ...   ...       ...      ...
177    0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  0.00000  0.000000  2.802809      0.0    0.0   0.0  0.000000      0.0
178    0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  0.00000  2.332806  5.605619      0.0    0.0   0.0  0.000000      0.0
179    0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  2.90817  0.000000  0.000000      0.0    0.0   0.0  0.000000      0.0
180    0.0      0.0        0.0  0.0     0.0      0.0  0.000000     0.0       0.0  ...   0.000000  0.00000  0.000000  0.000000      0.0    0.0   0.0  0.000000      0.0
181    0.0      0.0        0.0  0.0     0.0      0.0  6.416549     0.0       0.0  ...   4.412247  2.90817  0.000000  0.000000      0.0    0.0   0.0  0.000000      0.0

[182 rows x 3000 columns]
(182, 3000)
System check identified no issues (0 silenced).

You have 15 unapplied migration(s). Your project may not work properly until you apply the migrations for app(s): admin, auth, contenttypes, sessions.
Run 'python manage.py migrate' to apply them.
May 31, 2022 - 22:42:07
Django version 2.1.7, using settings 'SearchEngine.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

C:\Windows\system32\cmd.exe
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:42
31-05-2022

Fig: server started and build a vector from dataset

NEW USER SIGNUP HERE

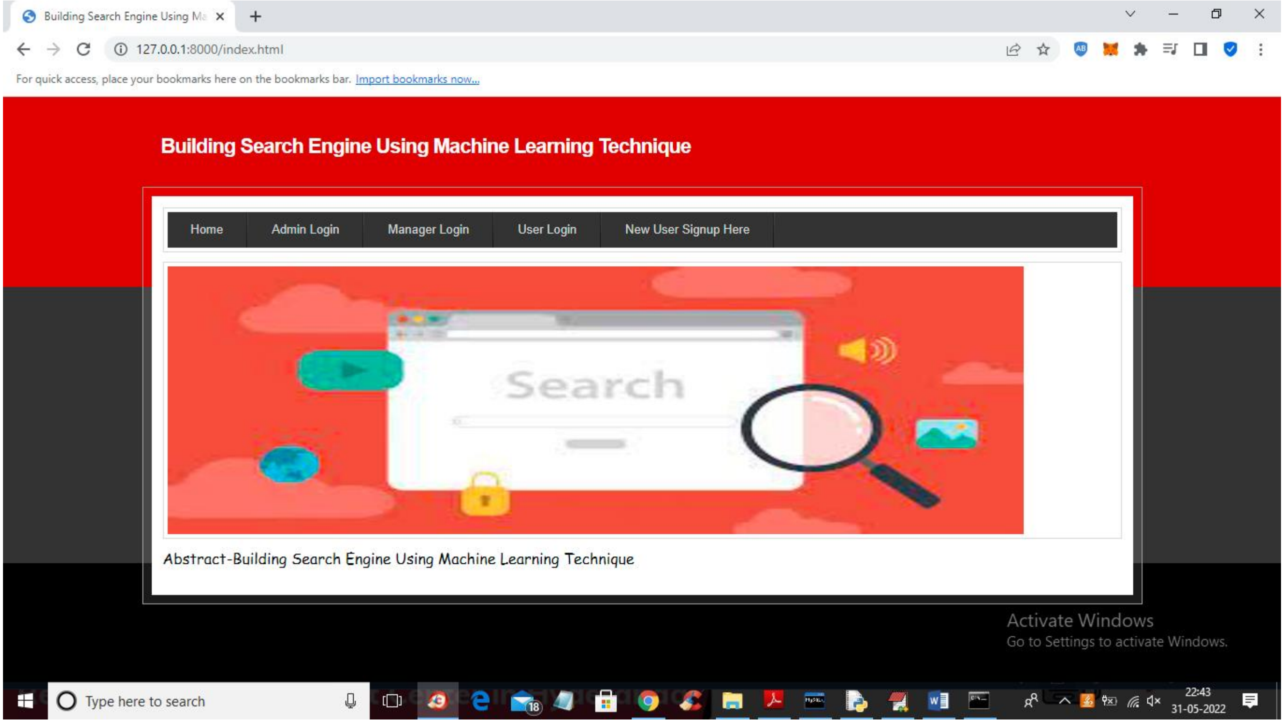
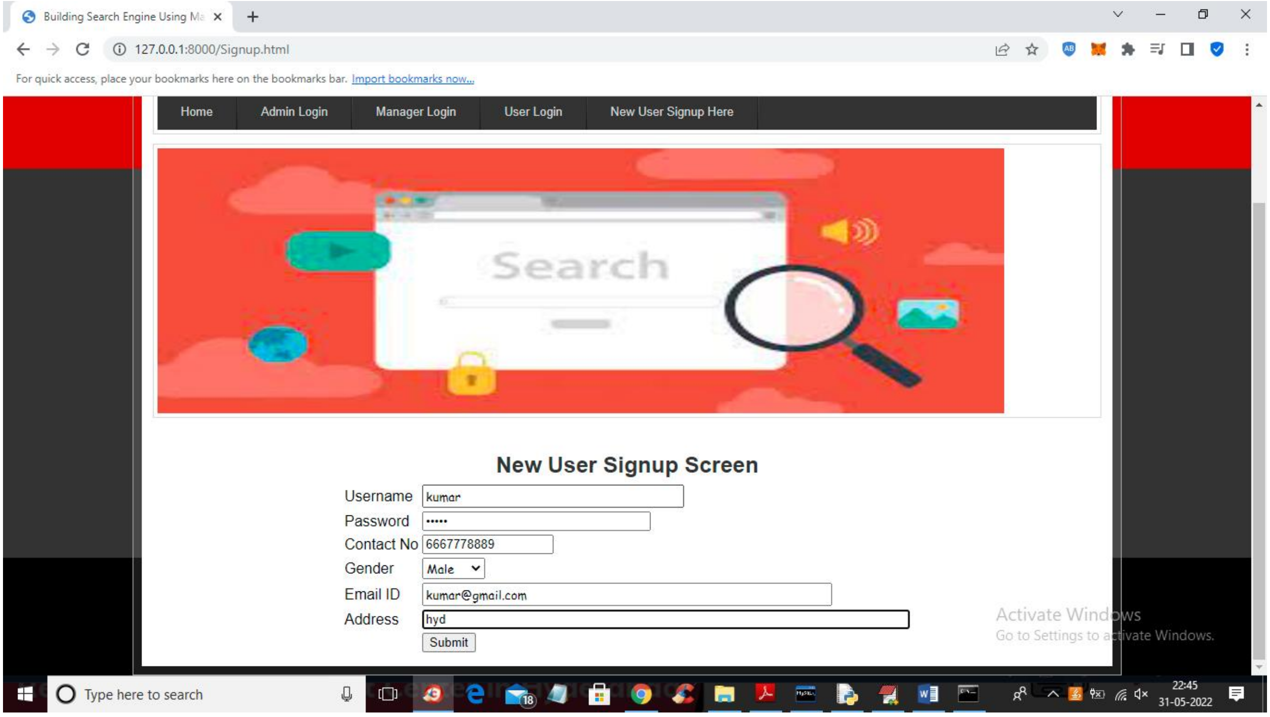

Building Search Engine Using Ma
127.0.0.1:8000/index.html
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
Home
Admin Login
Manager Login
User Login
New User Signup Here
Search
Abstract-Building Search Engine Using Machine Learning Technique
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:43
31-05-2022

Fig: New User Signup Here

USER IS SIGNING UP



Building Search Engine Using Ma
127.0.0.1:8000/Signup.html
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Home
Admin Login
Manager Login
User Login
New User Signup Here
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:45
31-05-2022

New User Signup Screen

Username kumar

Password •••••

Contact No 6667778889

Gender Male

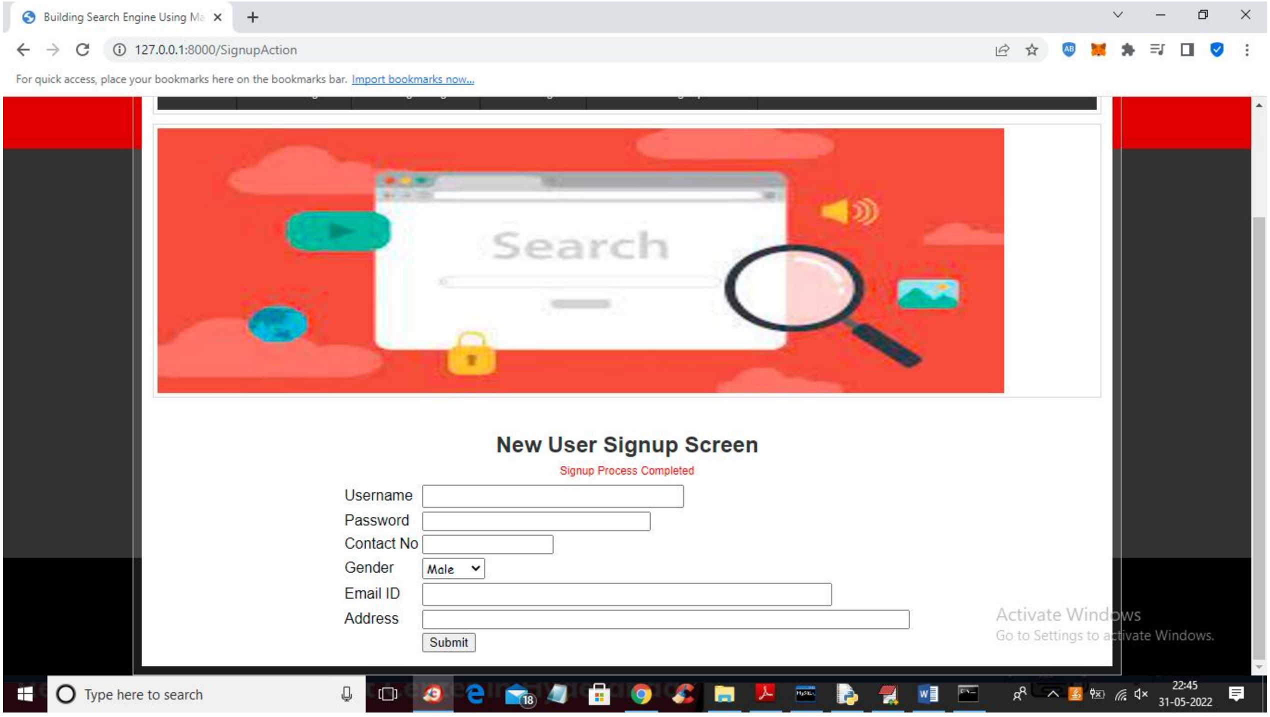
Email ID kumar@gmail.com

Address hyd

Submit

Fig: user is signing up

NEW USER SIGNUP SCREEN


Building Search Engine Using Ma
127.0.0.1:8000/SignupAction
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:45
31-05-2022

New User Signup Screen

Signup Process Completed

Username

Password

Contact No

Gender Male

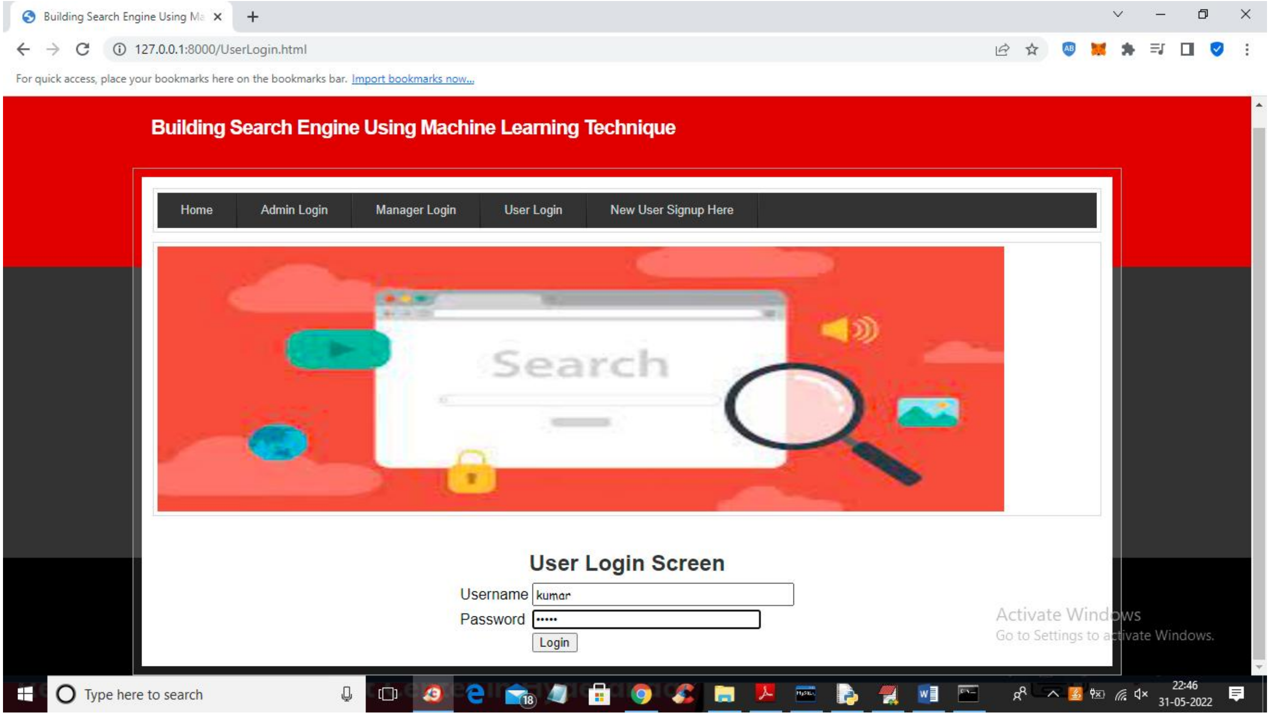
Email ID

Address

Submit

Fig: New user signup Screen

USER LOGIN SCREEN


Building Search Engine Using Ma
127.0.0.1:8000/UserLogin.html
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
Home
Admin Login
Manager Login
User Login
New User Signup Here
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:46
31-05-2022

User Login Screen

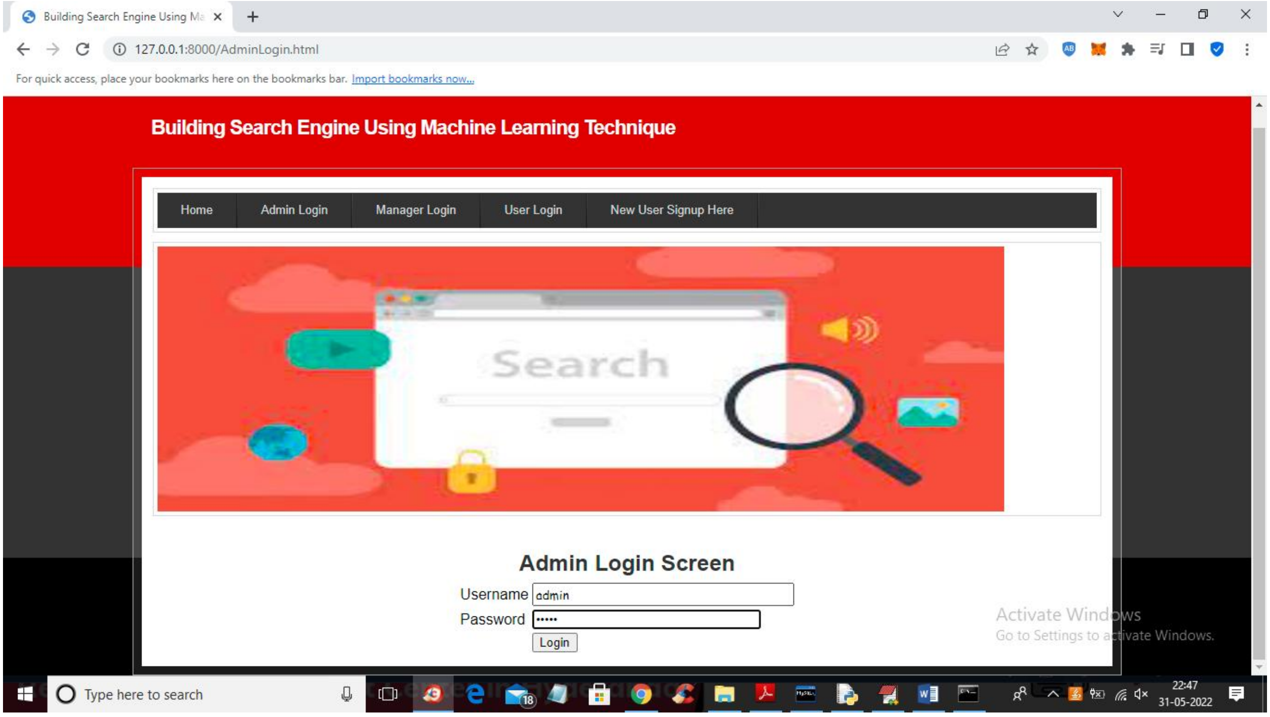
Username kumar

Password •••••

Login

Fig: User Login Screen

ADMIN LOGIN SCREEN



Building Search Engine Using Ma
127.0.0.1:8000/AdminLogin.html
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
Home
Admin Login
Manager Login
User Login
New User Signup Here
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:47
31-05-2022

Admin Login Screen

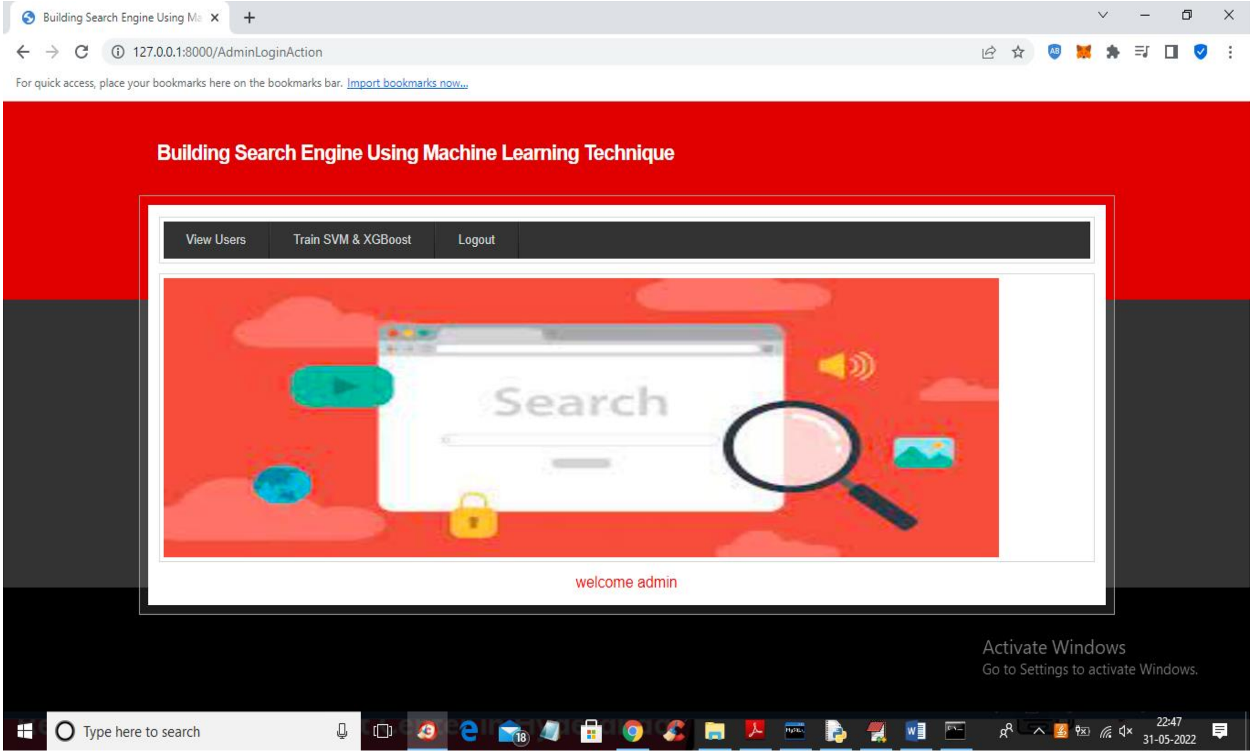
Username admin

Password •••••

Login

Fig: Admin Login Screen

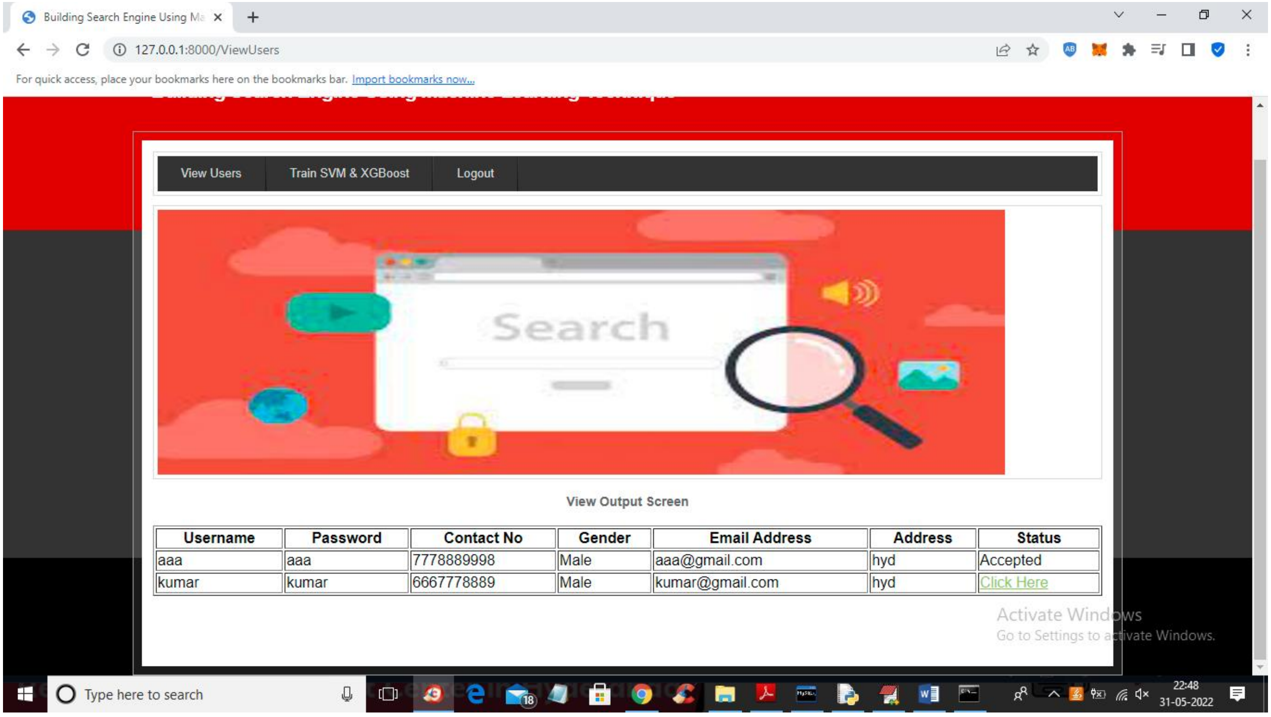
WELCOME ADMIN



Building Search Engine Using Ma
127.0.0.1:8000/AdminLoginAction
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
View Users
Train SVM & XGBoost
Logout
Search
welcome admin
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:47
31-05-2022

Fig: Welcome Admin

VIEW OUTPUT SCREEN

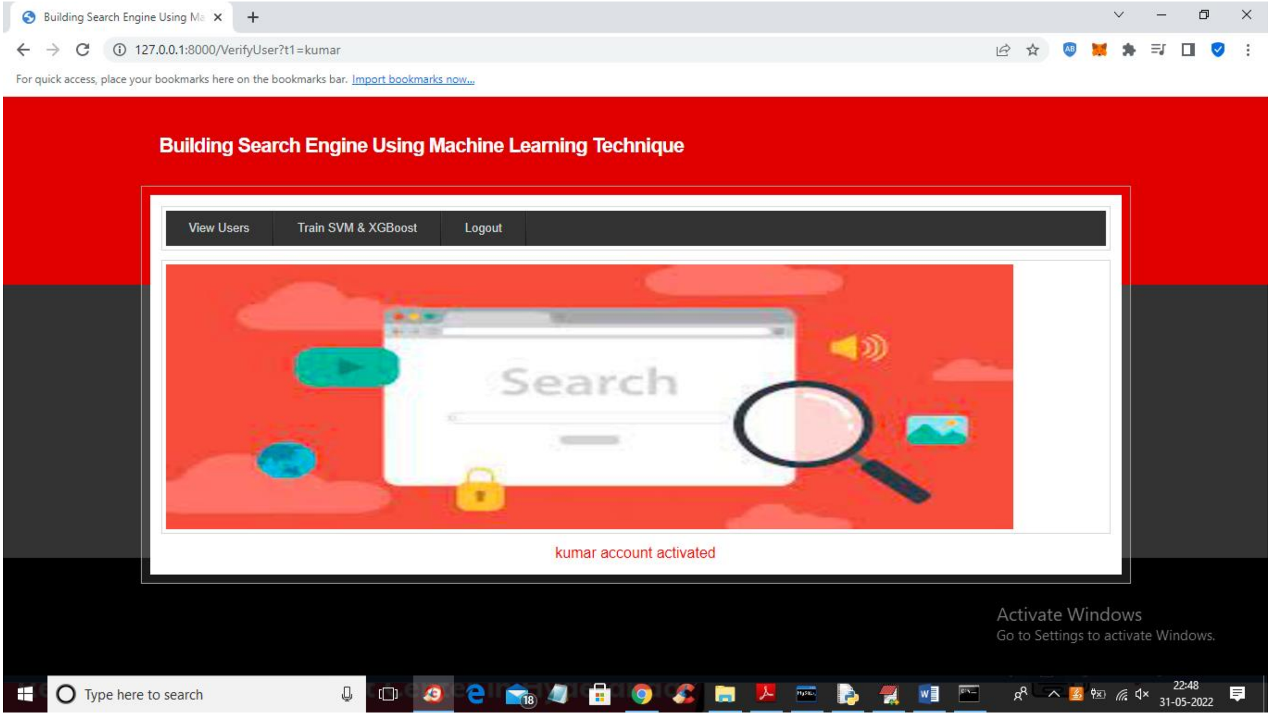

Building Search Engine Using Ma
127.0.0.1:8000/ViewUsers
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
View Users
Train SVM & XGBoost
Logout
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:48
31-05-2022

View Output Screen

Username	Password	Contact No	Gender	Email Address	Address	Status
aaa	aaa	7778889998	Male	aaa@gmail.com	hyd	Accepted
kumar	kumar	6667778889	Male	kumar@gmail.com	hyd	Click Here

Fig: View Output Screen

ADMIN ACTIVATED USER ACCOUNT

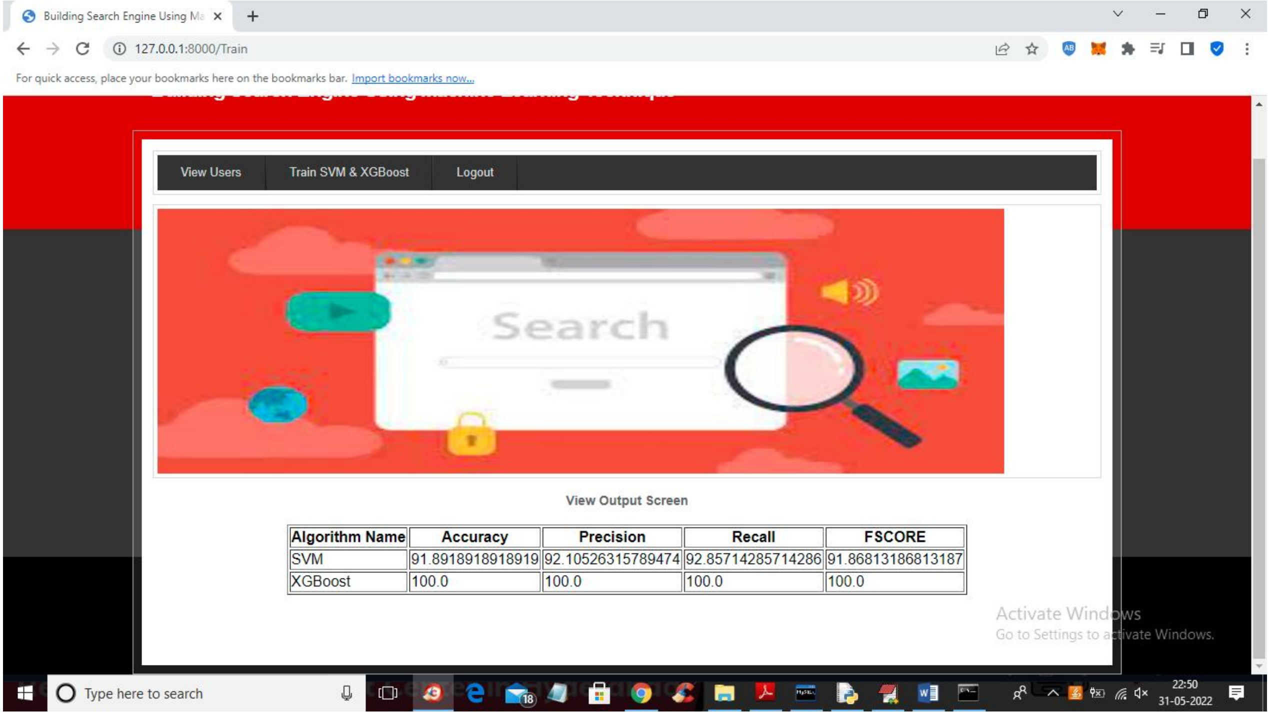


Building Search Engine Using Ma
127.0.0.1:8000/VerifyUser?t1=kumar
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
View Users
Train SVM & XGBoost
Logout
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:48
31-05-2022

kumar account activated

Fig: admin activated user account

VIEW OUTPUT SCREEN

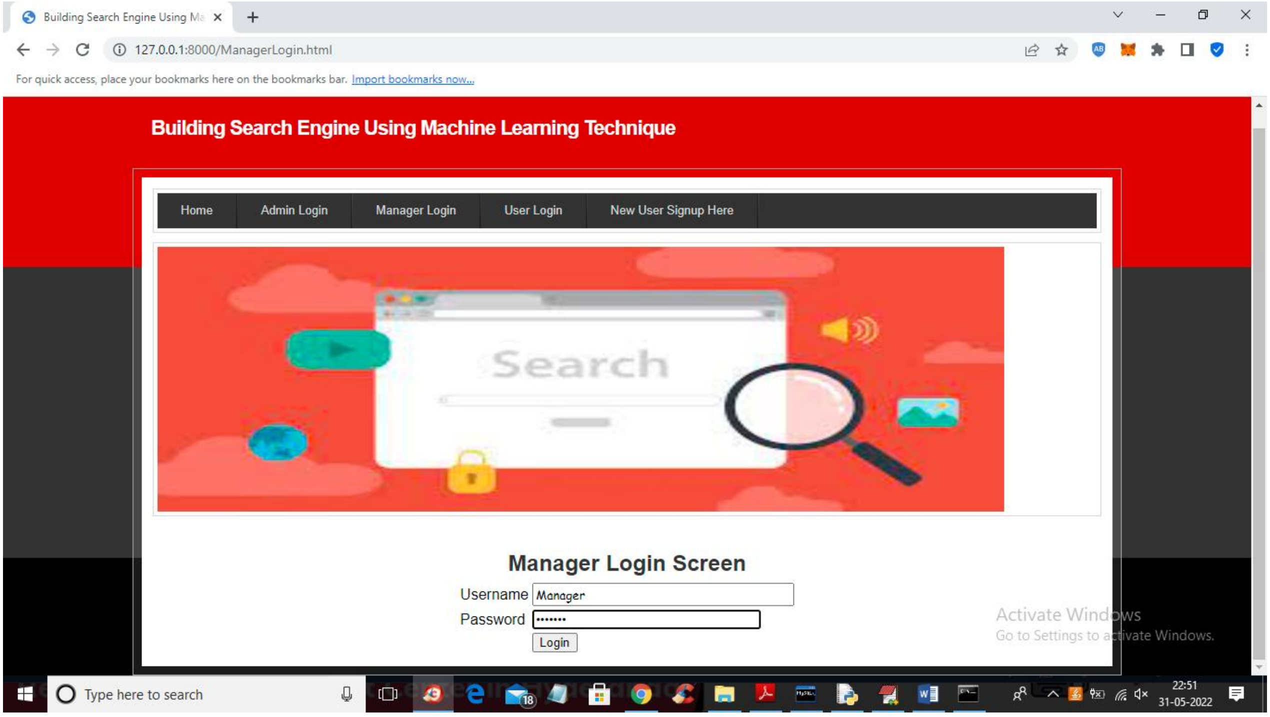

Building Search Engine Using Ma
127.0.0.1:8000/Train
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
View Users
Train SVM & XGBoost
Logout
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:50
31-05-2022

View Output Screen

Algorithm Name	Accuracy	Precision	Recall	FSCORE
SVM	91.8918918918919	92.10526315789474	92.85714285714286	91.86813186813187
XGBoost	100.0	100.0	100.0	100.0

Fig: view Output Screen

MANAGER LOGIN SCREEN



127.0.0.1:8000/ManagerLogin.html

Building Search Engine Using Machine Learning Technique

Home
Admin Login
Manager Login
User Login
New User Signup Here
Search

Manager Login Screen

Username Manager

Password •••••••

Login
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:51
31-05-2022

Fig: Manager Login Screen

MANAGER LOGIN

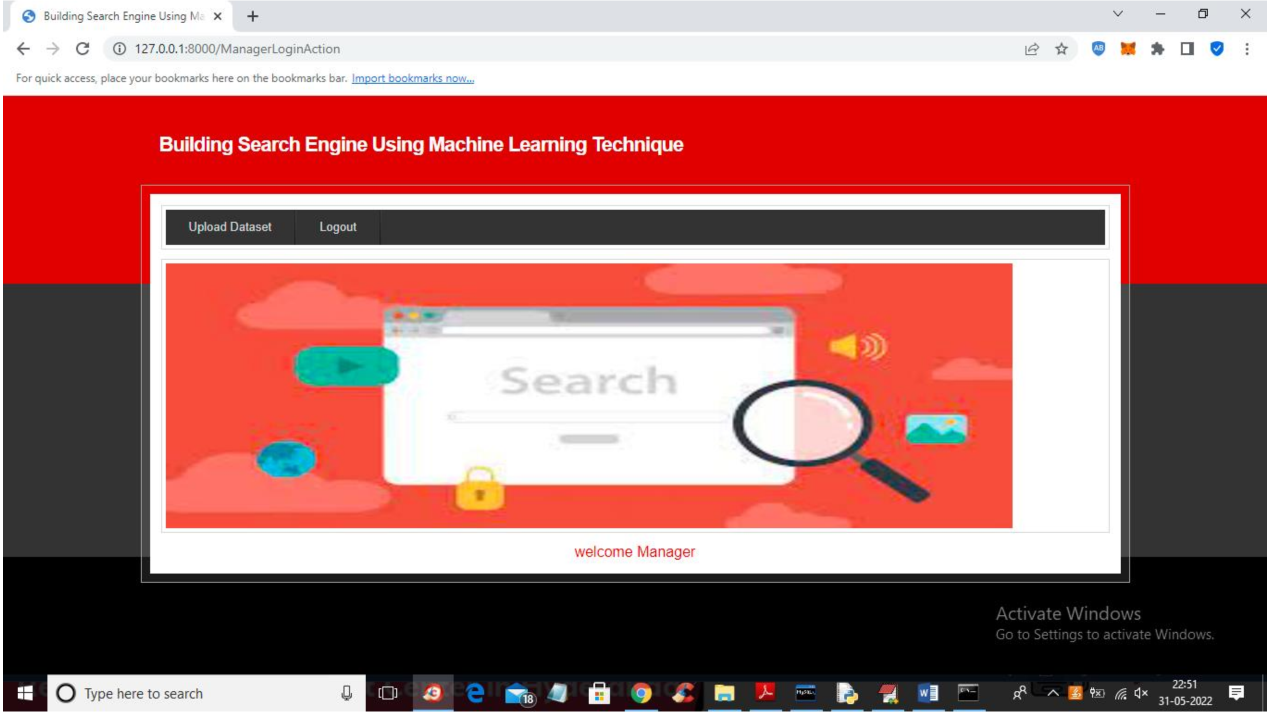
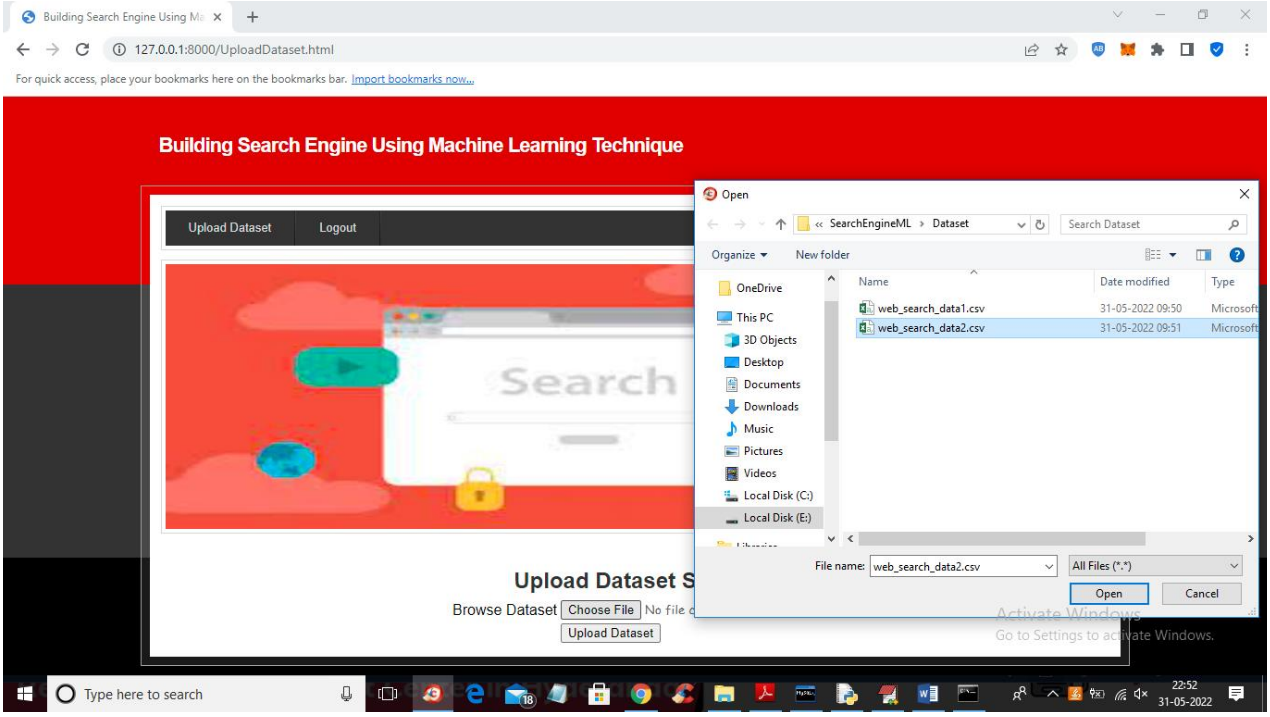

Building Search Engine Using Ma
127.0.0.1:8000/ManagerLoginAction
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
Upload Dataset
Logout
Search
welcome Manager
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:51
31-05-2022

Fig: Manager Login

UPLOAD DATASET



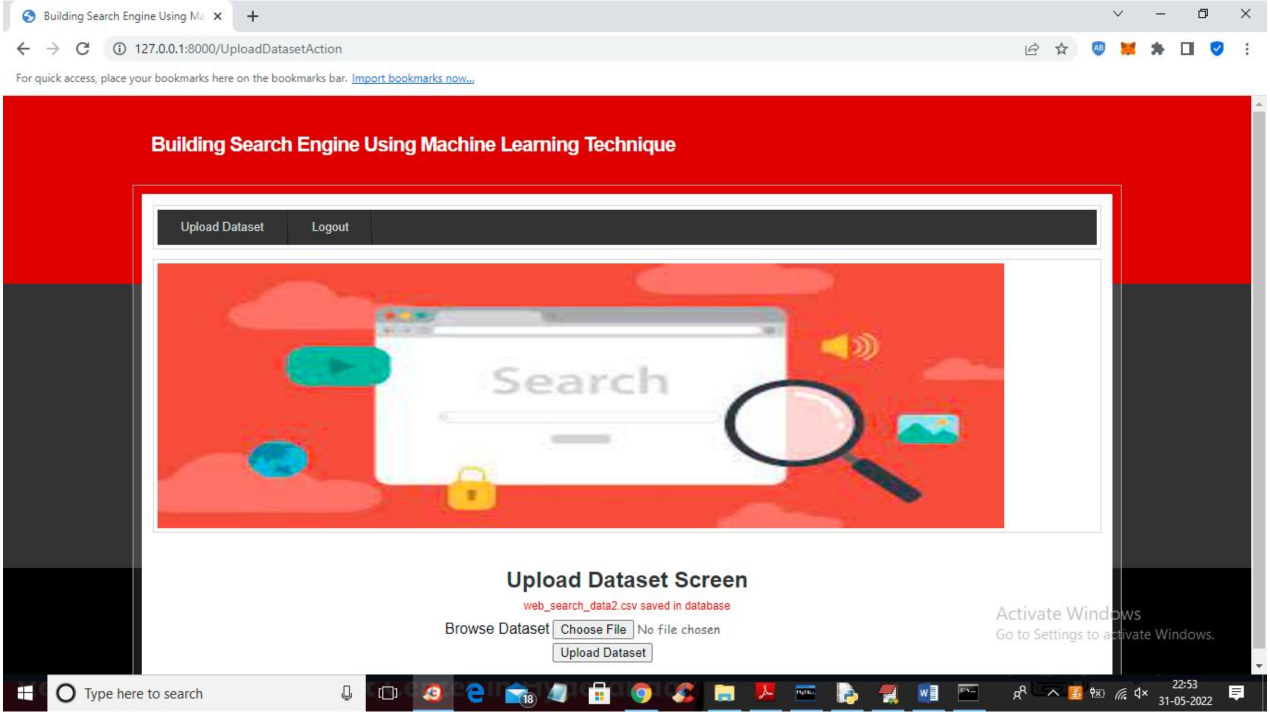
Name	Date modified	Type
web_search_data1.csv	31-05-2022 09:50	Microsoft
web_search_data2.csv	31-05-2022 09:51	Microsoft

File name: web_search_data2.csv

All Files (*.*)

Fig: Upload Dataset

DATASET FILE SAVED IN DATABASE



127.0.0.1:8000/UploadDatasetAction

Building Search Engine Using Machine Learning Technique

Upload Dataset Screen

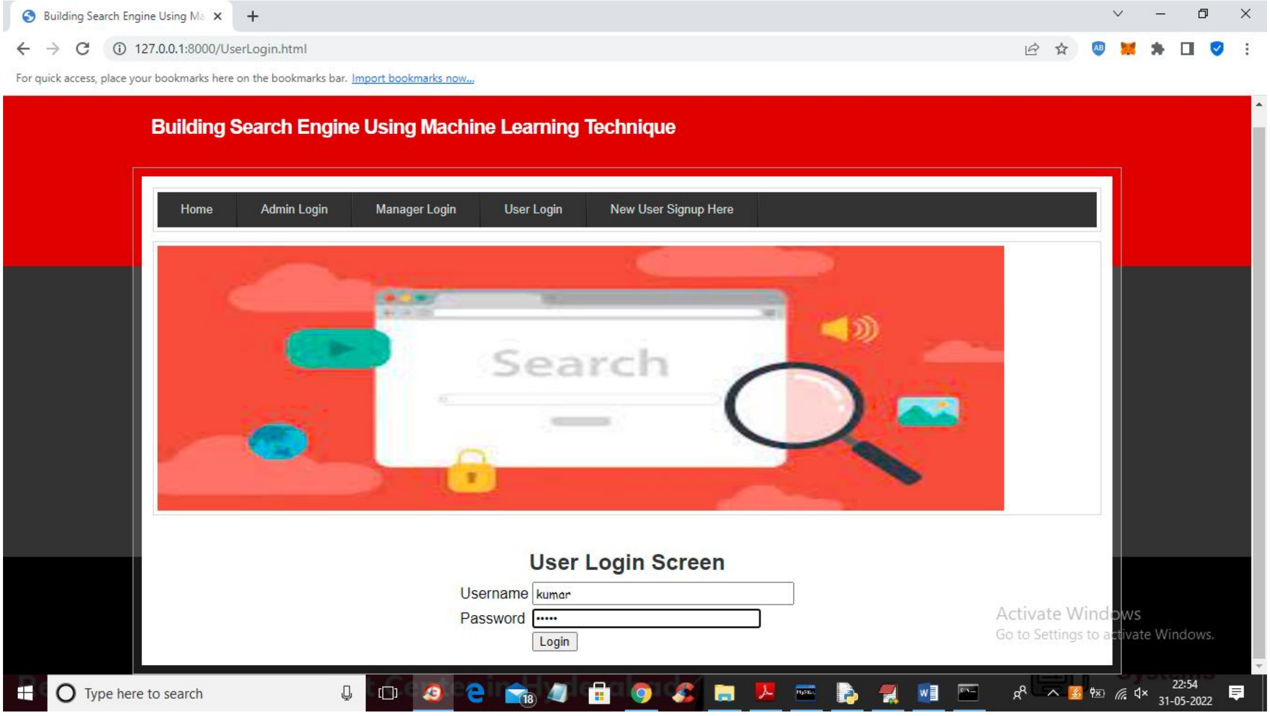
web_search_data2.csv saved in database

Browse Dataset Choose File No file chosen

Building Search Engine Using Ma
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Upload Dataset
Logout
Search
Upload Dataset
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:53
31-05-2022

Fig: dataset file saved in database

USER LOGIN SCREEN



Building Search Engine Using Ma
127.0.0.1:8000/UserLogin.html
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
Home
Admin Login
Manager Login
User Login
New User Signup Here
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:54
31-05-2022

User Login Screen

Username kumar

Password •••••

Login

Fig: User Login Screen

SEARCH WITH PAGE RANK LINK TO SEARCH ANY DATA

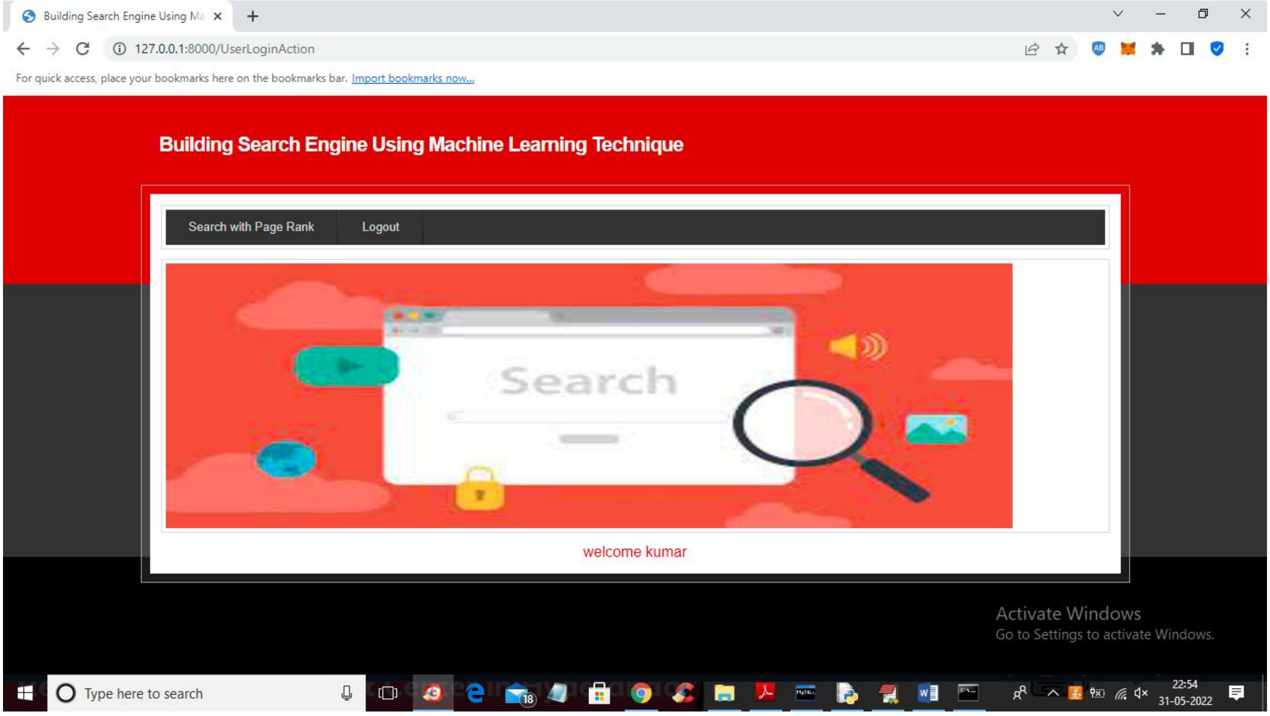
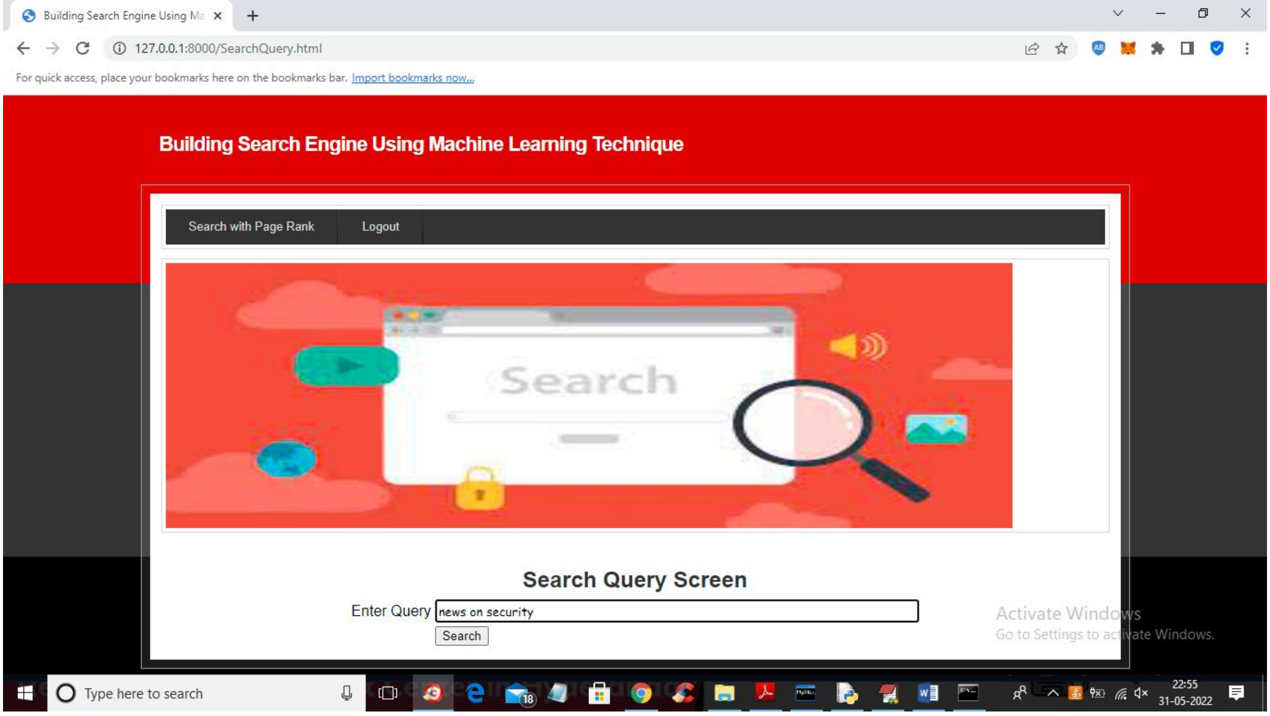

Building Search Engine Using Ma
127.0.0.1:8000/UserLoginAction
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
Search with Page Rank
Logout
Search
welcome kumar
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:54
31-05-2022

Fig: Search with Page Rank link to search any data

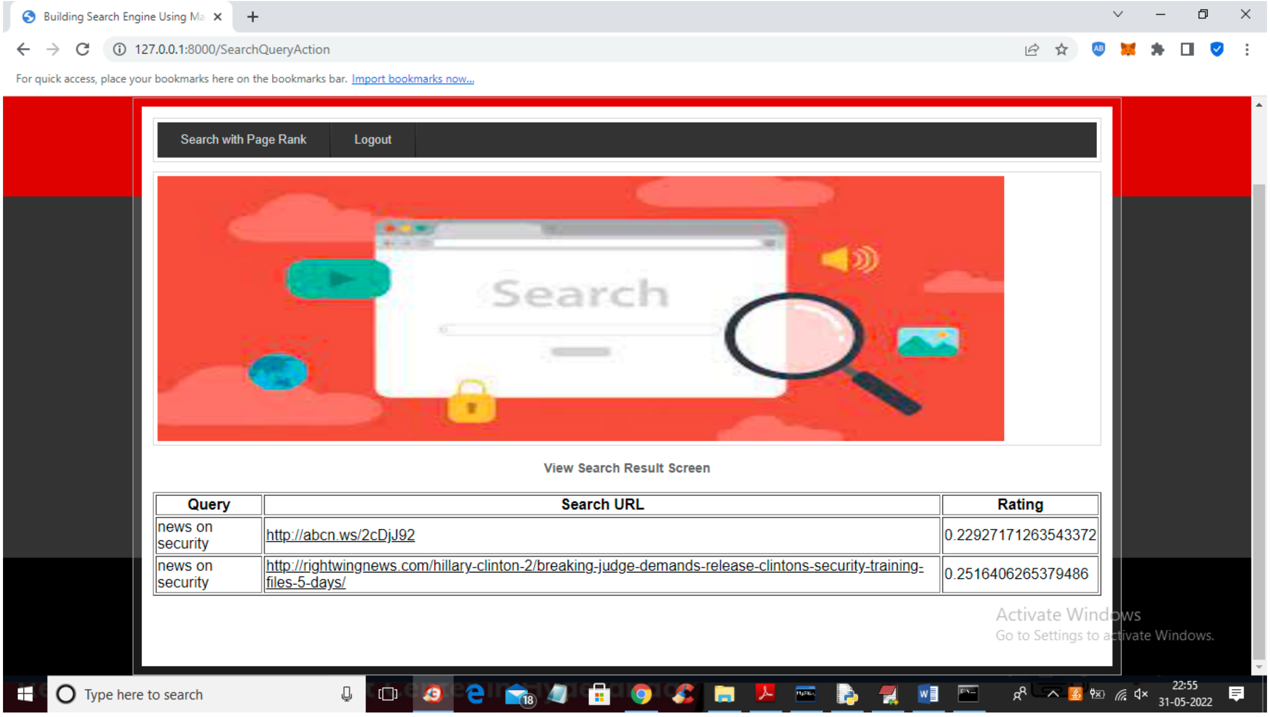
SEARCH QUERY SCREEN



Building Search Engine Using Ma
127.0.0.1:8000/SearchQuery.html
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
Search with Page Rank
Logout
Search
Search Query Screen
Enter Query
news on security
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:55
31-05-2022

Fig: Search Query Screen

VIEW SEARCH RESULT SCREEN

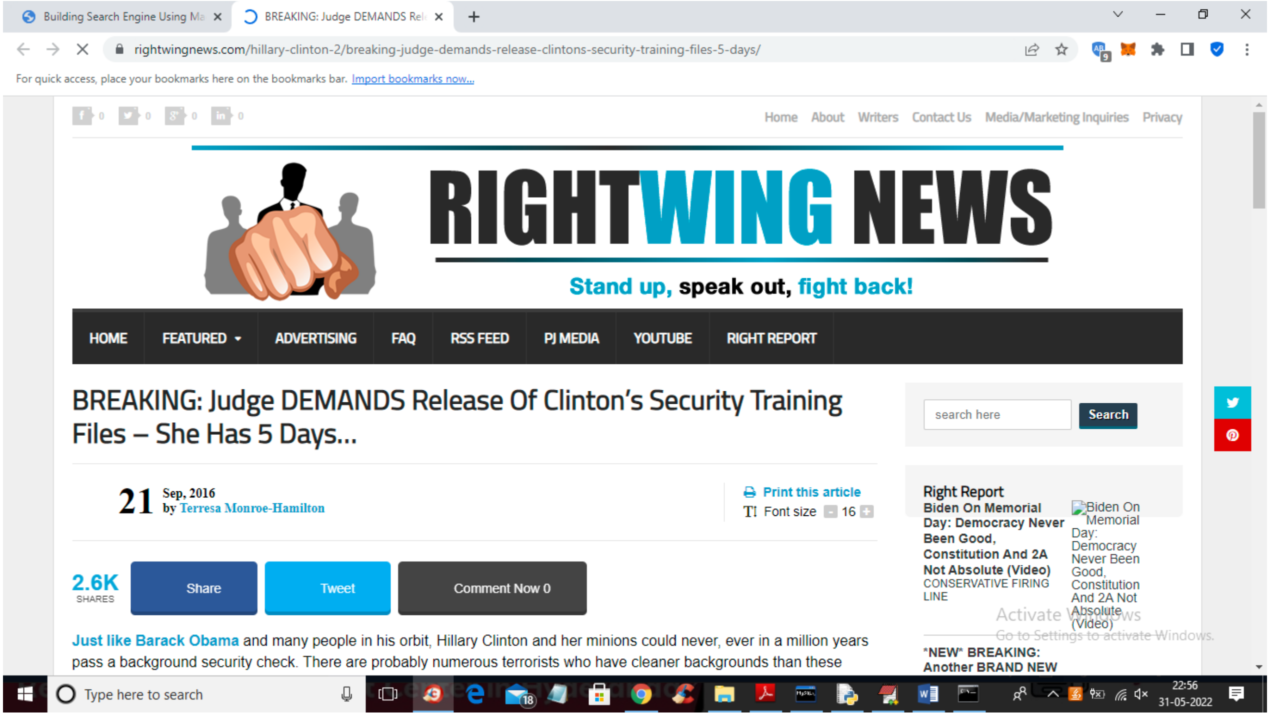

Building Search Engine Using Ma
127.0.0.1:8000/SearchQueryAction
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Search with Page Rank
Logout
Search
Type here to search
22:55
31-05-2022

View Search Result Screen

Query	Search URL	Rating
news on security	http://abcn.ws/2cDjJ92	0.22927171263543372
news on security	http://rightwingnews.com/hillary-clinton-2/breaking-judge-demands-release-clintons-security-training-files-5-days/	0.2516406265379486

Fig: View Search Result Screen

USER GIVEN QUERY AVAILABLE IN DATASET



Building Search Engine Using Ma
BREAKING: Judge DEMANDS Rel
rightwingnews.com/hillary-clinton-2/breaking-judge-demands-release-clintons-security-training-files-5-days/
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Home About Writers Contact Us Media/Marketing Inquiries Privacy
RIGHTWING NEWS
Stand up, speak out, fight back!
HOME FEATURED ADVERTISING FAQ RSS FEED PJ MEDIA YOUTUBE RIGHT REPORT
search here
Search
Print this article
Font size 16
2.6K SHARES
Share
Tweet
Comment Now 0
Type here to search
22:56
31-05-2022

BREAKING: Judge DEMANDS Release Of Clinton's Security Training Files – She Has 5 Days...

21 Sep, 2016
by Terresa Monroe-Hamilton

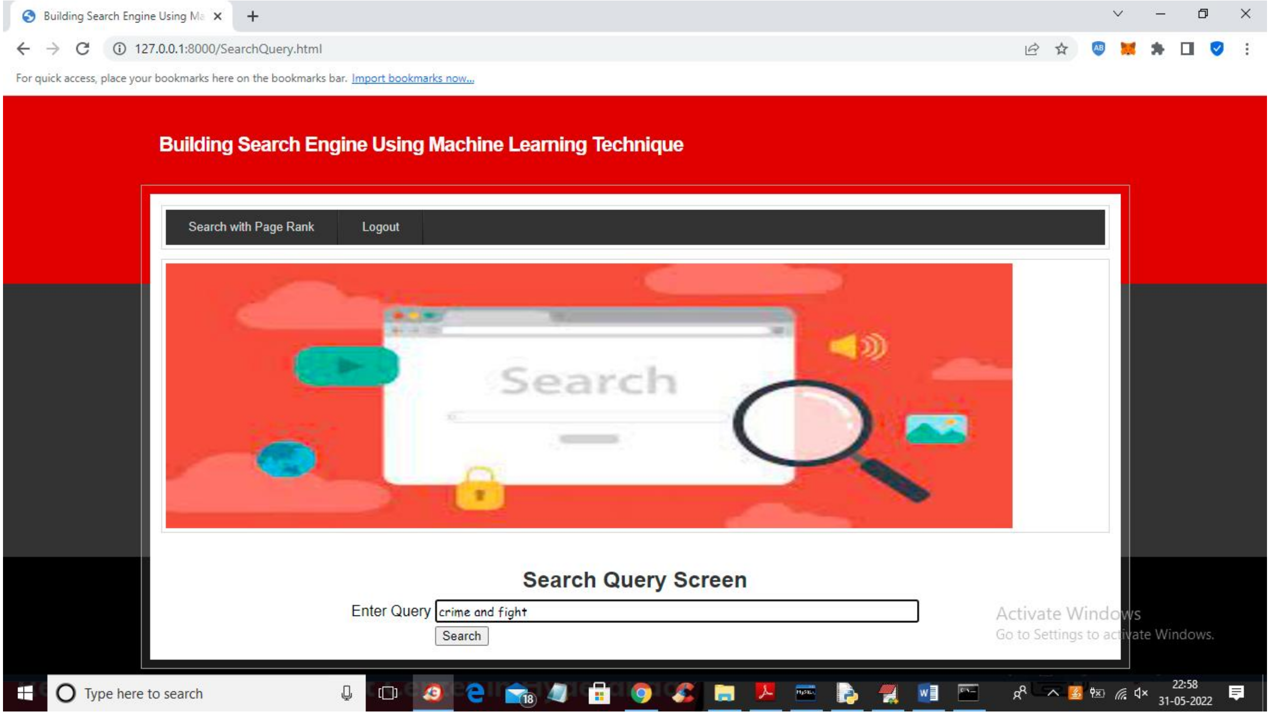
Just like Barack Obama and many people in his orbit, Hillary Clinton and her minions could never, ever in a million years pass a background security check. There are probably numerous terrorists who have cleaner backgrounds than these

Right Report
Biden On Memorial Day: Democracy Never Been Good, Constitution And 2A Not Absolute (Video)
CONSERVATIVE FIRING LINE

NEW BREAKING: Another BRAND NEW

Fig: User given Query available in dataset

SEARCH QUERY SCREEN


Building Search Engine Using Ma
127.0.0.1:8000/SearchQuery.html
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Search with Page Rank
Logout
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:58
31-05-2022

Building Search Engine Using Machine Learning Technique

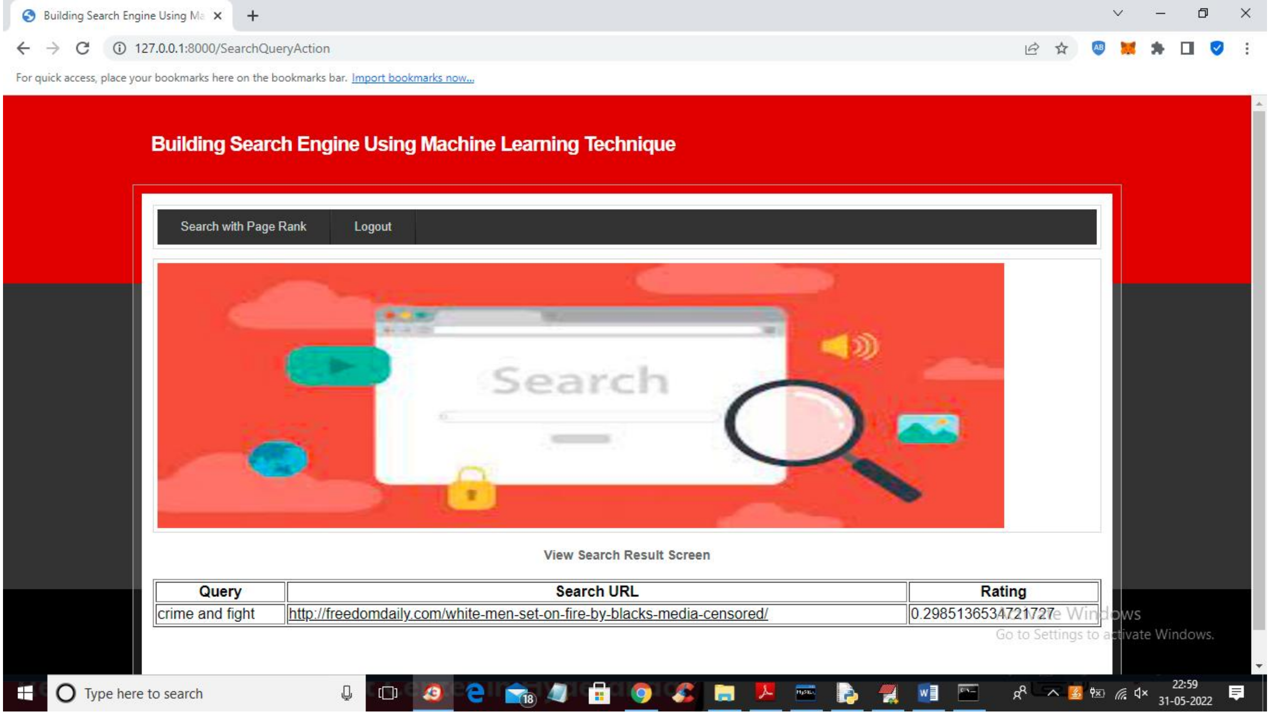
Search Query Screen

Enter Query crime and fight

Search

Fig: Search Query Screen

VIEW SEARCH RESULT SCREEN


Building Search Engine Using Ma
127.0.0.1:8000/SearchQueryAction
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Search with Page Rank
Logout
Search
Type here to search
22:59
31-05-2022

Building Search Engine Using Machine Learning Technique

View Search Result Screen

Query	Search URL	Rating
crime and fight	http://freedomdaily.com/white-men-set-on-fire-by-blacks-media-censored/	0.2985136534721727

Fig: View Search Result Screen

9.CONCLUSION

Search engines are very useful for finding out more relevant URLs for given keywords. Due to this, user time is reduced for searching the relevant web page. For this, Accuracy is a very important factor. From the above observation, it can be concluded that XGBoost is better in terms of accuracy than SVM and ANN. Thus, Search engines built using XGBoost and PageRank algorithms will give better accuracy.

10. REFERENCES

JOURNALS

[1] Manika Dutta, K. L. Bansal, “A Review Paper on Various Search Engines (Google, Yahoo, Altavista, Ask and Bing)”, International Journal on Recent and Innovation Trends in Computing and Communication, 2016.

[2] Gunjan H. Agre, Nikita Mahajan, “Keyword Focused Web Crawler”, International Conference on Electronic and Communication Systems, IEEE, 2015. [3] Tuhina Sen, Dev Kumar Chaudhary, “Contrastive Study of Simple PageRank, HITS and Weighted PageRank Algorithms: Review”, International Conference on Cloud Computing, Data Science & Engineering, IEEE, 2017.

[4] Michael Chau, Hsinchu Chen, “A machine learning approach to web page filtering using content and structure analysis”, Decision Support Systems 44 (2008) 482–494, science Direct,2008.

[5] Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, “Comparative Study of Page Rank and Weighted Page Rank Algorithm”, International Journal of Innovative Research in Computer and Communication Engineering, February 2014.

[6] K. R. Srinath, “Page Ranking Algorithms – A Comparison”, International Research Journal of Engineering and Technology (IRJET), Dec2017.

[7] S. Prabha, K. Doraiswamy, J. Indhumathi, “Comparative Analysis of Different Page Ranking Algorithms”, International Journal of Computer and Information Engineering, 2014.

[8] Dilip Kumar Sharma, A. K. Sharma, “A Comparative Analysis of Web Page Ranking Algorithms”, International Journal on Computer Science and Engineering, 2010.

[9] Vijay Chauhan, Arunima Jaiswal, Junaid Khalid Khan, “Web Page Ranking Using Machine Learning Approach”, International Conference on Advanced Computing Communication Technologies, 2015.

[10] Amanjot Kaur Sandhu, Tawie s. Liu., “Wikipedia Search Engine: Interactive Information Retrieval Interface Design”, International Conference on Industrial and Information Systems, 2014.

[11] Neha Sharma, Rashi Agarwal, Narendra Kohli, “Review of features and machine learning techniques for web searching”, International Conference on Advanced Computing Communication Technologies, 2016.

[12] Swear Liang Yong, Markus Hagen Buchner, Ah Chung Tsoi, “Ranking Web Pages using Machine Learning Approaches”, International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

[13] B. Jaganathan, Kalyani Desikan, “Weighted Page Rank Algorithm based on In-Out Weight of Webpages”, Indian Journal of Science and Technology, Dec-2015.

TEXTBOOKS

- Programming Python, Mark Lutz
- Head First Python, Paul Barry
- Core Python Programming, R. Nageswara Rao
- Learning with Python, Allen B. Downey

WEBSITES

- https://www.w3schools.com/python/
- https://www.tutorialspoint.com/python/index.htm
- https://www.javatpoint.com/python-tutorial
- https://www.learnpython.org/
- https://www.pythontutorial.net/

Term Paper

Developing a Search Engine Using XGB Algorithm

Mr. A. HEMANTHA KUMAR[1], Y. VARSHITHA[2]

[1] Associate Professor of MCA, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS) Gudur (M), Tirupati (Dt), AP

[2]PG Scholar, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS) Gudur (M), Tirupati (Dt), AP

ABSTRACT_ Building a Search Engine using Machine Learning Techniques The internet is the largest and most extravagant source of data. Search engines are widely used to retrieve information from the Internet. Search engines provide a simple interface for searching for user queries and providing results in the form of the web address of the relevant web page, but utilizing traditional search engines to find relevant information has become quite difficult. This study proposed a search engine that uses Machine Learning techniques to display more relevant web sites at the top of user queries.

1. INTRODUCTION

Internet is really a trap of individual frameworks and servers which are associated with various innovation and strategies. Each site includes the stacks of site pages that are being made and sent on the server. So on the off chance that a client needs something, the individual in question requirements to type a watchword. Watchword is a bunch of words removed from client search input. Search input given by a client might be linguistically erroneous. Here comes the real requirement for web crawlers. Web crawlers give you a basic point of interaction to look through client questions and show the outcomes. • Web crawlers help in gathering information about a site and the connections connected with them. We are just involving web crawlers for gathering information and data from WWW and putting away it in our data set.

• Indexer which orchestrates each term on each page and stores the ensuing rundown of terms in a colossal vault. • Question Motor is basically used to answer to the client's watchword and show the compelling result for their catchphrase. In the question motor, the Page positioning calculation positions the URL by involving various calculations in the inquiry motor. • This paper uses AI Procedures to find the highest level of reasonable web address for the given catchphrase. The result of the Page Rank calculation is given as contribution to the AI calculation.

2.LITERATURE SURVEY

2.1 S. Su, Y. Sun, X. Gao, J. Qiu* and Z. Tian*. A Correlation-change based Feature Selection Method for IoT Equipment Anomaly Detection. Applied Sciences.

In the era of the fourth industrial revolution, there is a growing trend to deploy sensors on industrial equipment, and analyze the industrial equipment's running status according to the sensor data. Thanks to the rapid development of IoT technologies [1], sensor data could be easily fetched from industrial equipment, and analyzed to produce further value for industrial control at the edge of the network or at data centers. Due to the considerable development of deep learning in recent years, a common practice of such analysis is to conduct deep learning [2,3,4]. Such methods select a subset of all fetched sensor data stream as the input features, and generate equipment predictions. As a result, the performance of the learning model was seriously impacted by the features selected, thus feature selection plays a critical role for such methods.

To select an appropriate set of features for the learning model, researchers aim to select the most relevant features to the prediction model to improve the prediction performance, or to select the most informative features to conduct data reduction. Unfortunately, both kinds of methods have intrinsic drawbacks when applied in the online scenarios. The former kind of methods seriously depends on predefined evaluation criteria, such as feature relevance metrics [5] or a predefined learning model [6]. Thus, such method are limited to certain dataset, and are not suitable for online scenarios which involve dynamical and unsupervised feature selection. The later kind of methods right fits in the online scenarios. However, data reduction mainly aims to improve the efficiency (but not accuracy) of the prediction model, which is not the most concerning factor of online industrial equipment status analysis.

To relieve the dependency of predefined evaluation criteria, researchers switch to select the features which can indicate the online sensor data's characters, such as features which are smoothest on the graph [7], or the features with highest clusterability [8,9]. In this paper, we focus on the features with correlation changes such as smoothness and clusterability, which are important characters for traditional pattern recognition fields like image processing and voice recognition [7,8,9]. We believe that correlation changes can significantly pinpoint status changes in industrial environment. As far as we know, this is the first work focusing on correlation changes for online feature selection.

2.2.X. Yu, Z. Tian, J. Qiu, F. Jiang. A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for Smart Mobile Devices. Wireless Communications and Mobile Computing, https://doi.org/10.1155/2018/5823439.

With the development of Internet and information technology, smart mobile devices appear in our daily lives, and the problem of information leakage on smart mobile devices will follow which has become more and more serious [1, 2]. All kinds of private or sensitive information, such as intellectual property and financial data, might be distributed to unauthorized entity intentionally or accidentally. And that it is impossible to prevent from spreading once the confidential information has leaked.

According to survey reports [3, 4], most of the threats to information security are caused by internal data leakage. These internal threats consist of approximate 29% private or sensitive accidental data leakage, approximate 16% theft of intellectual property, and approximate 15% other thefts including customer information, and financial data. Further, the consensus of approximate 67% organizations shows that the damage caused from internal threats is more serious than those form outside.

Although laws and regulations have been passed to punish various behaviors of intentional data leakage, it is still hard to prevent data leakage effectively. Confidential data can be easily disguised by rephrasing confidential contents or embedding confidential contents in nonconfidential contents [5, 6]. In order to avoid the problems arising from data leakage, lots of software and hardware solutions have been developed which are discussed in the following chapter.

In this paper, we present CBDLP, a data leakage prevention model based on confidential terms and their context terms, which can detect the rephrased confidential contents effectively. In CBDLP, a graph structure with confidential terms and their context involved is adopted to represent documents of the same class, and then the confidentiality score of the document to be detected is calculated to justify whether confidential contents is involved or not. Based on the attribute reduction method from rough set theory, we further propose a pruning method. According to the importance of the confidential terms and their context, the graph structure of each cluster is updated after pruning. The motivation of the paper is to develop a solution which can prevent intentional or accidental data leakage from insider effectively. As mixed-confidential documents are very common, it is very important to accurately detect the documents containing confidential contents even when most of the confidential contents have been rephrased.

2.3 Y. Sun, M. Li, S. Su, Z. Tian, W. Shi, M. Han. Secure Data Sharing Framework via Hierarchical Greedy Embedding in Darknets. ACM/Springer Mobile Networks an

Geometric routing, which combines greedy embedding and greedy forwarding, is a promising approach for efficient data sharing in darknets. However, the security of data sharing using geometric routing in darknets is still an issue that has not been fully studied. In this paper, we propose a Secure Data Sharing framework (SeDS) for uture darknets via hierarchical greedy embedding. SeDS adopts a hierarchical topology and uses a set of secure nodes to protect the whole topology. To support geometric routing in the hierarchical topology, a two-level bit-string prefix embedding approach (Prefix-T) is first proposed, and then a greedy forwarding strategy and a data mapping approach are combined with Prefix-T for data sharing. SeDS guarantees that the publication or request of a data item can always pass through the corresponding secure node, such that security strategies can be performed. The experimental results show that SeDS provides scalable and efficient end-to-end communication and data sharing.

2.4 Z. Wang, C. Liu, J. Qiu, Z. Tian, C., Y. Dong, S. Su Automatically Traceback RDP-based Targeted Ransomware Attacks. Wireless Communications and Mobile Computing. 2018. https://doi.org/10.1155/2018/7943586.

With the popularization of new energy electric vehicles (EVs), the recommendation algorithm is widely used in the relatively new feld of charge piles. At the same time, the construction of charging infrastructure is facing increasing demand and more severe challenges. With the ubiquity of Internet of vehicles (IoVs), inter-vehicle communication can share information about the charging experience and trafc condition to help achieving better charging recommendation and higher energy efciency. The recommendation of charging piles is of great value. However, the existing methods related to such recommendation consider inadequate reference factors and most of them are generalized for all users, rather than personalized for specifc populations

3. PROPOSED SYSTEM

In this research, the author uses machine learning methods known as SVM and XGBOOST to forecast search results for a given query and to design a search engine using machine learning algorithms. To train this method, the author uses website data, which is then turned into a numeric vector known as TFIDF (term frequency inverse document frequency). TFIDF vectors contain the average frequency of each word. The proposed search engine is highly helpful in locating more relevant URLs for provided keywords.

3.1 IMPLEMENTATION

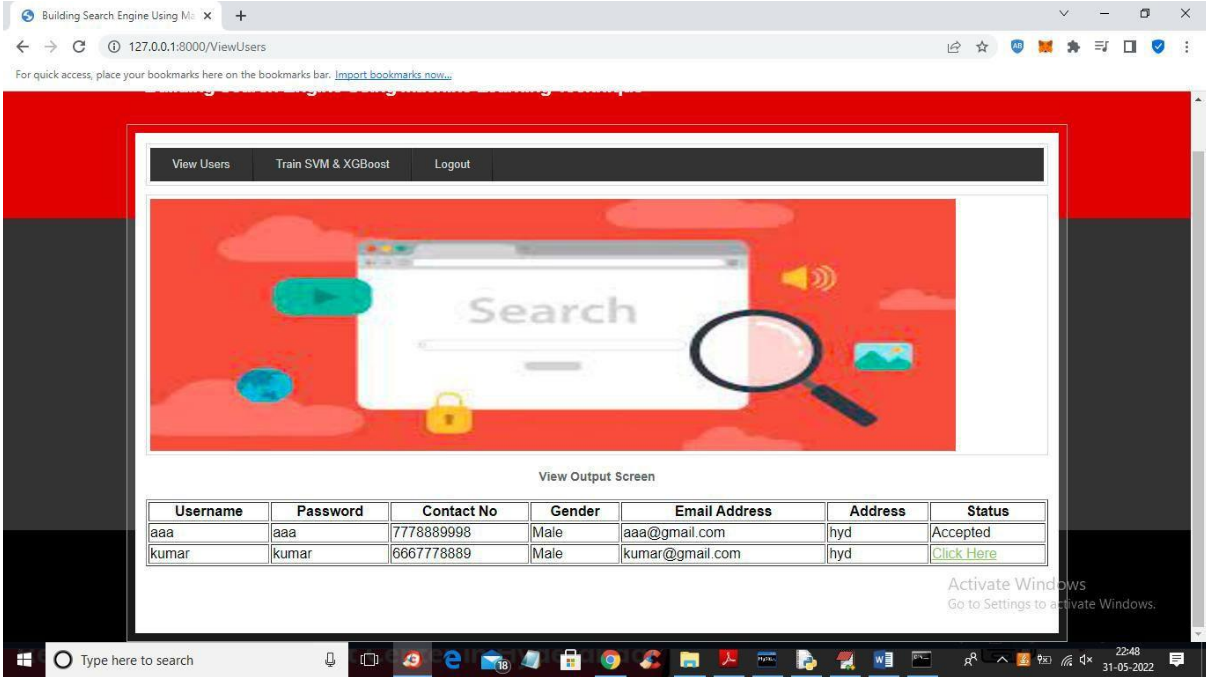
1) Admin module: admin can login to application using username and password as admin and then accept or activate new users registration and then train SVM and XGBOOST algorithm

2) Manager module: manager can login to application by using username and password as Manager and Manager and then upload dataset to application

3) New User Signup: using this module new user can signup with the application

4) User Login: user can login to application and then perform search by giving query.

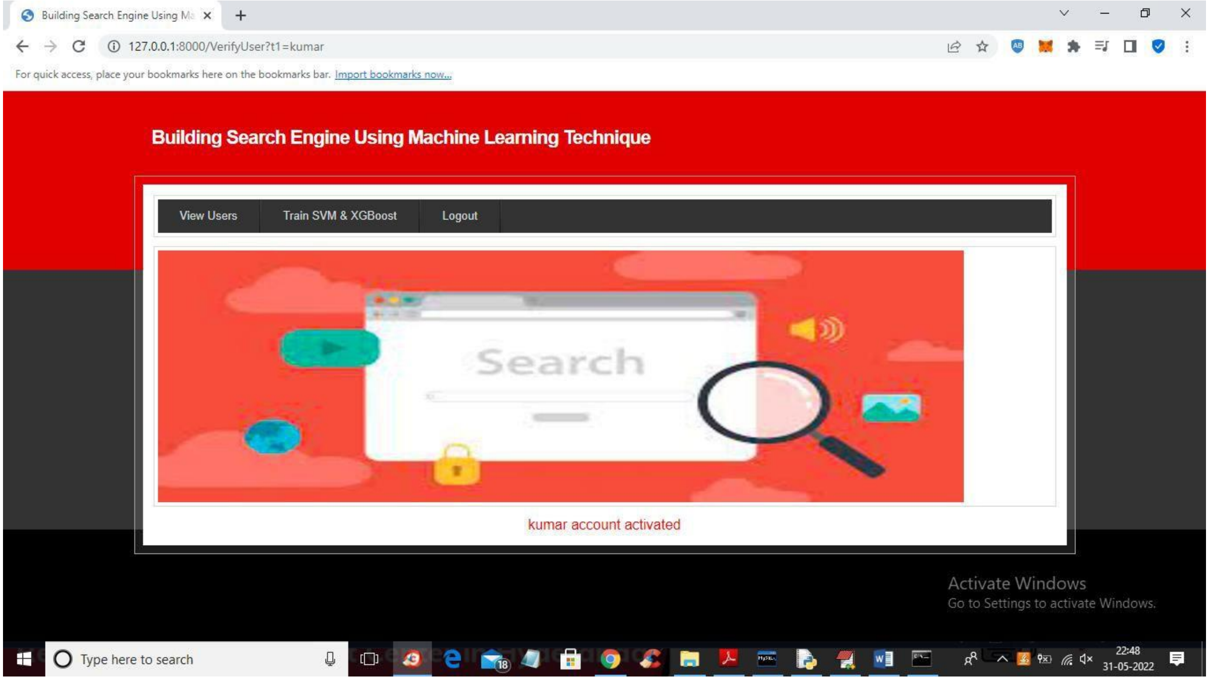
4. RESULTS AND DISCUSSION


Building Search Engine Using Ma
127.0.0.1:8000/ViewUsers
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
View Users
Train SVM & XGBoost
Logout
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:48
31-05-2022

View Output Screen

Username	Password	Contact No	Gender	Email Address	Address	Status
aaa	aaa	7778889998	Male	aaa@gmail.com	hyd	Accepted
kumar	kumar	6667778889	Male	kumar@gmail.com	hyd	Click Here

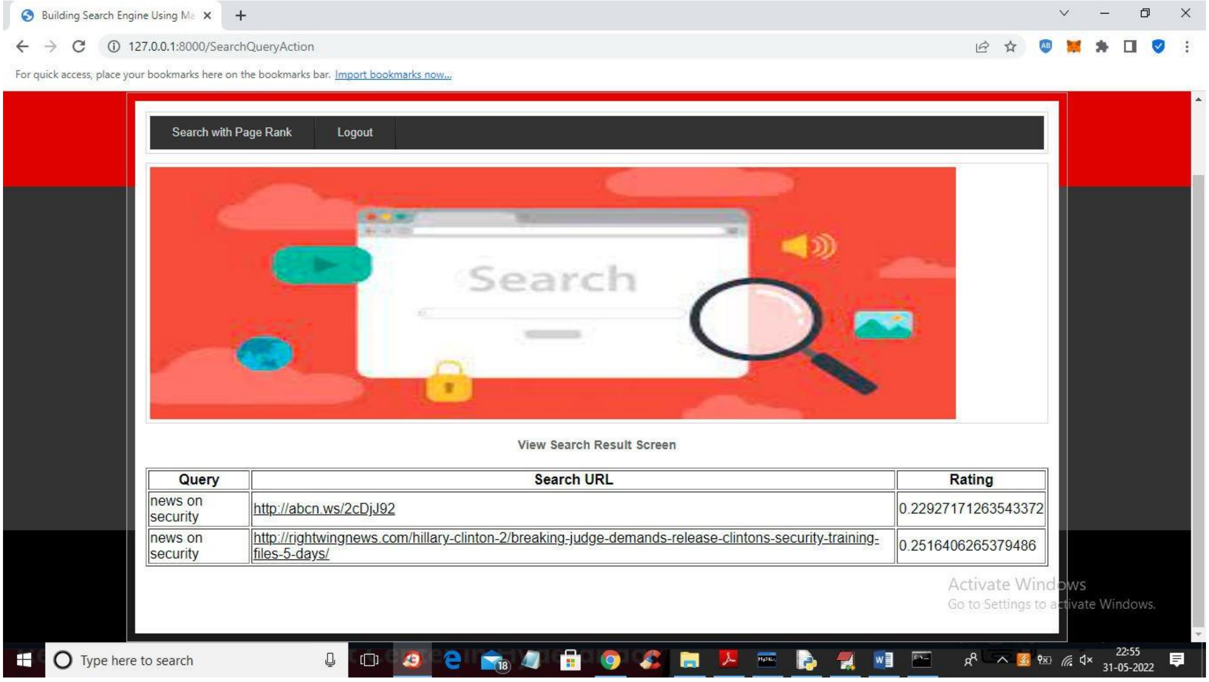
In above screen admin can click on 'Click Here' link to activate that user account



Building Search Engine Using M
127.0.0.1:8000/VerifyUser?t1=kumar
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Building Search Engine Using Machine Learning Technique
View Users
Train SVM & XGBoost
Logout
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:48
31-05-2022

kumar account activated

In above screen we can see admin activated kumar user account and now admin can click on 'Train SVM & XGBOOST' link to train machine learning SVM and XGBOOST algorithm and get below output

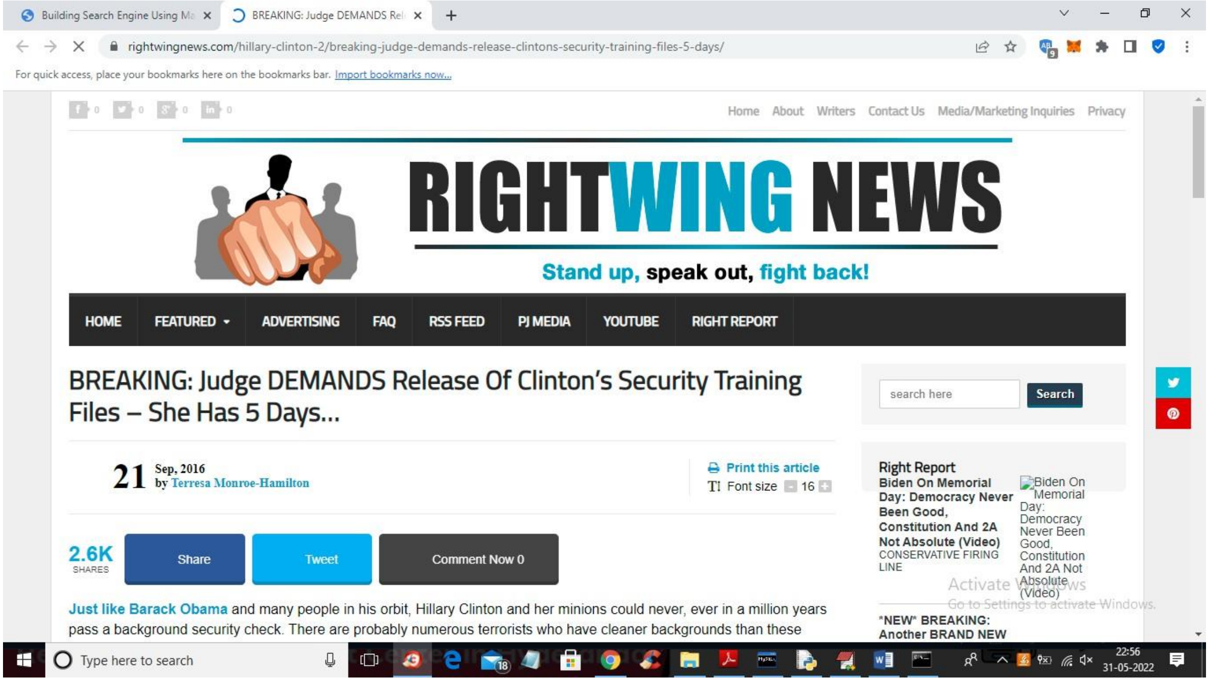


Building Search Engine Using M
127.0.0.1:8000/SearchQueryAction
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Search with Page Rank
Logout
Search
Activate Windows
Go to Settings to activate Windows.
Type here to search
22:55
31-05-2022

View Search Result Screen

Query	Search URL	Rating
news on security	http://abcn.ws/2cDjJ92	0.22927171263543372
news on security	http://rightwingnews.com/hillary-clinton-2/breaking-judge-demands-release-clintons-security-training-files-5-days/	0.2516406265379486

In above screen machine learning algorithm predicts two URLS for given query and user can click on those URLS to visit page



rightwingnews.com/hillary-clinton-2/breaking-judge-demands-release-clintons-security-training-files-5-days/

BREAKING: Judge DEMANDS Release Of Clinton's Security Training Files – She Has 5 Days...

21 Sep, 2016 by Terresa Monroe-Hamilton

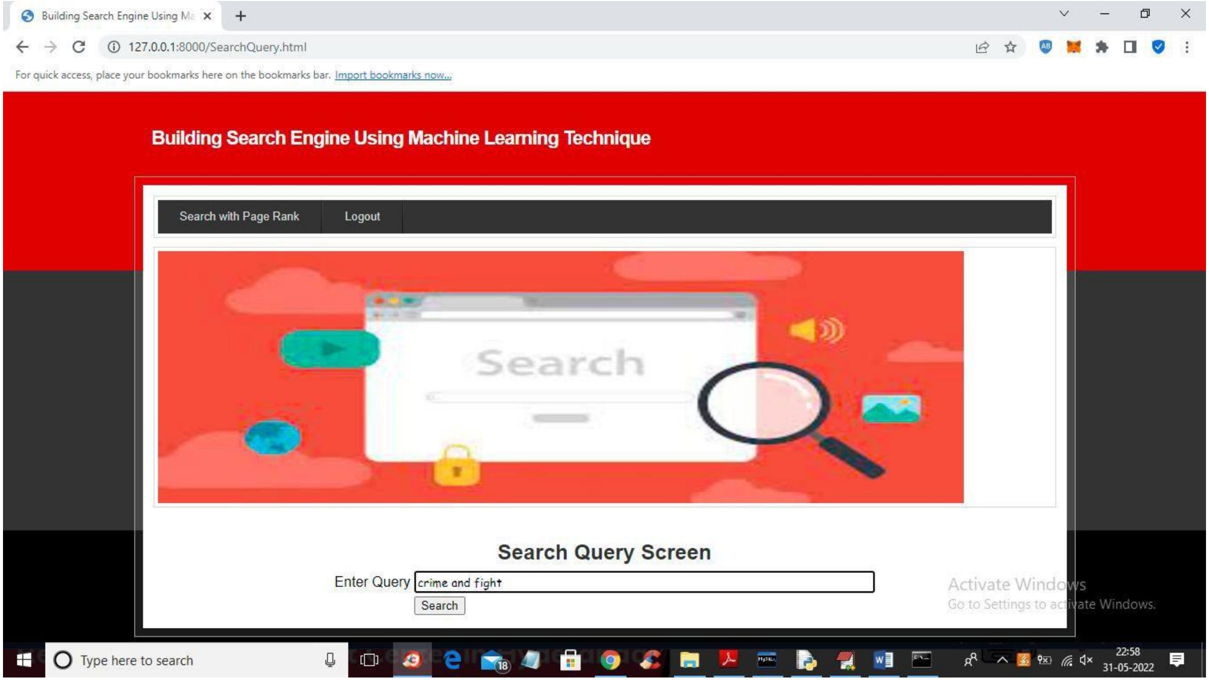
Just like Barack Obama and many people in his orbit, Hillary Clinton and her minions could never, ever in a million years pass a background security check. There are probably numerous terrorists who have cleaner backgrounds than these

Right Report
Biden On Memorial Day: Democracy Never Been Good, Constitution And 2A Not Absolute (Video)
CONSERVATIVE FIRING LINE

"NEW" BREAKING: Another BRAND NEW

Building Search Engine Using Ma
BREAKING: Judge DEMANDS Rel
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Home About Writers Contact Us Media/Marketing Inquiries Privacy
RIGHTWING NEWS
Stand up, speak out, fight back!
HOME FEATURED ADVERTISING FAQ RSS FEED PJ MEDIA YOUTUBE RIGHT REPORT
search here
Search
Print this article
Font size 16
2.6K SHARES
Share
Tweet
Comment Now 0
Type here to search
22:56
31-05-2022

In above screen by clicking on URL link user can visit and view page. Similarly user can give any query and if query available in dataset then he will get output



127.0.0.1:8000/SearchQuery.html

Building Search Engine Using Machine Learning Technique

Search Query Screen

Enter Query crime and fight

Building Search Engine Using Ma
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Search with Page Rank
Logout
Search
Search
Type here to search
22:58
31-05-2022

For above query we got below result

5,CONCLUSION

Search engines are quite effective for discovering more relevant URLs for specific keywords. As a result, the amount of time users spend searching for relevant web pages is reduced. Accuracy is quite crucial in this regard. Based on the observations above, it is possible to conclude that XGBoost outperforms SVM and ANN in terms of accuracy. Thus, search engines developed with the XGBoost and PageRank algorithms will provide greater accuracy.

REFERENCES

[1] Manika Dutta, K. L. Bansal, “A Review Paper on Various Search Engines (Google, Yahoo, Altavista, Ask and Bing)”, International Journal on Recent and Innovation Trends in Computing and Communication, 2016.

[2] Gunjan H. Agre, Nikita V.Mahajan, “Keyword Focused Web Crawler”, International Conference on Electronic and Communication Systems, IEEE, 2015. [3] Tuhena Sen, Dev Kumar Chaudhary, “Contrastive Study of Simple PageRank, HITS and Weighted PageRank Algorithms: Review”, International Conference on Cloud Computing, Data Science & Engineering, IEEE, 2017.

[4] Michael Chau, Hsinchun Chen, “A machine learning approach to web page filtering using content and structure analysis”, Decision Support Systems 44 (2008) 482–494,scienceDirect,2008.

[5] Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, “Comparative Study of Page Rank and Weighted Page Rank Algorithm”, International Journal of Innovative Research in Computer and Communication Engineering, February 2014.

[6] K. R. Srinath, “Page Ranking Algorithms – A Comparison”, International Research Journal of Engineering and Technology (IRJET), Dec2017.

[7] S. Prabha, K. Duraiswamy, J. Indhumathi, “Comparative Analysis of Different Page Ranking Algorithms”, International Journal of Computer and Information Engineering, 2014.

[8] Dilip Kumar Sharma, A. K. Sharma, “A Comparative Analysis of Web Page Ranking Algorithms”, International Journal on Computer Science and Engineering, 2010.

[9] Vijay Chauhan, Arunima Jaiswal, Junaid Khalid Khan, “Web Page Ranking Using Machine Learning Approach”, International Conference on Advanced Computing Communication Technologies, 2015.

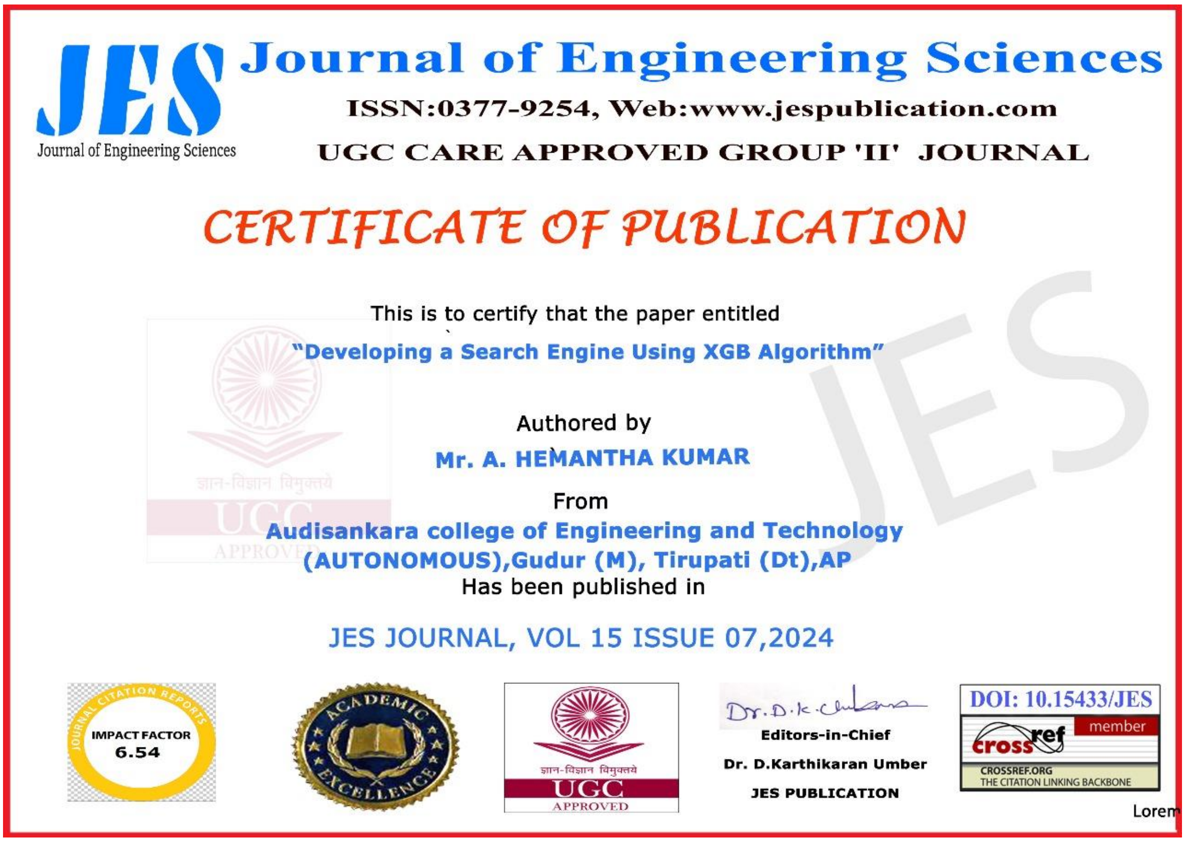
[10] Amanjot Kaur Sandhu, Tiewei s. Liu., "Wikipedia Search Engine: Interactive Information Retrieval Interface Design", International Conference on Industrial and Information Systems, 2014.

[11] Neha Sharma, Rashi Agarwal, Narendra Kohli, "Review of features and machine learning techniques for web searching", International Conference on Advanced Computing Communication Technologies, 2016.

Author Profiles

Mr. A. HEMANTHA KUMAR currently he is working Associate Professor in Audisankara College of Engineering & Technology Gudur(M), Tirupati (DT), he is done MCA from Priyadarshini Post Graduation Centre in 2000, M.E at Sathyabama University Chennai in 2006, Pursuing PHD at Bharath University Chennai.

Y. VARSHITHA is pursuing MCA from Audisankara college of Engineering &Technology (AUTONOMOUS), Gudur, Affiliated to JNTUA in 2024. Andhra Pradesh, India.



Journal of Engineering Sciences
ISSN:0377-9254, Web:www.jespublication.com
UGC CARE APPROVED GROUP 'II' JOURNAL

CERTIFICATE OF PUBLICATION

This is to certify that the paper entitled

"Developing a Search Engine Using XGB Algorithm"

Authored by

Mr. A. HEMANTHA KUMAR

From

Audisankara college of Engineering and Technology (AUTONOMOUS),Gudur (M), Tirupati (Dt),AP

Has been published in

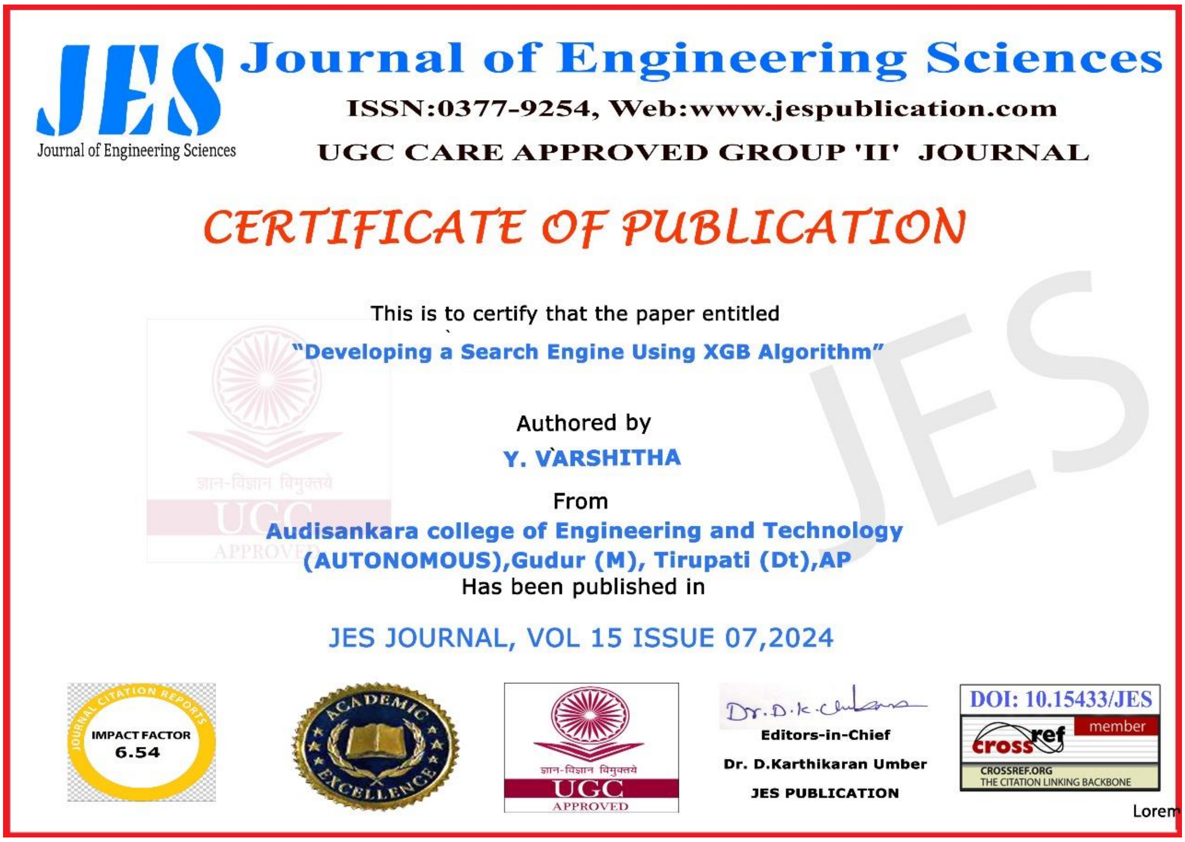
JES JOURNAL, VOL 15 ISSUE 07,2024

IMPACT FACTOR 6.54

Editors-in-Chief
Dr. D.Karthikaran Umber
JES PUBLICATION

DOI: 10.15433/JES

Lorem



JES
Journal of Engineering Sciences

Journal of Engineering Sciences
ISSN:0377-9254, Web:www.jespublication.com
UGC CARE APPROVED GROUP 'II' JOURNAL

CERTIFICATE OF PUBLICATION

This is to certify that the paper entitled

"Developing a Search Engine Using XGB Algorithm"

Authored by

Y. VARSHITHA

From

Audisankara college of Engineering and Technology (AUTONOMOUS),Gudur (M), Tirupati (Dt),AP

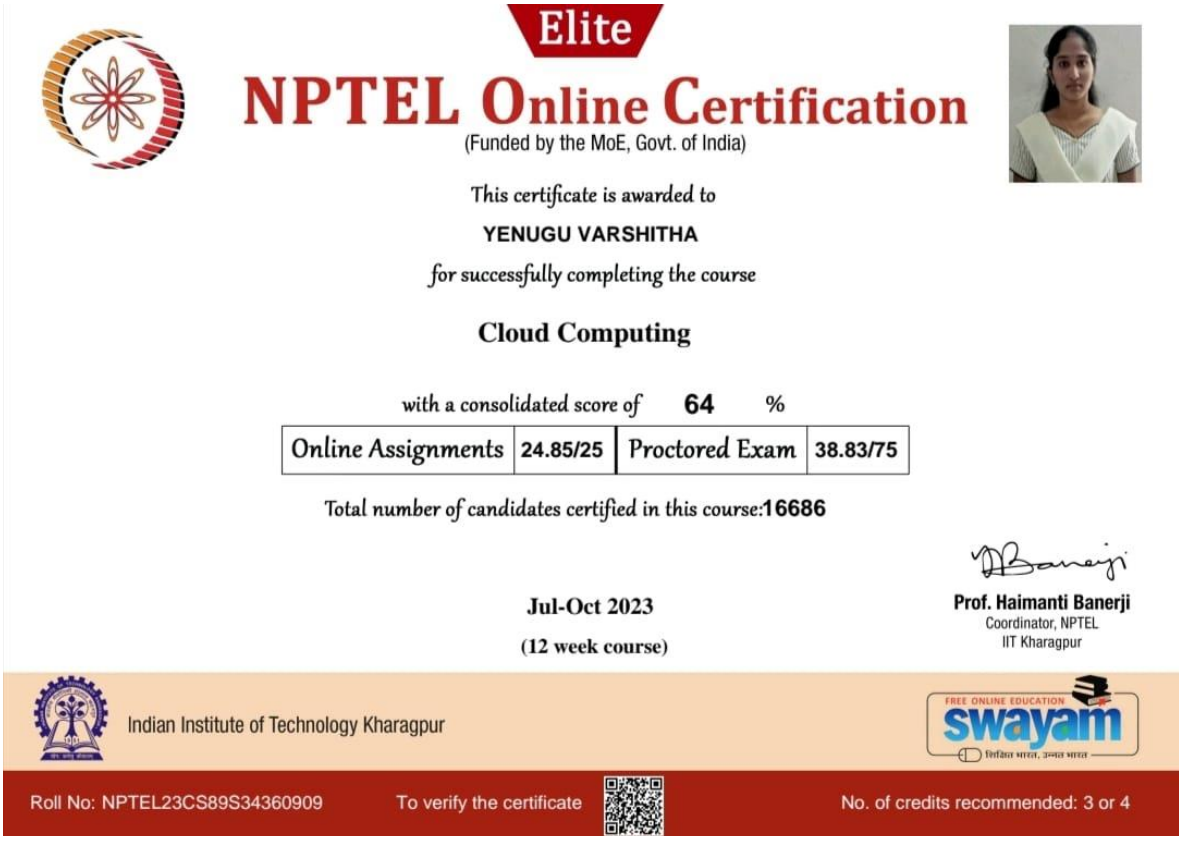
Has been published in

JES JOURNAL, VOL 15 ISSUE 07,2024

Dr.D.k.
Editors-in-Chief
Dr. D.Karthikaran Umber
JES PUBLICATION

DOI: 10.15433/JES
crossref member
CROSSREF.ORG
THE CITATION LINKING BACKBONE

Lorem



Elite

NPTEL Online Certification

(Funded by the MoE, Govt. of India)

This certificate is awarded to

YENUGU VARSHITHA

for successfully completing the course

Cloud Computing

with a consolidated score of **64** %

Online Assignments	**24.85/25**	Proctored Exam	**38.83/75**

*Total number of candidates certified in this course:***16686**

Jul-Oct 2023

(12 week course)

HBanerji

Prof. Haimanti Banerji
Coordinator, NPTEL
IIT Kharagpur

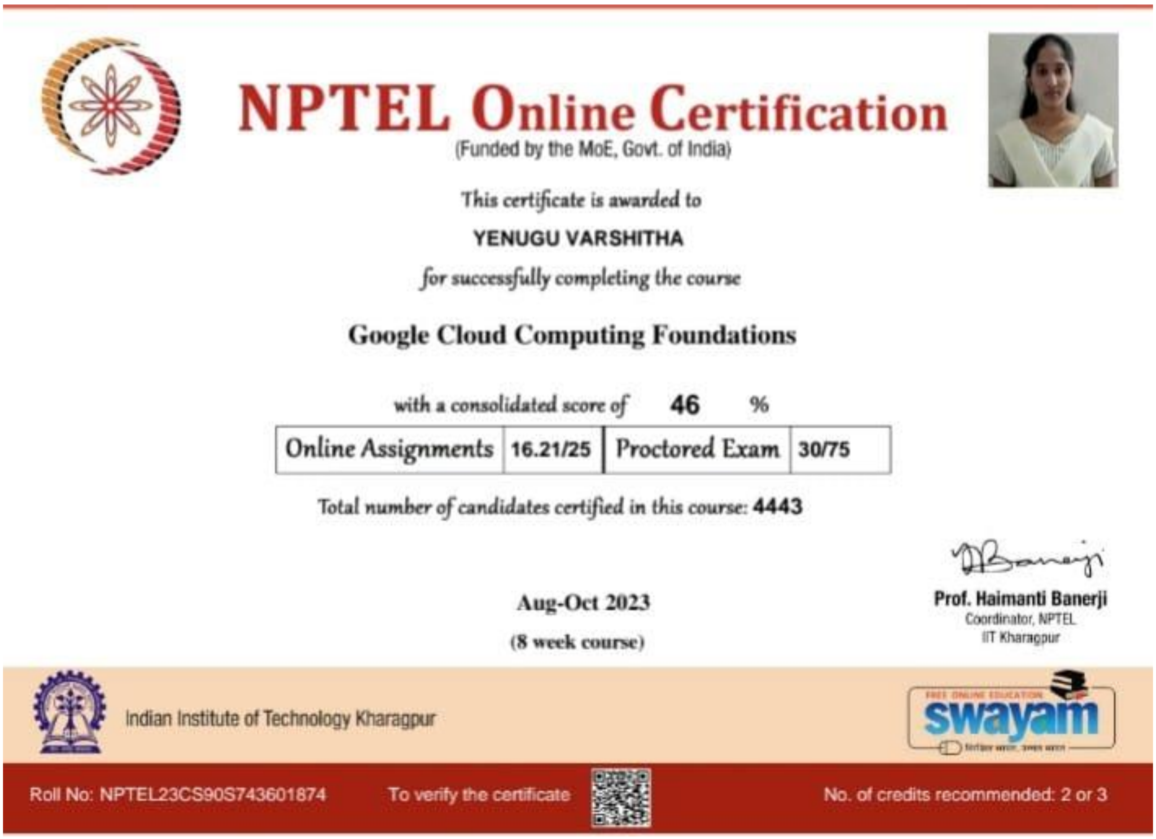
Indian Institute of Technology Kharagpur

FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत

Roll No: NPTEL23CS89S34360909

To verify the certificate

No. of credits recommended: 3 or 4



NPTEL Online Certification

(Funded by the MoE, Govt. of India)

This certificate is awarded to

YENUGU VARSHITHA

for successfully completing the course

Google Cloud Computing Foundations

with a consolidated score of **46** %

Online Assignments	16.21/25	Proctored Exam	30/75

Total number of candidates certified in this course: **4443**

Aug-Oct 2023

(8 week course)

DBanerji

Prof. Haimanti Banerji
Coordinator, NPTEL
IIT Kharagpur

Indian Institute of Technology Kharagpur

FREE ONLINE EDUCATION
swayam

Roll No: NPTEL23CS90S743601874

To verify the certificate

No. of credits recommended: 2 or 3

THANK YOU