An abstract on

UNDERSTANDING SHORT TEXTS THROUGH SEMANTIC ENRICHMENT AND SEMANTIC HASHING

Submitted in partial fulfillment of the requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY

in

Computer Science & Engineering

by

Ramya S	**(164G1A0575)**
Rekha G	**(164G1A0579)**
Ramya Sai N	**(164G1A0576)**
Sai Kiran K	**(164G1A0586)**
Rama Mohan Chowdary K	**(164G1A0574)**

Under the Guidance of

Mr .T. Murali Krishna **MCA, M. Phil, M. Tech, (PhD)**

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY

ANANTHAPURAMU

(Accredited by NAAC with 'A' Grade & Accredited by NBA, Affiliated to JNTUA, Approved by AICTE, New Delhi)

2019-2020

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY:ANANTAPURAMU
(Accredited by NAAC with 'A' Grade, Affiliated to JNTUA, Approved by AICTE, New Delhi)

Certificate

This is to certify that the project report entitled Understanding Short Text Through Semantic Enrichment and Semantic Hashing is the bonafide work carried out by **Ramya S** bearing Roll Number 164G1A0575**, Rekha G** bearing Roll Number 164G1A0579 , **Ramya Sai N** bearing Roll Number 164G1A0576, Sai Kiran K bearing Roll Number 164G1A0586 and **Rama Mohan Chowdary K** bearing Roll Number 164G1A0574 in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** during the academic year 2019-2020.

Guide	**Head of the Department**
Mr. T. Murali Krishna MCA, M. Phil, M. Tech, (PhD) Assistant Professor	Dr. G.K.V. Narasimha Reddy, Ph.D Professor & HOD
Date:	**EXTERNAL EXAMINER**

Ananthapuramu

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that we have now the opportunity to express my gratitude for all of them.

It is with immense pleasure that we would like to express my indebted gratitude to my Guide **Mr. T. Murali Krishna , Assistant Professor, Computer Science & Engineering**, who has guided me a lot and encouraged me in every step of the project work. We thank him for the stimulating guidance, constant encouragement and constructive criticism which have made possible to bring out this project work.

We express our deep-felt gratitude to **Mr. R.Sandeep Kumar, (Ph.D),Assistant Professor,** project coordinator valuable guidance and unstinting encouragement enable us to accomplish our project successfully in time.

We are very much thankful to **Dr. G.K.V. Narasimha Reddy, Ph.D , Professor & Head of the Department, Computer Science & Engineering,** for his kind support and for providing necessary facilities to carry out the work.

We wish to convey my special thanks to **Dr.T.Hitendra Sarma, Ph.D, Principal** of **Srinivasa Ramanujan Institute of Technology** for giving the required information in doing our project work. Not to forget, we thank all other faculty and non-teaching staff, and my friends who had directly or indirectly helped and supported us in completing our project in time.

We also express our sincere thanks to the Management for providing excellent facilities**.**

Finally, we wish to convey our gratitude to our family who fostered all the requirements and facilities that we need.

Project Associates

DECLARATION

We Ramya S bearing reg no : 164G1A0575, Rekha G bearing reg no : 164G1A0579, Ramya Sai N bearing reg no : 164G1A0576, Sai Kiran K bearing reg no : 164G1A0586, Rama Mohan Chowdary K bearing reg no: 164G1A0574, students of SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY, Rotarypuram , hereby declare that the dissertation entitled "UNDERSTANDING SHORT TEXT THROUGH SEMANTIC ENRICHMENT AND SEMANTIC HASHING" embodies the report of our project work carried out by us during IV Year Bachelor of Technology under the guidance of Mr. T. MURALI KRISHNA, MCA, M. Phil, M. Tech, (PhD), Department of CSE and this work has been submitted for the partial fulfillment of the requirements for the award of Bachelor of Technology degree.

The results embodied in this project report have not been submitted to any other Universities of Institute for the award of Degree.

Ramya S	Reg no: 164G1A0575
Rekha G	Reg no: 164G1A0579
Ramya Sai N	Reg no: 164G1A0576
Sai Kiran K	Reg no: 164G1A0586
Rama Mohan Chowdary K	Reg no: 164G1A0574

CONTENTS

ABSTRACT

Everyday billions of short texts are generated in an enormous volume in the form of search queries, news titles, tags, chat bots, social media posts etc. Understanding short texts retrieval, classification and processing become a very difficult task. Clustering short texts (such as news titles) by their meaning is a challenging task. The semantic hashing approach encodes the meaning of a text into a compact binary code. Thus, to tell if two texts have similar meanings, we only need to check if they have similar codes. The encoding is created by a deep neural network, which is trained on texts. To cluster short texts by their meanings, we propose to add more semantic signals to short texts. Specifically, for each term in a short text, we obtain its concepts and co-occurring terms from a database to enrich the short text. Furthermore, we introduce a simplified deep learning network for semantic hashing.

Many approaches have been proposed to facilitate short text understanding by enriching the short text. These approaches are based on semantic hashing models.

Our approach is based on semantic network for enriching a short text. We present a novel mechanism to semantically enrich short texts with both concepts and co-occurring terms.

KEY WORDS: Short text, semantic enrichment, semantic hashing, deep neural network, clustering.

LIST OF FIGURES

LIST OF SCREENS

LIST OF ABBREVIATIONS

IR	Information Retrieval
SRS	Software Requirement Specification
HTML	Hyper Text Mark-up Language
CSS	Cascading Style Sheets
JVM	Java Virtual Machine
ODBC	Open Database Connectivity
SQL	Structured Query Language
API	Application Program Interface
JDBC	Java Database Connectivity
JDK	Java Development Kit
IDE	Integrated Development Environment
Java EE	Java Enterprise Edition
MSI	Microsoft Installer
UML	Unified Modeling Language

CHAPTER: 1

INTRODUCTION

1.1 Motivation

A wide range of applications handle short texts. For example, news recommendation system needs to process the news titles which may be not strictly syntactical; in web search, queries consist of a very small number of keywords. Short texts introduce new challenges to many text related tasks including information retrieval (IR), classification, and clustering. Unlike long documents, two short texts that have similar meaning do not necessarily share many words. For example, the meanings of "upcoming apple products" and "new iphone and ipad" are closely related, but they share no common words. The lack of sufficient statistical information leads to difficulties in effectively measuring similarity, and as a result, many existing text analytics algorithms do not apply to short texts directly. More importantly, the lack of statistical information also means problems that can be safely ignored when we handle long documents become critical for short texts. The word "apple" gives rise to different meanings in "apple product" and "apple tree". Due to the scarcity of contextual information, these ambiguous words make short texts hard to understand by machines.

1.2 Problem Definition

Short texts introduce new challenges to many text related tasks including information retrieval (IR), classification, and clustering, two short texts that have similar meaning For example, the meanings of "upcoming apple products" and "new iphone and ipad" are closely related, but they share no common words.

The word "apple" gives rise to different meanings in "apple product" and "apple tree". Due to the scarcity of contextual information, these ambiguous words make short texts hard to understand by machines.

1.3 Objective of the Project

The objective of the project is to facilitate short text understanding by enriching the short text representations.

CHAPTER: 2

LITERATURE SURVEY

2.1 Introduction

Title: Principles of hash-based text retrieval

Abstract

Hash-based similarity search reduces a continuous similarity relation to the binary concept "similar or not similar": two feature vectors are considered as similar if they are mapped on the same hash key. From its runtime performance this principle is unequaled-while being unaffected by dimensionality concerns at the same time. Similarity hashing is applied with great success for near similarity search in large document collections, and it is considered as a key technology for near-duplicate detection and plagiarism analysis. These papers reveals the design principles behind hash-based search methods and presents them in a unified way. We introduce new stress statistics that are suited to analyze the performance of hash-based search methods, and we explain the rationale of their effectiveness. Based on these insights, we show how optimum hash functions for similarity search can be derived. We also present new results of a comparative study between different hash-based search methods.

Title: Probase: A probabilistic taxonomy for text understanding

Abstract

Knowledge is indispensable to understanding. The ongoing information explosion highlights the need to enable machines to better understand electronic text in human language. Much work has been devoted to creating universal ontology's or taxonomies for this purpose. However, none of the existing ontology's has the needed depth and breadth for “universal understanding”. In this paper, we present a universal, probabilistic taxonomy that is more comprehensive than any existing ones. It contains 2.7 million concepts harnessed automatically from a corpus of 1.68 billion web pages. Unlike traditional taxonomies that treat knowledge as black and white, it uses probabilities to model inconsistent, ambiguous and uncertain information it contains. We present details of how the taxonomy is constructed, its probabilistic modeling, and its potential applications in text understanding.

In this paper, we presented a framework which automatically inferences an open-domain, probabilistic taxonomy from the entire web. This taxonomy, to the best of our knowledge, is currently the largest and the most comprehensive in terms of the number of concepts included. Its probabilistic model allows the integration of both precise and ambiguous knowledge and even tolerates inconsistencies and errors which are common on the Web. More importantly, this model enables probabilistic inference between concepts and instances which will benefit a wide range of applications that require text understanding.

Title: A web based kernel function for measuring the similarity of short text Snippets

Abstract

Determining the similarity of short text snippets, such as search queries, works poorly with traditional document similarity measures (e.g., cosine), since there are often few, if any, terms in common between two short text snippets. We address this problem by introducing a novel method for measuring the similarity between short text snippets (even those without any overlapping terms) by leveraging web search results to provide greater context for the short texts. In this paper, we define such a similarity kernel function, mathematically analyze some of its properties, and provide examples of its efficacy. We also show the use of this kernel function in a large-scale system for suggesting related queries to search engine users.

We have presented a new kernel function for measuring the semantic similarity between pairs of short text snippets. We have shown, both anecdotally and in a human-evaluated query suggestion system, that this kernel is an effective measure of similarity for short texts, and works well even when the short texts being considered have no common terms. Moreover, we have also provided a theoretical analysis of the kernel function that shows that it is well-suited for use with the web. There are several lines of future work that this kernel lays the foundation for. The first is improvement in the generation of query expansions with the goal of improving the match score for the kernel function. The second is the incorporation of this kernel into other kernel-based machine learning methods to determine its ability to provide improvement in tasks such as classification and clustering of text.

Title: Semantic Enrichment of Short Texts through Conducive Formula

Abstract

Web search queries are regularly short and equivocal. Accurate current categorization of user queries regard increased efficiency, quickness, and return potential in general-purpose

web search systems. To characterize these queries into certain target classes is a difficult task. Such categorization becomes critical if the system is to return results not just from a general web collection but from topic-specific databases. Semantic analysis is used for the semantic word collections. For each word co-occurring and conceptualized terms are defined. Random forest algorithm is used for semantic hashing of the queries. The efficiency of retrieving the result is more efficient than the traditional algorithms.

Title: Using deep learning for short text understanding

Abstract

Classifying short texts to one category or clustering semantically related texts is challenging, and the importance of both is growing due to the rise of microblogging platforms, digital news feeds, and the like. We can accomplish this classifying and clustering with the help of a deep neural network which produces compact binary representations of a short text, and can assign the same category to texts that have similar binary representations. But problems arise when there is little contextual information on the short texts, which makes it difficult for the deep neural network to produce similar binary codes for semantically related texts. We propose to address this issue using semantic enrichment. This is accomplished by taking the nouns, and verbs used in the short texts and generating the concepts and co-occurring words with the help of those terms. The nouns are used to generate concepts within the given short text, whereas the verbs are used to prune the ambiguous context (if any) present in the text. The enriched text then goes through a deep neural network to produce a prediction label for that short text representing it’s category.

2.2 Existing System

- Many applications have been proposed to facilitate short text understanding by enriching the short text.
- The lack of sufficient statistical information leads to difficulties in effectively measuring similarity, and as a result, many existing text analytics algorithms do not apply to short texts directly.
- Semantic hashing models are used for understanding short text.

2.2 Disadvantages of Existing System

- Search-based approaches may work well for so-called head queries, but for tail or unpopular queries, it is very likely that some of the top search results are irrelevant, which means the enriched short text is likely to contain a lot of noise.
- On the other hand, methods based on external resources are constrained by the coverage of these resources. Take Word Net for example, Word Net does not contain information for proper nouns, which prevents it to understand entities such as "USA" or "IBM."
- For ordinary words such as "cat", Word Net contains detailed information about its various senses. However, much of the knowledge is of linguistic value, and is rarely evoked in daily usage. For example, the sense of "cat" as gossip or woman is rarely encountered.
- Unfortunately, Word Net does not weight senses based on their usage, and these rarely used senses often give rise to misinterpretation of short texts. In summary, without knowing the distribution of the senses, it is difficult to build an inferencing mechanism to choose appropriate senses for a word in a context.

2.3 Proposed System

- We present a novel mechanism to semantically enrich short texts with both concepts and co-occurring terms.
- For this we use deep neural networks model for semantic hashing.
- We show significant improvements over existing approaches, which confirm that concepts and co-occurring terms effectively enrich short texts, and enable better understanding of them.

CHAPTER: 3

FEASIBILITY STUDY

3.1 Introduction

Feasibility analysis reduces the development risks. The major areas are considered in feasibility analysis are as follows.

The feasibility study concerns with the consideration made to verify whether the system is to be developed in all terms. Once an idea to develop software is put forward the question that arises first will pertain to the feasibility aspects. It involves developing and understanding of the selected program. There are different aspects in the feasibility study.

1. Technical Feasibility
2. Economical Feasibility
3. Social Feasibility

3.1.1 Technical Feasibility

Technical feasibility study is the complete study of the project in terms of input, processes, output, fields, programs and procedures. It is a very effective tool for long term planning and trouble shooting. The technical feasibility study should most essentially support the financial information of an organization.

Understand the different technologies involved in the proposed system before commencing the project we have to be very clear about what are the technologies that are to be required for the development of the new system.

3.1.2 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.

3.1.3 Social Feasibility

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

CHAPTER: 4
REQUIREMENTS

4.1 Introduction

Software Requirement Specification is the starting point of the software developing activity. As systems now are more complex it became evident that the goal of the entire system cannot be easily comprehended. Hence the need for the requirement phase arose. The software is initiated by the client needs. The software requirements specification lays out functional and non-functional requirements, and it may include a set of use cases that describe user interactions that the software must provide to the user for perfect interaction. The SRS is the means of translating the ideas of the minds of the clients (the input) into a formal document (the output of the requirement phase).

To run our project on a specified computer the following are the Software and hardware requirement.

4.1.1 Software Requirements

• Operating system	:	Windows 7/8/9.
• Coding Language	:	JAVA/J2EE
• Tool	:	Netbeans 7.2.1
• Scripting Languages	:	HTML,CSS
• Database	:	MYSQL

4.1.2 Hardware Requirements

• Processor	:	Core i3
• Hard Disk	:	120 GB.
• RAM	:	3GB.

CHAPTER: 5

ANALYSIS

5.1 Introduction

Java technology is both a programming language and a platform. The Java programming language is a high-level language that can be characterized by all of the following buzzwords:

- Simple
- Architecture neutral
- Object oriented
- Portable
- Distributed
- High performance
- Interpreted
- Multithreaded
- Robust
- Dynamic
- Secure

With most programming languages, you either compile or interpret a program so that you can run it on your computer. The Java programming language is unusual in that a program is both compiled and interpreted. With the compiler, first you translate a program into an intermediate language called *Java byte codes* —the platform-independent codes interpreted by the interpreter on the Java platform. The interpreter parses and runs each Java byte code instruction on the computer. Compilation happens just once; interpretation occurs each time the program is executed. The following figure illustrates how this works.

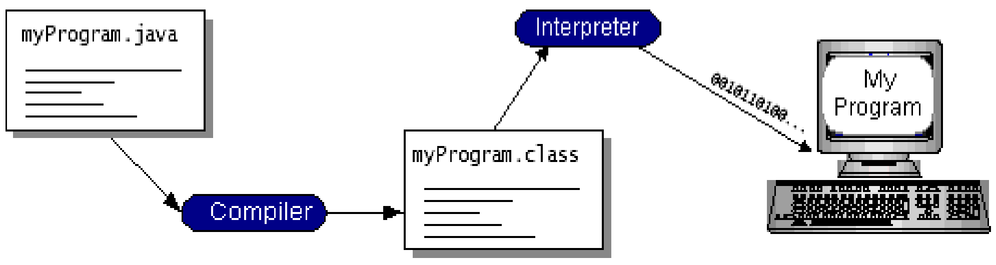
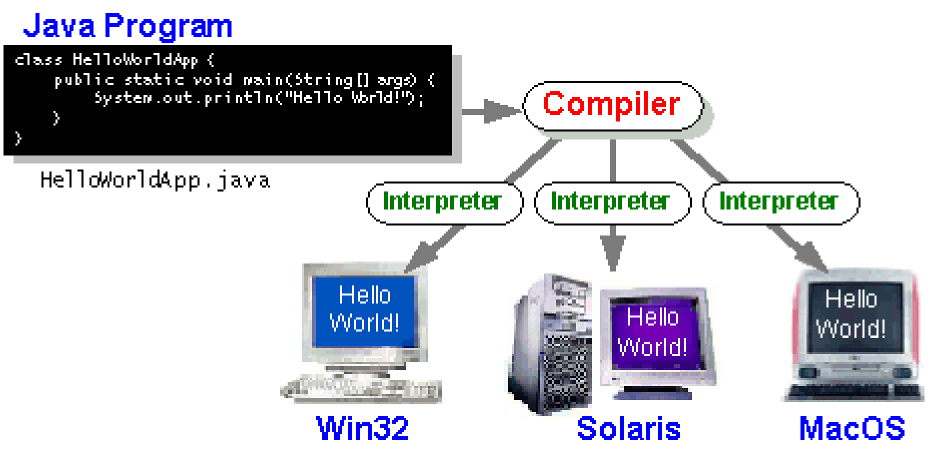


Fig 5.1.1: Figure illustrating how Java works

You can think of Java byte codes as the machine code instructions for the *Java Virtual Machine* (Java VM). Every Java interpreter, whether it's a development tool or a Web browser that can run applets, is an implementation of the Java VM. Java byte codes help make “write once, run anywhere” possible. You can compile your program into byte codes on any platform that has a Java compiler. The byte codes can then be run on any implementation of the Java VM. That means that as long as a computer has a Java VM, the same program written in the Java programming language can run on Windows 2000, a Solaris workstation, or on an iMac.



```
class HelloWorldApp {
    public static void main(String[] args) {
        System.out.println("Hello World!");
    }
}
```

Fig 5.1.2: Illustrating write once, run anywhere

ODBC

Microsoft Open Database Connectivity (ODBC) is a standard programming interface for application developers and database systems providers. Before ODBC became a *de facto* standard for Windows programs to interface with database systems, programmers had to use proprietary languages for each database they wanted to connect to. Now, ODBC has made the choice of the database system almost irrelevant from a coding perspective, which is as it should be. Application developers have much more important things to worry about than the syntax that is needed to port their program from one database to another when business needs suddenly change.

Through the ODBC Administrator in Control Panel, you can specify the particular database that is associated with a data source that an ODBC application program is written to use. Think of an ODBC data source as a door with a name on it. Each door will lead you to a particular database. For example, the data source named Sales Figures might be a SQL Server database, whereas the Accounts Payable data source could refer to an Access database. The physical database referred to by a data source can reside anywhere on the LAN.

The ODBC system files are not installed on your system by Windows 95. Rather, they are installed when you setup a separate database application, such as SQL Server Client or Visual Basic 4.0. When the ODBC icon is installed in Control Panel, it uses a file called ODBCINST.DLL. It is also possible to administer your ODBC data sources through a stand-alone program called ODBCADM.EXE. There is a 16-bit and a 32-bit version of this program and each maintains a separate list of ODBC data sources. From a programming perspective, the beauty of ODBC is that the application can be written to use the same set of function calls to interface with any data source, regardless of the database vendor. The source code of the application doesn’t change whether it talks to Oracle or SQL Server. We only mention these two as an example. There are ODBC drivers available for several dozen popular database systems. Even Excel spreadsheets and plain text files can be turned into data sources. The operating system uses the Registry information written by ODBC Administrator to determine which low-level ODBC drivers are needed to talk to the data source (such as the interface to Oracle or SQL Server). The loading of the ODBC drivers is transparent to the ODBC application program. In a client/server environment, the ODBC API even handles many of the network issues for the application programmer.

The advantages of this scheme are so numerous that you are probably thinking there must be some catch. The only disadvantage of ODBC is that it isn’t as efficient as talking directly to the native database interface. ODBC has had many detractors make the charge that it is too slow. Microsoft has always claimed that the critical factor in performance is the quality of the driver software that is used. In our humble opinion, this is true. The availability of good ODBC drivers has improved a great deal recently. And anyway, the criticism about performance is somewhat analogous to those who said that compilers would never match the speed of pure assembly language. Maybe not, but the compiler (or ODBC) gives you the opportunity to write cleaner programs, which means you finish sooner. Meanwhile, computers get faster every year.

JDBC

In an effort to set an independent database standard API for Java; Sun Microsystems developed Java Database Connectivity, or JDBC. This consistent interface is achieved through the use of "plug-in" database connectivity modules, or *drivers*. If a database vendor wishes to have JDBC support, he or she must provide the driver for each platform that the database and Java run on.

To gain a wider acceptance of JDBC, Sun based JDBC's framework on ODBC. As you discovered earlier in this chapter, ODBC has widespread support on a variety of platforms. Basing JDBC on ODBC will allow vendors to bring JDBC drivers to market much faster than developing a completely new connectivity solution.

JDBC was announced in March of 1996. It was released for a 90 day public review that ended June 8, 1996. Because of user input, the final JDBC v1.0 specification was released soon after.

The remainder of this section will cover enough information about JDBC for you to know what it is about and how to use it effectively. This is by no means a complete overview of JDBC. That would fill an entire book.

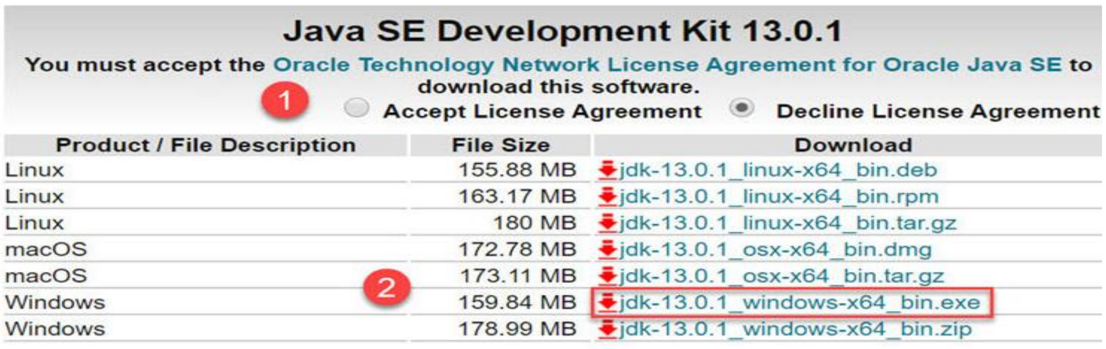
5.2 JAVA Installation

Following are steps to install Java in Windows.

STEP 1: Go to link**,** Click on Download JDK.

STEP 2: Next,

1. Accept License Agreement.
2. Download latest Java JDK for your version (32 or 64 bit) of java for Windows.



Java SE Development Kit 13.0.1

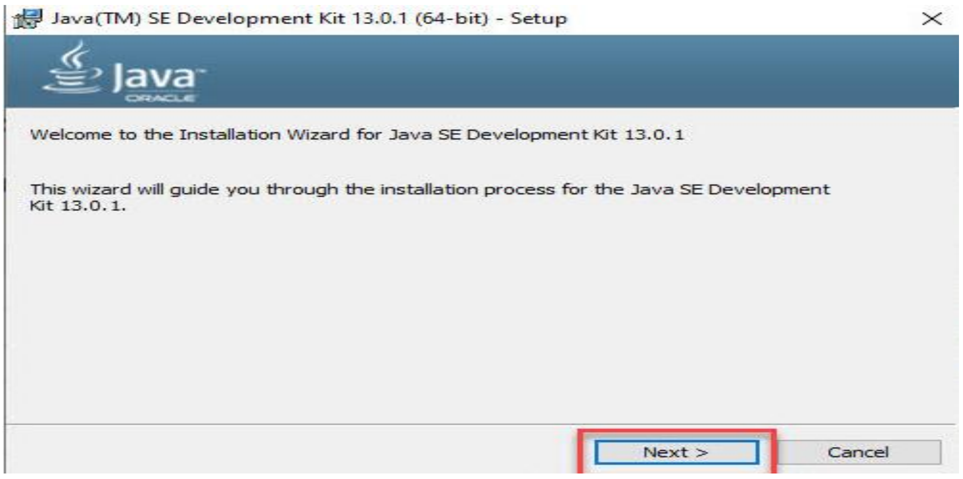
You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.

1 ○ Accept License Agreement [✓ ● Decline License Agreement]

Product / File Description	File Size	Download
Linux	155.88 MB	jdk-13.0.1_linux-x64_bin.deb
Linux	163.17 MB	jdk-13.0.1_linux-x64_bin.rpm
Linux	180 MB	jdk-13.0.1_linux-x64_bin.tar.gz
macOS	172.78 MB	jdk-13.0.1_osx-x64_bin.dmg
macOS	173.11 MB	jdk-13.0.1_osx-x64_bin.tar.gz
Windows	159.84 MB	jdk-13.0.1_windows-x64_bin.exe
Windows	178.99 MB	jdk-13.0.1_windows-x64_bin.zip

Fig 5.2.1: Download window

STEP 3: Once the download is complete, run the exe for install JDK. Click Next.



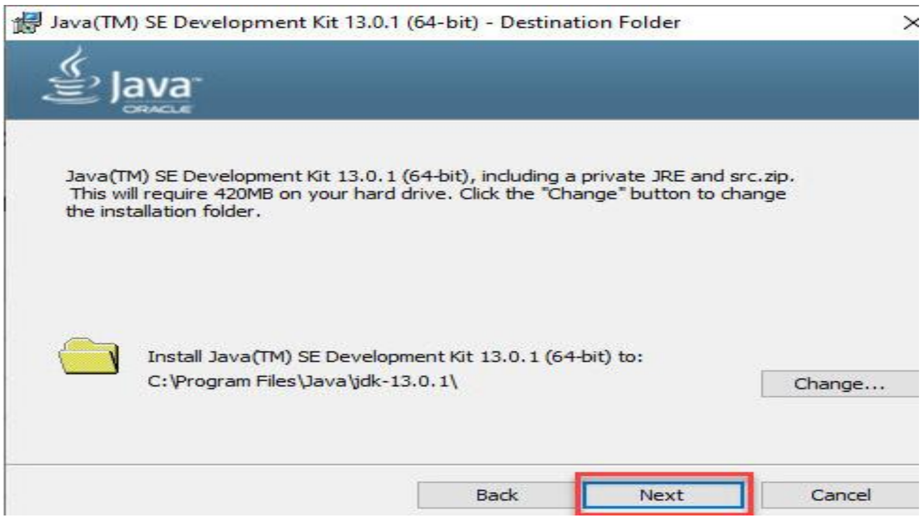
Java(TM) SE Development Kit 13.0.1 (64-bit) - Setup

Welcome to the Installation Wizard for Java SE Development Kit 13.0.1

This wizard will guide you through the installation process for the Java SE Development Kit 13.0.1.

Fig 5.2.2: Installation window

STEP 4: Select the PATH for Java installation and click next.



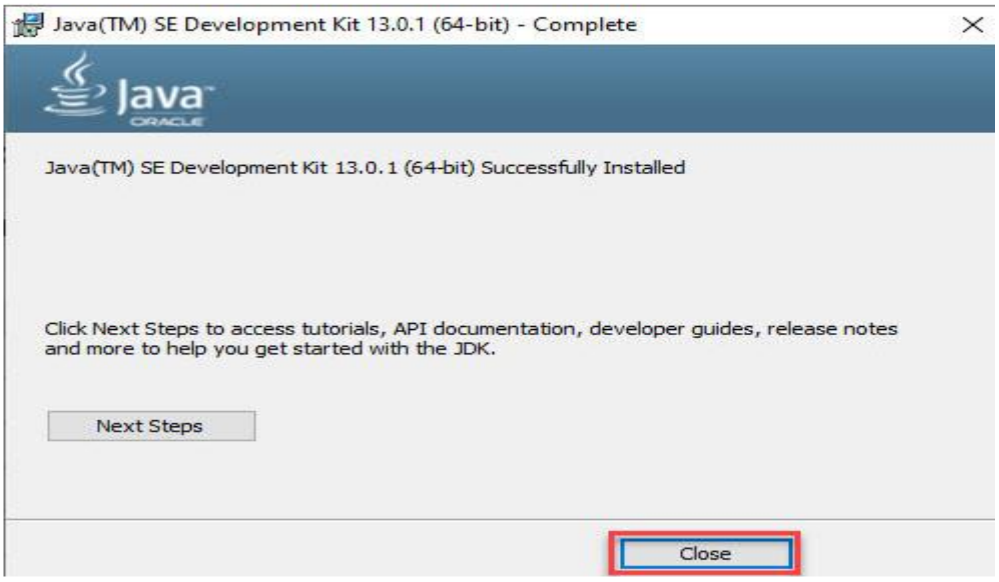
Java(TM) SE Development Kit 13.0.1 (64-bit) - Destination Folder

Java(TM) SE Development Kit 13.0.1 (64-bit), including a private JRE and src.zip. This will require 420MB on your hard drive. Click the "Change" button to change the installation folder.

Install Java(TM) SE Development Kit 13.0.1 (64-bit) to:
C:\Program Files\Java\jdk-13.0.1\

Fig 5.2.3: Selecting path for Java installation

STEP 5: Once installation is complete click Close.



Java(TM) SE Development Kit 13.0.1 (64-bit) - Complete

Java(TM) SE Development Kit 13.0.1 (64-bit) Successfully Installed

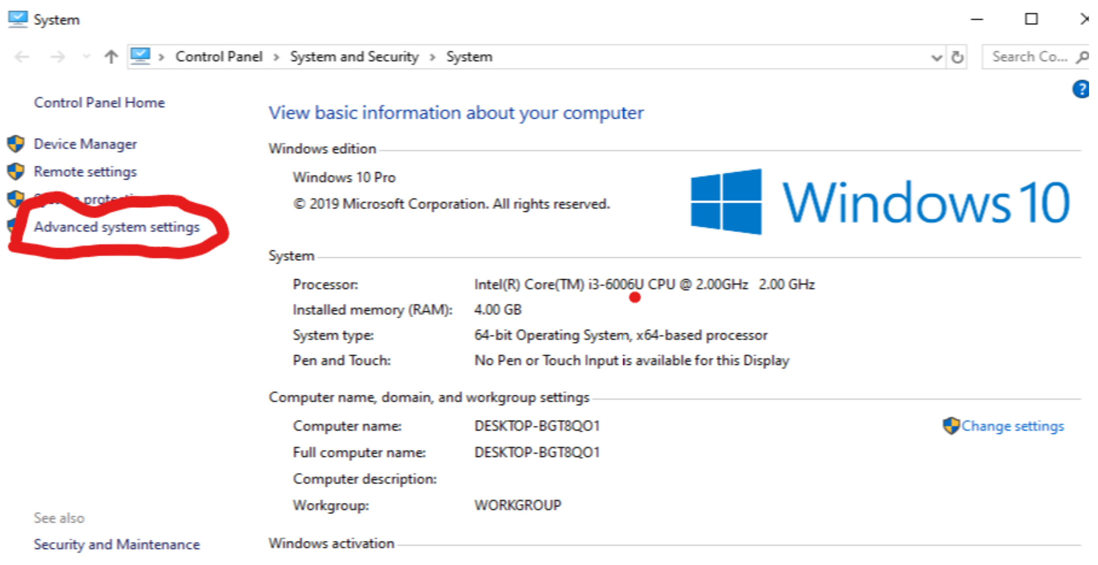
Click Next Steps to access tutorials, API documentation, developer guides, release notes and more to help you get started with the JDK.

Fig 5.2.4: Installation completed window

5.2.1 Setting Environment Variables in Java: Path and Classpath

STEP 1: Right Click on the My Computer and Select the properties.

STEP 2: Click on advanced system settings.



View basic information about your computer

Windows edition

Windows 10 Pro

© 2019 Microsoft Corporation. All rights reserved.

System

Processor:	Intel(R) Core(TM) i3-6006U CPU @ 2.00GHz 2.00 GHz
Installed memory (RAM):	4.00 GB
System type:	64-bit Operating System, x64-based processor
Pen and Touch:	No Pen or Touch Input is available for this Display

Computer name, domain, and workgroup settings

Computer name:	DESKTOP-BGT8QO1
Full computer name:	DESKTOP-BGT8QO1
Computer description:	
Workgroup:	WORKGROUP

Windows activation

Fig 5.2.1.1: Window to select Advanced system settings

STEP 3: Click on Environment Variables.

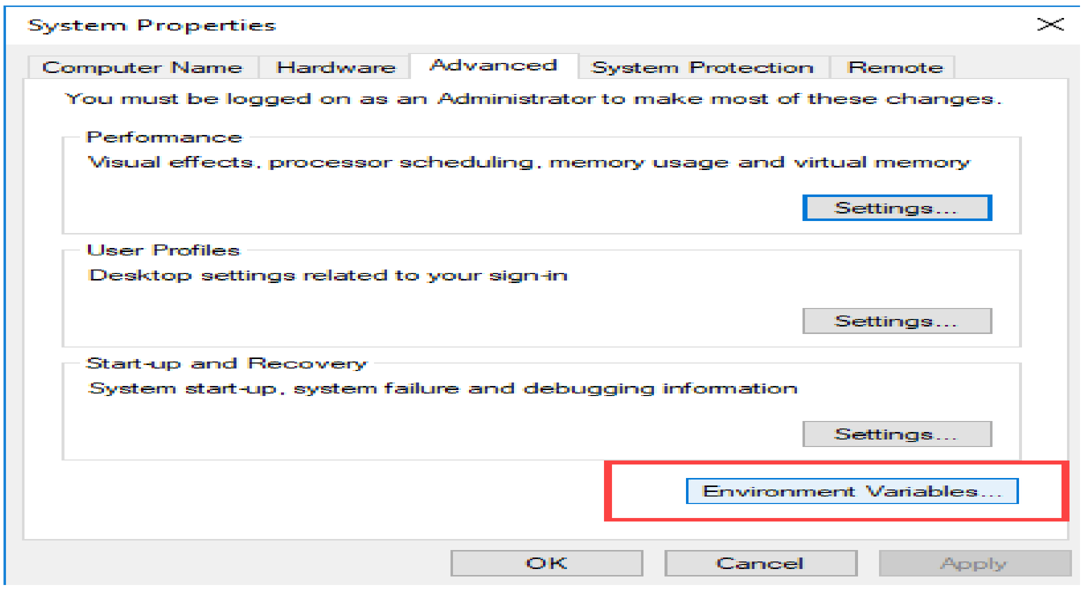
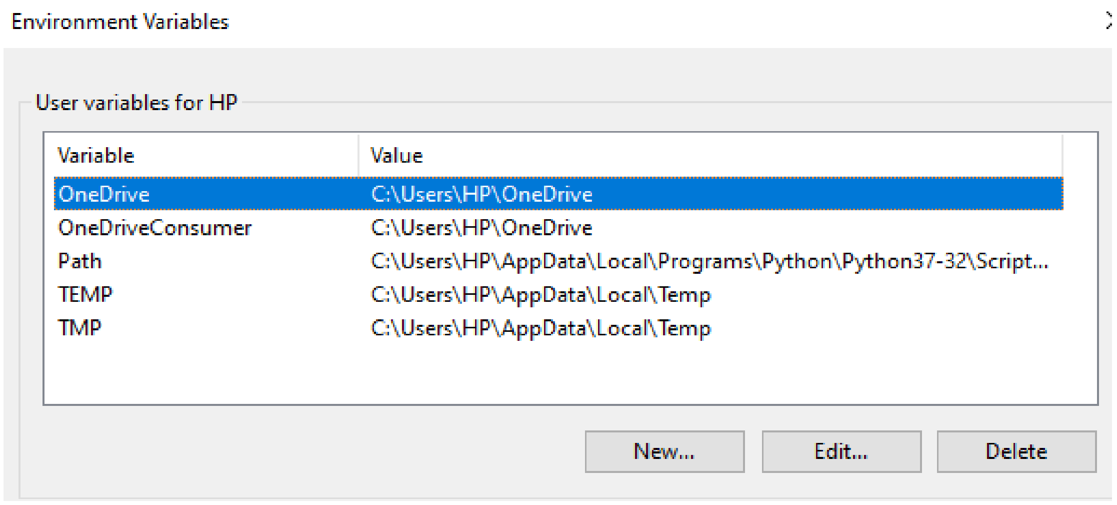

System Properties
Computer Name
Hardware
Advanced
System Protection
Remote
You must be logged on as an Administrator to make most of these changes.
Performance
Visual effects, processor scheduling, memory usage and virtual memory
Settings...
User Profiles
Desktop settings related to your sign-in
Settings...
Start-up and Recovery
System start-up, system failure and debugging information
Settings...
Environment Variables...
OK
Cancel
Apply

Fig 5.2.1.2: Window to select Environment Variables

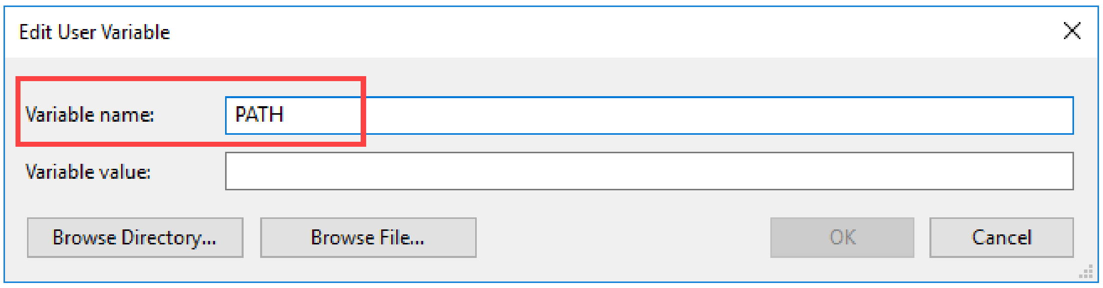
STEP 4: Click on new Button of User variables.



Variable	Value
OneDrive	C:\Users\HP\OneDrive
OneDriveConsumer	C:\Users\HP\OneDrive
Path	C:\Users\HP\AppData\Local\Programs\Python\Python37-32\Script...
TEMP	C:\Users\HP\AppData\Local\Temp
TMP	C:\Users\HP\AppData\Local\Temp

Fig 5.2.1.3: Window to create User variables

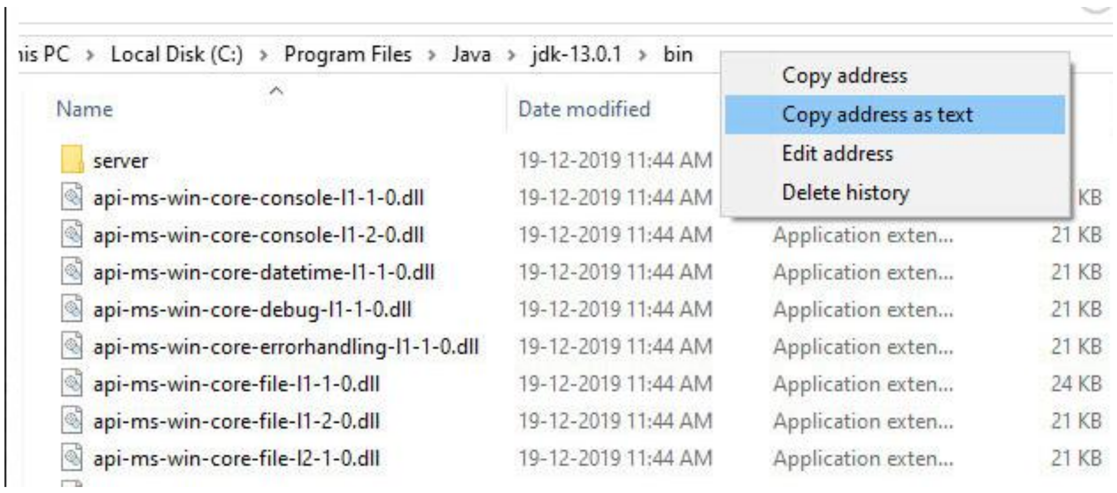
STEP 5: Type PATH in the Variable name.



Variable name:	PATH
Variable value:	

Fig 5.2.1.4: Window to set Variable name

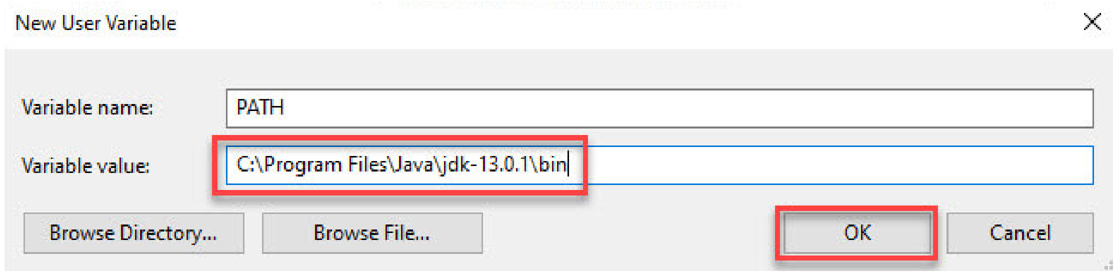
STEP 6: Copy the path of bin folder which is installed in JDK folder.



Name	Date modified	Type	Size
server	19-12-2019 11:44 AM		
api-ms-win-core-console-l1-1-0.dll	19-12-2019 11:44 AM		KB
api-ms-win-core-console-l1-2-0.dll	19-12-2019 11:44 AM	Application exten...	21 KB
api-ms-win-core-datetime-l1-1-0.dll	19-12-2019 11:44 AM	Application exten...	21 KB
api-ms-win-core-debug-l1-1-0.dll	19-12-2019 11:44 AM	Application exten...	21 KB
api-ms-win-core-errorhandling-l1-1-0.dll	19-12-2019 11:44 AM	Application exten...	21 KB
api-ms-win-core-file-l1-1-0.dll	19-12-2019 11:44 AM	Application exten...	24 KB
api-ms-win-core-file-l1-2-0.dll	19-12-2019 11:44 AM	Application exten...	21 KB
api-ms-win-core-file-l2-1-0.dll	19-12-2019 11:44 AM	Application exten...	21 KB

Fig 5.2.1.5: Window to copy the path

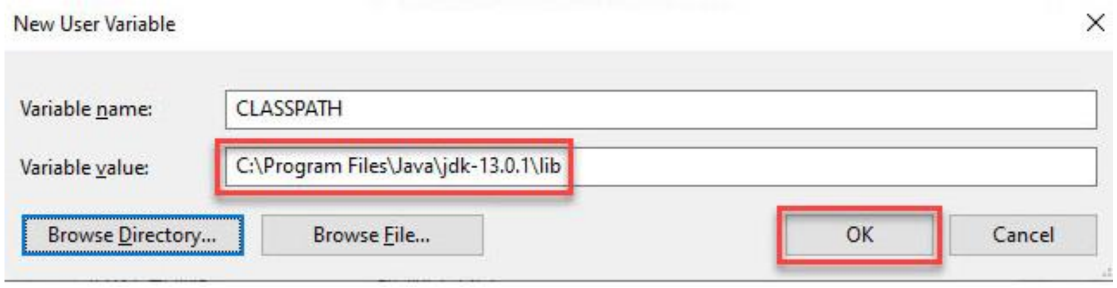
STEP 7: Paste Path of bin folder in Variable value and click on OK Button.



Variable name: PATH
Variable value: C:\Program Files\Java\jdk-13.0.1\bin

Fig 5.2.1.6: Path setting window

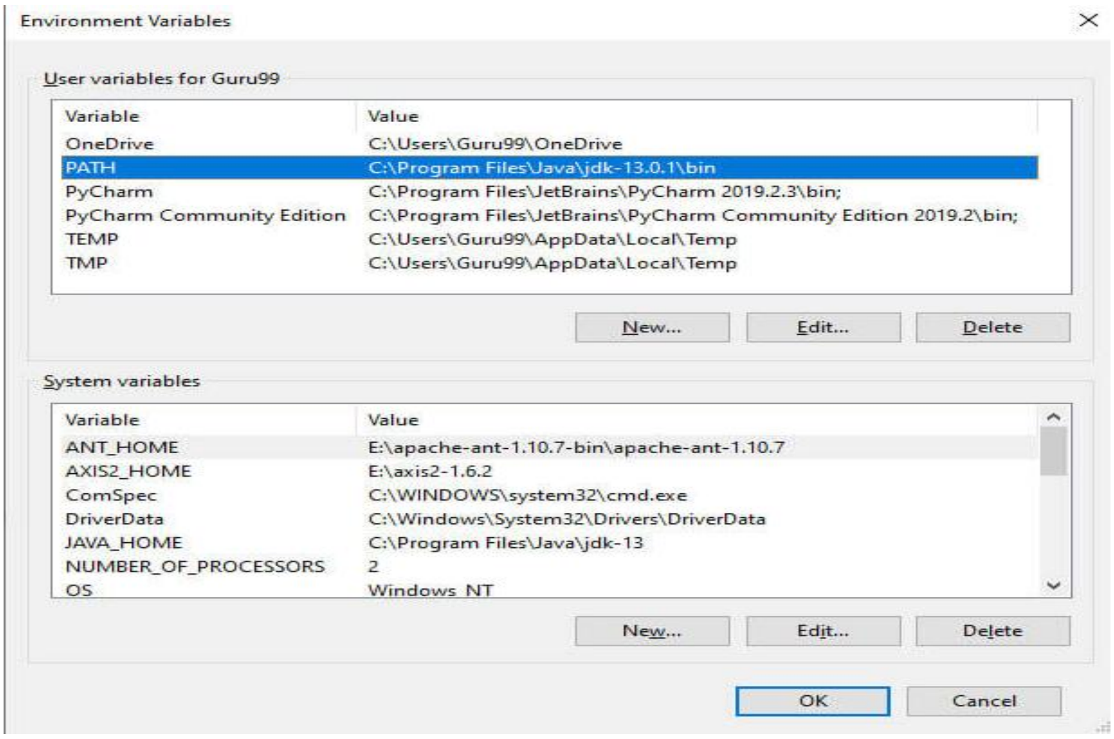
STEP 8: You can follow a similar process to set CLASSPATH.



Variable name: CLASSPATH
Variable value: C:\Program Files\Java\jdk-13.0.1\lib

Fig 5.2.1.7: Classpath setting window

STEP 9: Click on OK button.


Environment Variables

User variables for Guru99

Variable	Value
OneDrive	C:\Users\Guru99\OneDrive
PATH	C:\Program Files\Java\jdk-13.0.1\bin
PyCharm	C:\Program Files\JetBrains\PyCharm 2019.2.3\bin;
PyCharm Community Edition	C:\Program Files\JetBrains\PyCharm Community Edition 2019.2\bin;
TEMP	C:\Users\Guru99\AppData\Local\Temp
TMP	C:\Users\Guru99\AppData\Local\Temp

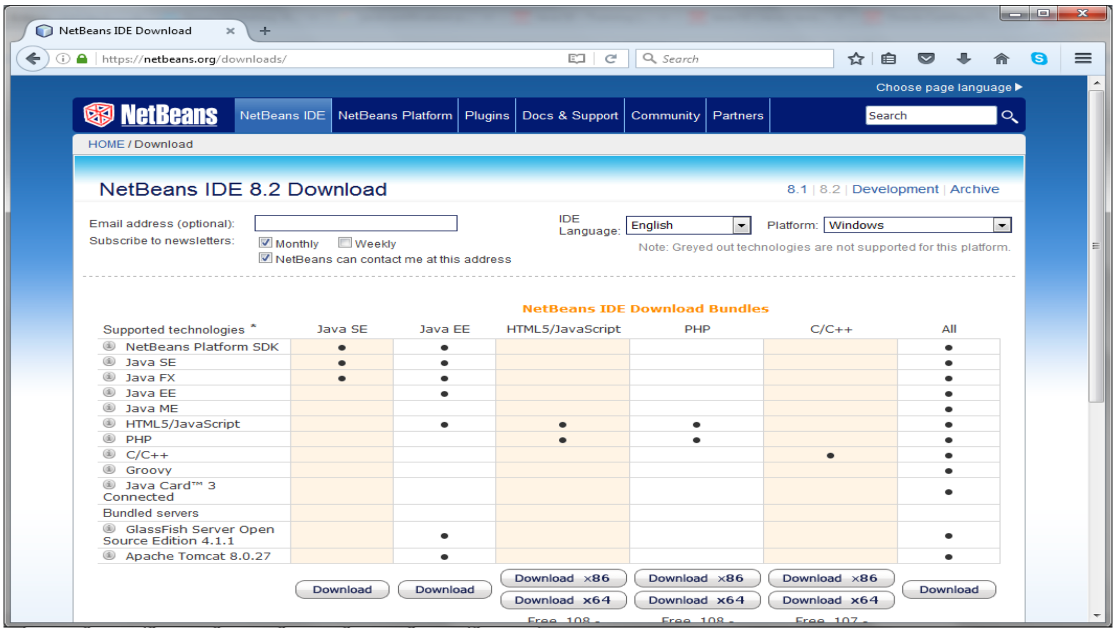
System variables

Variable	Value
ANT_HOME	E:\apache-ant-1.10.7-bin\apache-ant-1.10.7
AXIS2_HOME	E:\axis2-1.6.2
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
JAVA_HOME	C:\Program Files\Java\jdk-13
NUMBER_OF_PROCESSORS	2
OS	Windows_NT

Fig 5.2.1.8: Path setting completed window

5.3 NetBeans Installation

STEP 1: Go to https://netbeans.org/downloads to download the latest version of NetBeans IDE. You will see the following page:



NetBeans IDE 8.2 Download

Email address (optional):

Subscribe to newsletters: [✓ Monthly] Weekly

[✓ NetBeans can contact me at this address]

IDE Language: English

Platform: Windows

Note: Greyed out technologies are not supported for this platform.

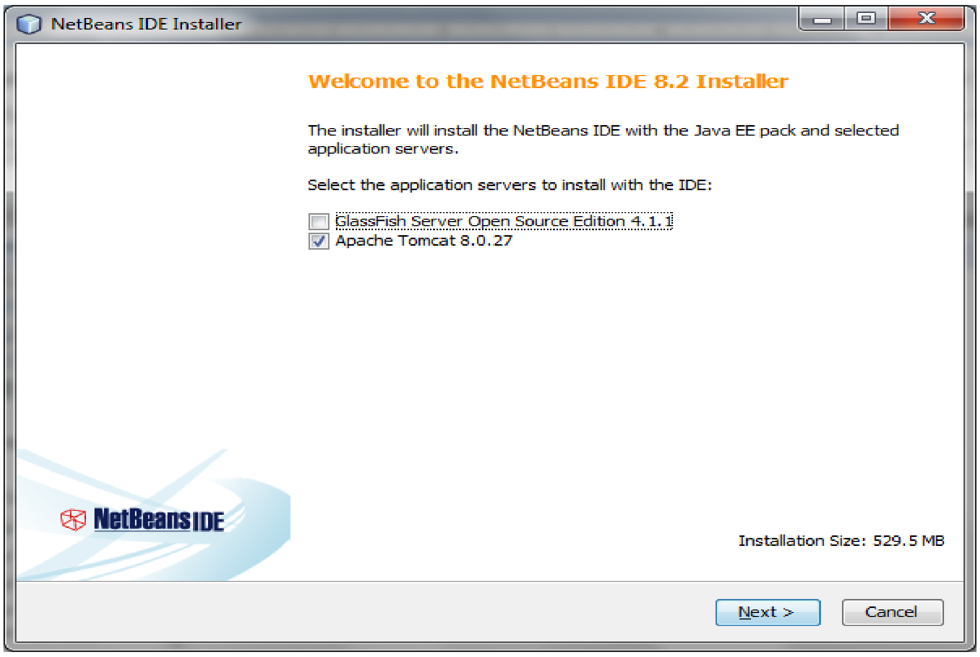
NetBeans IDE Download Bundles

Supported technologies *	Java SE	Java EE	HTML5/JavaScript	PHP	C/C++	All
NetBeans Platform SDK	●	●				●
Java SE	●	●				●
Java FX	●	●				●
Java EE		●				●
Java ME						●
HTML5/JavaScript		●	●	●		●
PHP			●	●		●
C/C++					●	●
Groovy						●
Java Card™ 3 Connected						●
Bundled servers						
GlassFish Server Open Source Edition 4.1.1		●				●
Apache Tomcat 8.0.27		●				●
	Download	Download	Download x86 Download x64	Download x86 Download x64	Download x86 Download x64	Download

Fig 5.3.1: Download window

STEP 2: Click the Download button in the column Java EE to download NetBeans installer for Java EE development. The file name of the installer program is something like netbeans-8.2-javaee-windows.exe.

STEP 3: Click on the installer file to start installing NetBeans IDE. You will be asked to install GlassFish and Apache Tomcat server:


NetBeans IDE Installer

Welcome to the NetBeans IDE 8.2 Installer

The installer will install the NetBeans IDE with the Java EE pack and selected application servers.

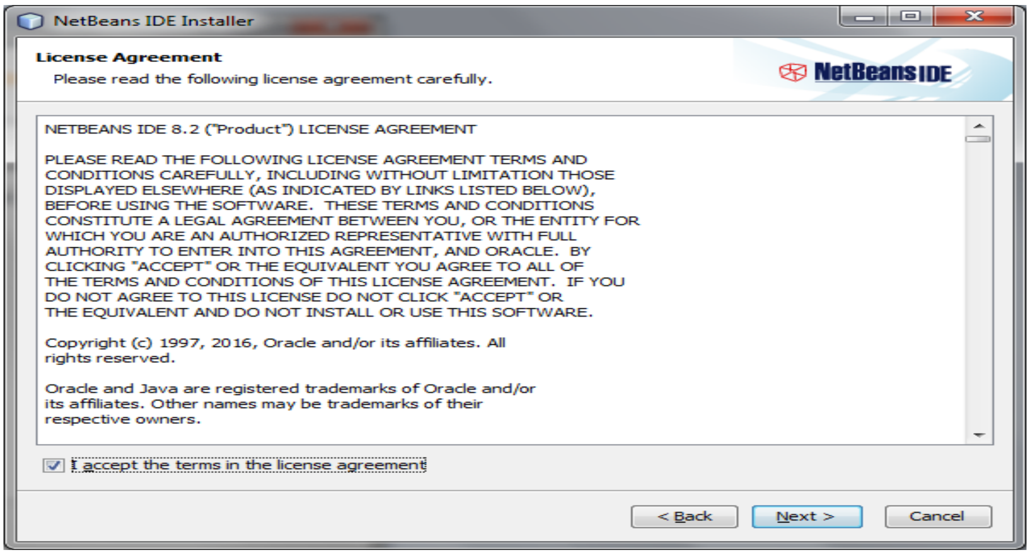
Select the application servers to install with the IDE:

☐ GlassFish Server Open Source Edition 4.1.1
☑ Apache Tomcat 8.0.27

Installation Size: 529.5 MB

Fig 5.3.2: Welcome window

STEP 4: Click Next. In the next screen, check ‘I accept the terms in the license agreement’:


NetBeans IDE Installer

License Agreement
Please read the following license agreement carefully.

NETBEANS IDE 8.2 ("Product") LICENSE AGREEMENT

PLEASE READ THE FOLLOWING LICENSE AGREEMENT TERMS AND CONDITIONS CAREFULLY, INCLUDING WITHOUT LIMITATION THOSE DISPLAYED ELSEWHERE (AS INDICATED BY LINKS LISTED BELOW), BEFORE USING THE SOFTWARE. THESE TERMS AND CONDITIONS CONSTITUTE A LEGAL AGREEMENT BETWEEN YOU, OR THE ENTITY FOR WHICH YOU ARE AN AUTHORIZED REPRESENTATIVE WITH FULL AUTHORITY TO ENTER INTO THIS AGREEMENT, AND ORACLE. BY CLICKING "ACCEPT" OR THE EQUIVALENT YOU AGREE TO ALL OF THE TERMS AND CONDITIONS OF THIS LICENSE AGREEMENT. IF YOU DO NOT AGREE TO THIS LICENSE DO NOT CLICK "ACCEPT" OR THE EQUIVALENT AND DO NOT INSTALL OR USE THIS SOFTWARE.

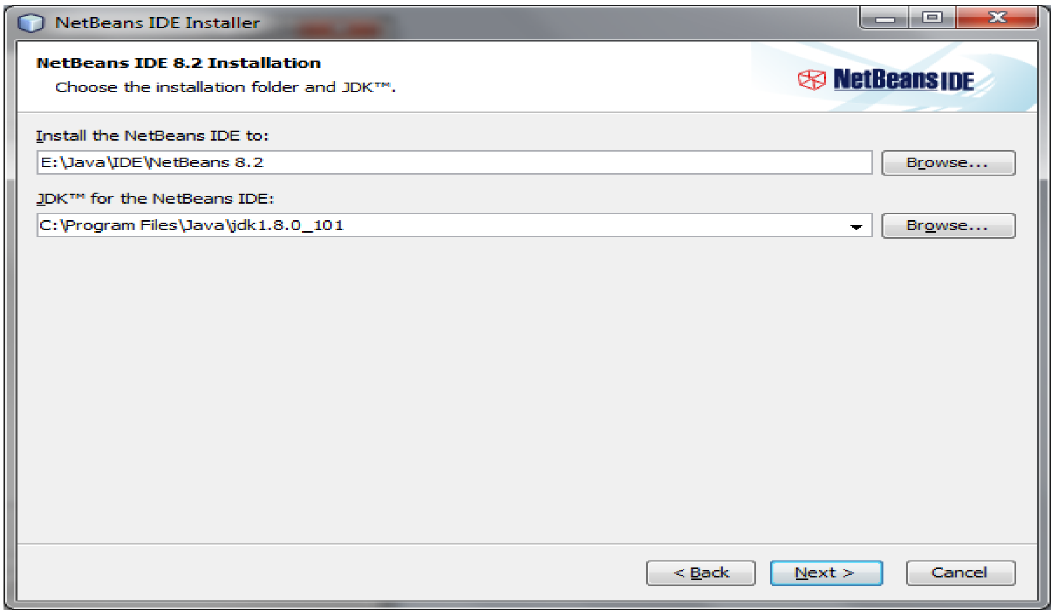
Copyright (c) 1997, 2016, Oracle and/or its affiliates. All rights reserved.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

☑ I accept the terms in the license agreement

Fig 5.3.3: License agreement window

STEP 5: Click Next. In the next screen, choose the installation directory and JDK version for the IDE:



NetBeans IDE Installer

NetBeans IDE 8.2 Installation

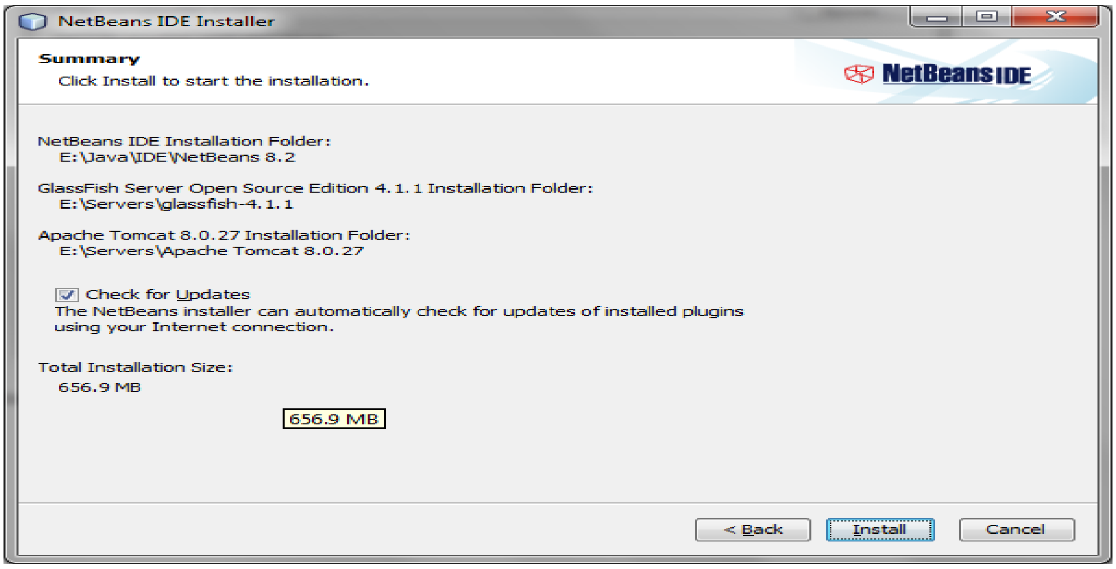
Choose the installation folder and JDK™.

Install the NetBeans IDE to:
E:\Java\IDE\NetBeans 8.2

JDK™ for the NetBeans IDE:
C:\Program Files\Java\jdk1.8.0_101

Fig 5.3.4: Window for choosing installation directory

STEP 6: Click on Install to start installation.



NetBeans IDE Installer

Summary

Click Install to start the installation.

NetBeans IDE Installation Folder:
E:\Java\IDE\NetBeans 8.2

GlassFish Server Open Source Edition 4.1.1 Installation Folder:
E:\Servers\glassfish-4.1.1

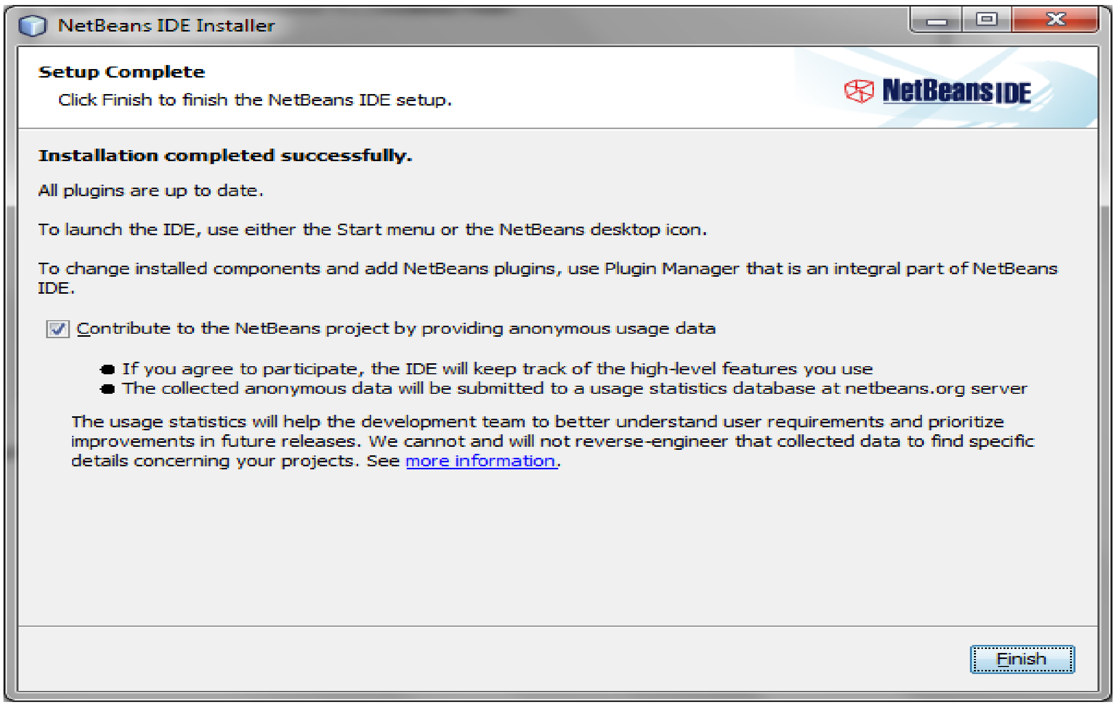
Apache Tomcat 8.0.27 Installation Folder:
E:\Servers\Apache Tomcat 8.0.27

☑ Check for Updates
The NetBeans installer can automatically check for updates of installed plugins using your Internet connection.

Total Installation Size:
656.9 MB

Fig 5.3.5: Installation Window

STEP 7: Wait until the setup complete, click on Finish.



NetBeans IDE Installer

Setup Complete

Click Finish to finish the NetBeans IDE setup.

Installation completed successfully.

All plugins are up to date.

To launch the IDE, use either the Start menu or the NetBeans desktop icon.

To change installed components and add NetBeans plugins, use Plugin Manager that is an integral part of NetBeans IDE.

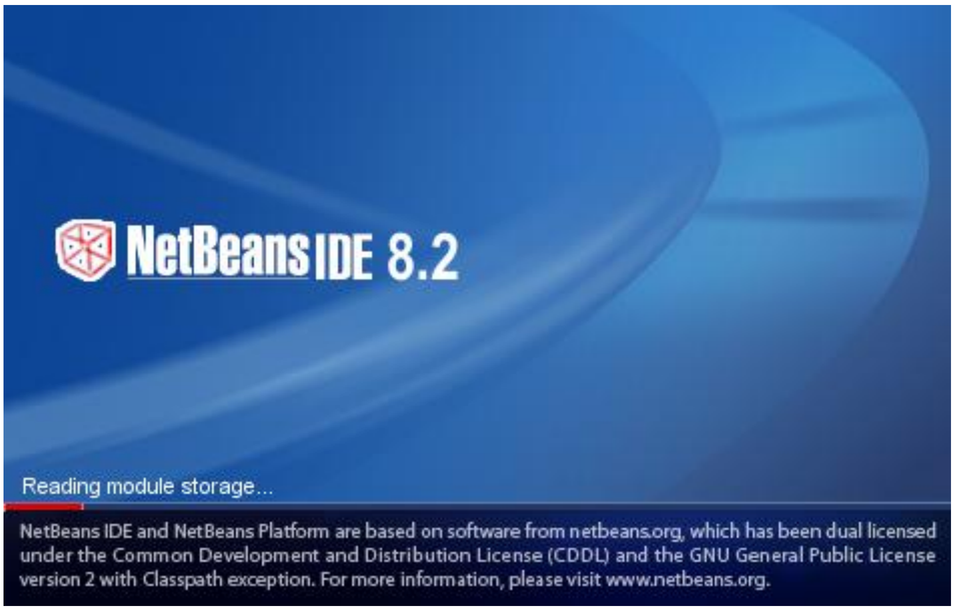
[✓] Contribute to the NetBeans project by providing anonymous usage data

- If you agree to participate, the IDE will keep track of the high-level features you use
- The collected anonymous data will be submitted to a usage statistics database at netbeans.org server

The usage statistics will help the development team to better understand user requirements and prioritize improvements in future releases. We cannot and will not reverse-engineer that collected data to find specific details concerning your projects. See more information.

Fig 5.3.6: Installation completed window

STEP 8: Now you can start NetBeans IDE from the start menu. The splash screen appears:



Reading module storage...

NetBeans IDE and NetBeans Platform are based on software from netbeans.org, which has been dual licensed under the Common Development and Distribution License (CDDL) and the GNU General Public License version 2 with Classpath exception. For more information, please visit www.netbeans.org.

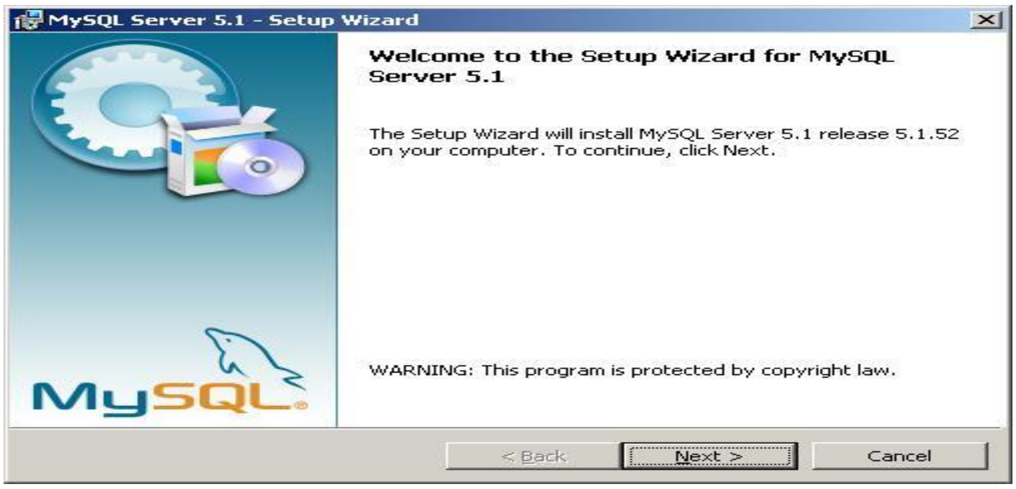
Fig 5.3.7: Splash screen

5.4 MySQL Installation

STEP 1: Download MySQL for Windows from the below link.

(**https://dev.mysql.com/downloads/mysql/5.1.html**)

STEP 2: Install MySQL. Double click the MSI installer to start installing MySQL. The following screen will appear, click on Next to continue.



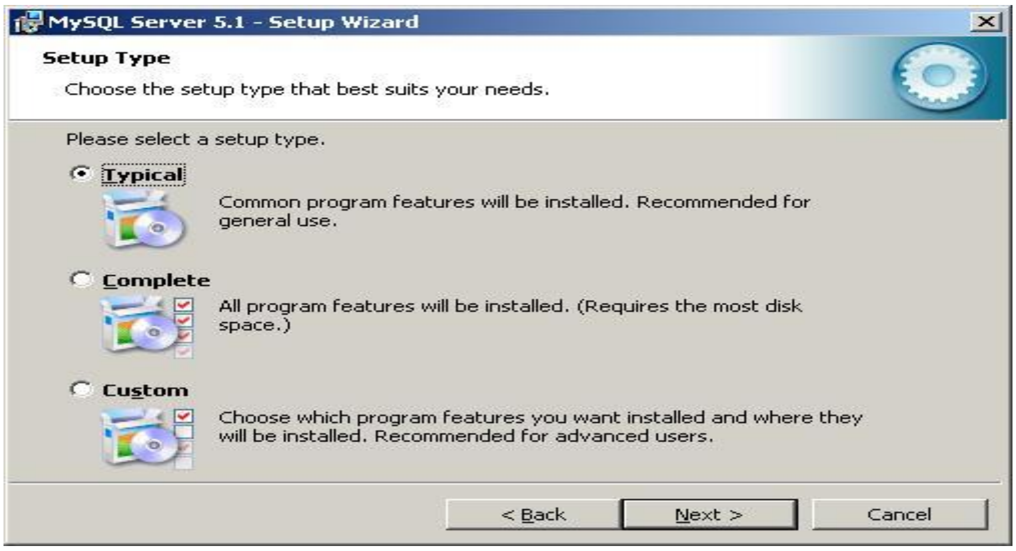
Welcome to the Setup Wizard for MySQL Server 5.1

The Setup Wizard will install MySQL Server 5.1 release 5.1.52 on your computer. To continue, click Next.

WARNING: This program is protected by copyright law.

Fig 5.4.1: Welcome window

STEP 3: Select a setup type - Typical, Complete, Custom. Select Typical and click Next.



Setup Type

Choose the setup type that best suits your needs.

Please select a setup type.

[✓ Typical]

Common program features will be installed. Recommended for general use.

Complete

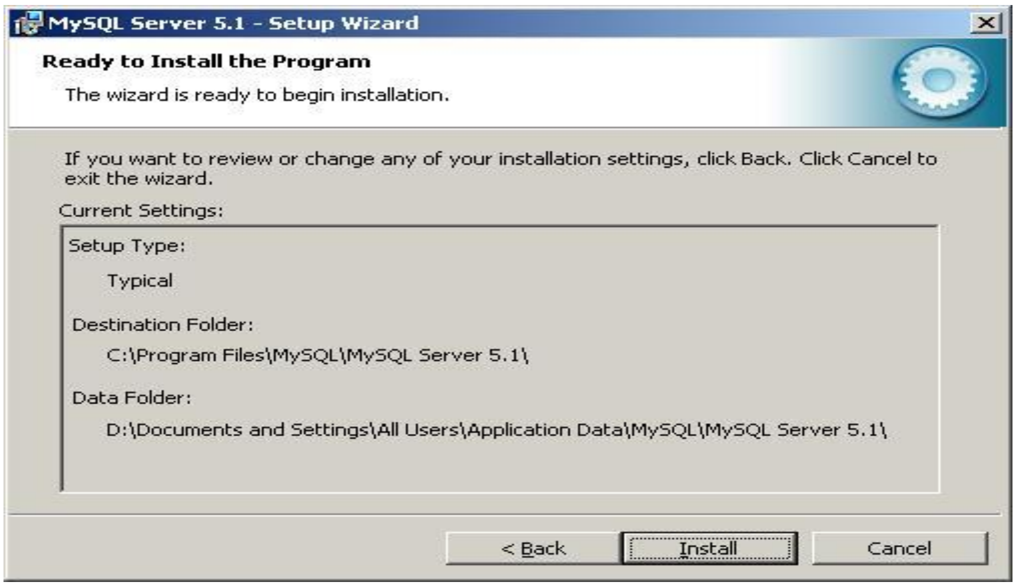
All program features will be installed. (Requires the most disk space.)

Custom

Choose which program features you want installed and where they will be installed. Recommended for advanced users.

Fig 5.4.2: Window to select setup type

STEP 4: Ready to install MySQL. Click on Install.



Ready to Install the Program

The wizard is ready to begin installation.

If you want to review or change any of your installation settings, click Back. Click Cancel to exit the wizard.

Current Settings:

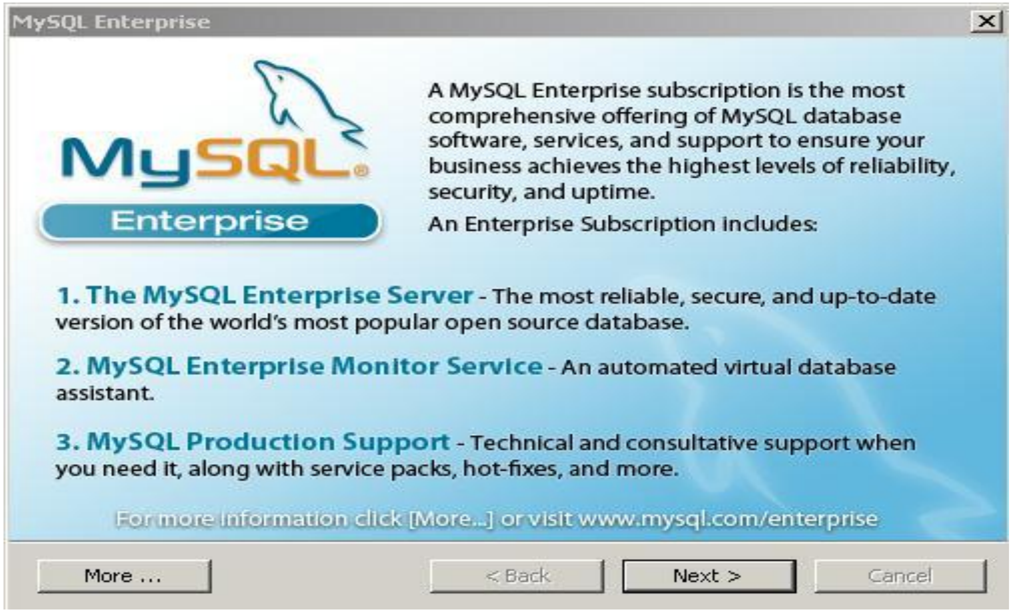
Setup Type:
Typical

Destination Folder:
C:\Program Files\MySQL\MySQL Server 5.1\

Data Folder:
D:\Documents and Settings\All Users\Application Data\MySQL\MySQL Server 5.1\

Fig 5.4.3: Installation window

STEP 5: More information about MySQL Enterprise Subscription. Click Next.



A MySQL Enterprise subscription is the most comprehensive offering of MySQL database software, services, and support to ensure your business achieves the highest levels of reliability, security, and uptime.

An Enterprise Subscription includes:

1. The MySQL Enterprise Server - The most reliable, secure, and up-to-date version of the world's most popular open source database.

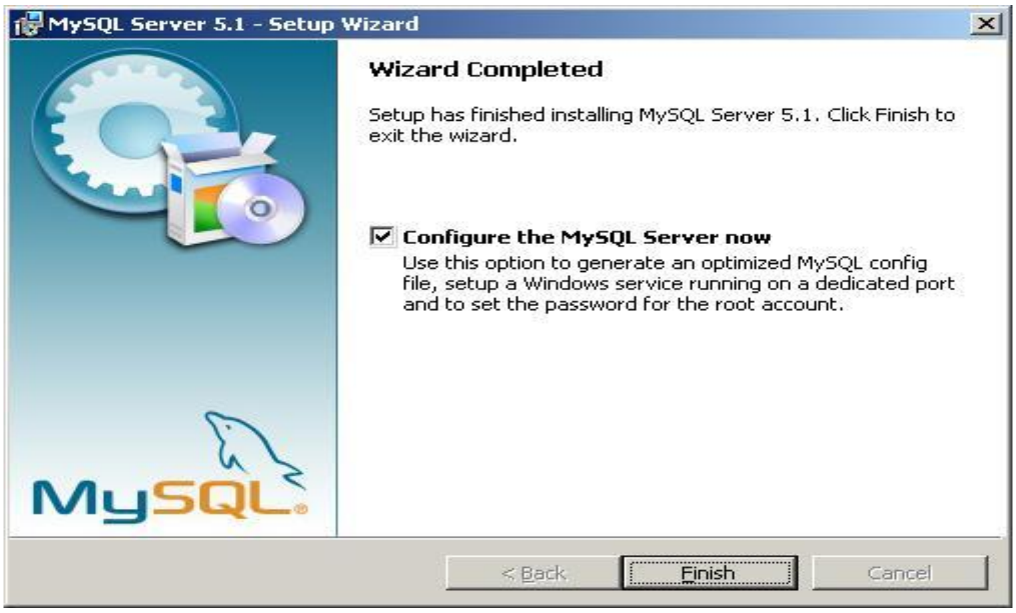
2. MySQL Enterprise Monitor Service - An automated virtual database assistant.

3. MySQL Production Support - Technical and consultative support when you need it, along with service packs, hot-fixes, and more.

For more information click [More...] or visit www.mysql.com/enterprise

Fig 5.4.4: MySQL Enterprise Subscription window

STEP 6: Setup Wizard Completed. Select the 'Configure the MySQL Server now' checkbox', click on Finish button.


MySQL Server 5.1 - Setup Wizard

Wizard Completed

Setup has finished installing MySQL Server 5.1. Click Finish to exit the wizard.

☑ Configure the MySQL Server now

Use this option to generate an optimized MySQL config file, setup a Windows service running on a dedicated port and to set the password for the root account.

Fig 5.4.5: Installation completed window

5.5 HTML

Hyper Text Mark-up Language is a structural mark-up language used to create and format web document. A mark-up language such as HTML is simply a collection of codes, called Elements that are used to indicate the structure and format of a document. A user agent, usually a web browser that renders the document, interprets the meaning of these codes to figure how to structure or display a document. HTML is not invention but it is an improved version of Standard Generalized Mark-up Language.

HTML in the following four stages:

- Level-0 included only the basic structural elements and assured that all browsers supported all features.
- Level-1 advanced features included highlighted text and graphics that were supported depending on the browser capability.
- Level –2 introduced the World Wide Web as an interactive medium and the feature of fill out forms on the Internet.
- Level-3 introduced frames, inline, video, sound, etc.

5.5.1 Importance of HTML

- HTML can be used to display any type of document on the host computer, which can be geographical at a different location.
- It is a versatile language and can be used on any platform or desktop.
- The appearance of a Web page is important, and HTML provides tags to make the document look attractive. Using graphics, fonts, different sizes, color, etc. can enhance the presentation of the document.

5.5.2 Functionality of HTML in the project

We know that this is a web-based project.

- Used to design the forms.
- Admin can communicate easily with server.

CHAPTER: 6

DESIGN

6.1 Introduction

The unified modeling language allows the software engineer to express an analysis model using the modeling notation that is governed by a set of syntactic semantic and pragmatic rules. A UML system is represented using five different views that describe the system from distinctly different perspective.

UML is specifically constructed through two different domains they are:

- UML Analysis modeling, this focuses on the user model and structural model views of the system.
- UML design modeling, which focuses on the behavioral modeling, implementation modeling and environmental model views.

Why Use UML in projects?

As the strategic value of software increases for many companies, the industry looks for techniques to automate the production of software and to improve quality and reduce cost and time-to-market. These techniques include component technology, visual programming, patterns and frameworks. Businesses also seek techniques to manage the complexity of systems as they increase in scope and scale. In particular, they recognize the need to solve recurring architectural problems, such as physical distribution, concurrency, replication, security, load balancing and fault tolerance. Additionally, the development for the World Wide Web, while making some things simpler, has exacerbated these architectural problems. The Unified Modeling Language (UML) was designed to respond to these needs.

Simply, Systems design refers to the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements which can be done easily through UML diagrams.

The following are the UML Diagrams

1. Class Diagram
2. Use Case Diagram
3. Sequence Diagram
4. Activity Diagram

5. Collaboration Diagram
6. Deployment Diagram
7. State Chart Diagram
8. Component Diagram

Class Diagram

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, and the relationships between the classes.

This is one of the most important of the diagrams in development. The diagram breaks the class into three layers. One has the name, the second describes its attributes and the third its methods. A padlock to left of the name represents the private attributes. The relationships are drawn between the classes. Developers use the Class Diagram to develop the classes. Analyses use it to show the details of the system.

Architects look at class diagrams to see if any class has too many functions and see if they are required to be split.

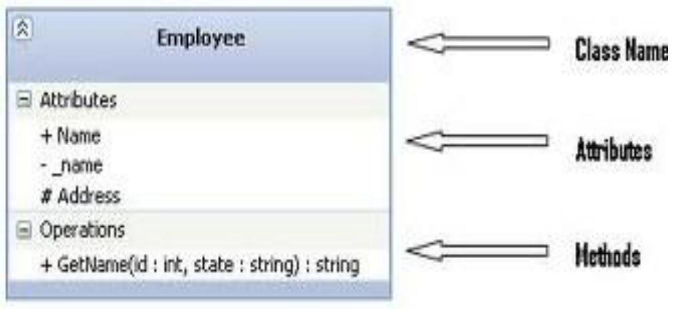


Fig 6.1: Class Structure

Use Case Diagram

In software engineering, a use case diagram in the Unified Modeling 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 main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. Use cases are used during requirements elicitation and analysis to represent the functionality of the system. Use cases focus on the behavior of the system from the external point of view. The actors are outside the boundary of the system, whereas the use cases are inside the boundary of the system.

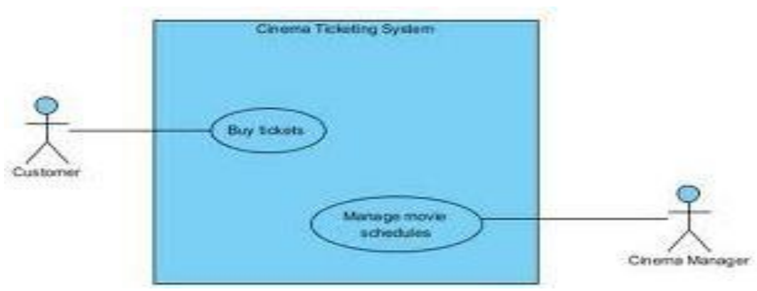


Fig 6.2: Actors and Usecases

Sequence Diagram

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called Event-trace diagrams, event scenarios, and timing diagrams.

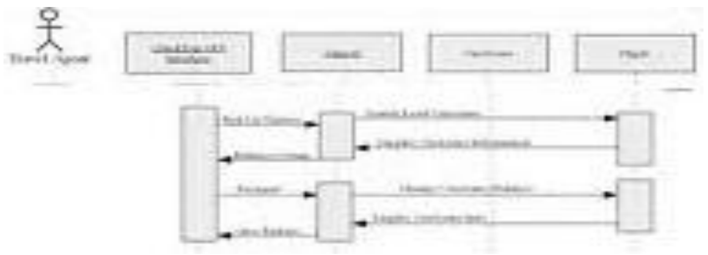


Fig 6.3: Objects and Timestamps

Activity Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

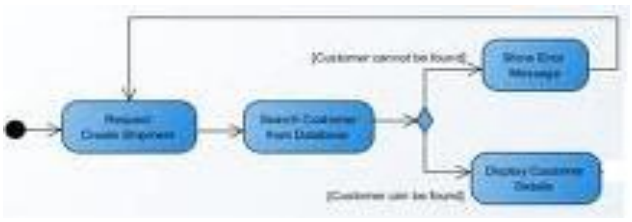


Fig 6.4: Activities

Collaboration Diagram

A Communication diagram models the interactions between objects or parts in terms of sequenced messages. Communication diagrams represent a combination of information taken from Class, Sequence, and Use Case Diagrams describing both the static structure and dynamic behavior of a system.

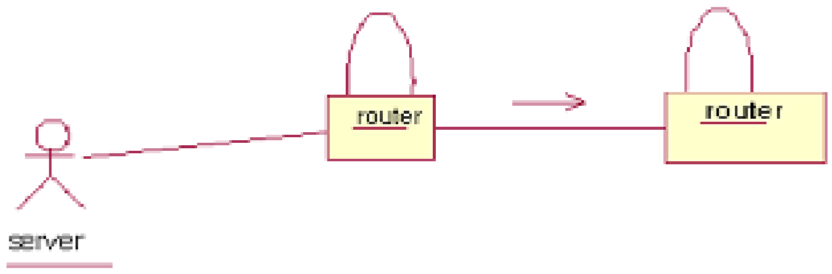


Fig 6.5: Collaborations

Deployment Diagram

A deployment diagram in the Unified Modeling Language models the physical deployment of artifacts on nodes. To describe a web site, for example, a deployment diagram would show what hardware components ("nodes") exist (e.g., a web server, an application server, and a database server), what software components ("artifacts") run on each node (e.g., web application, database), and how the different pieces are connected (e.g. JDBC).

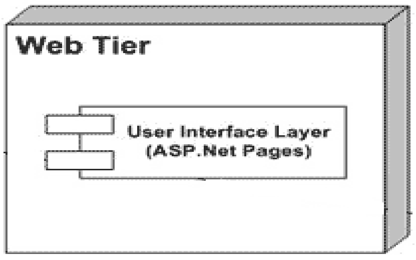


Fig 6.6: Deployment

State Chart Diagram

A state diagram is a type of diagram used in computer science and related fields to describe the behavior of systems. State diagrams require that the system described is composed of a finite number of states; sometimes, this is indeed the case, while at other times this is a reasonable abstraction. Many forms of state diagrams exist, which differ slightly and have different semantics.

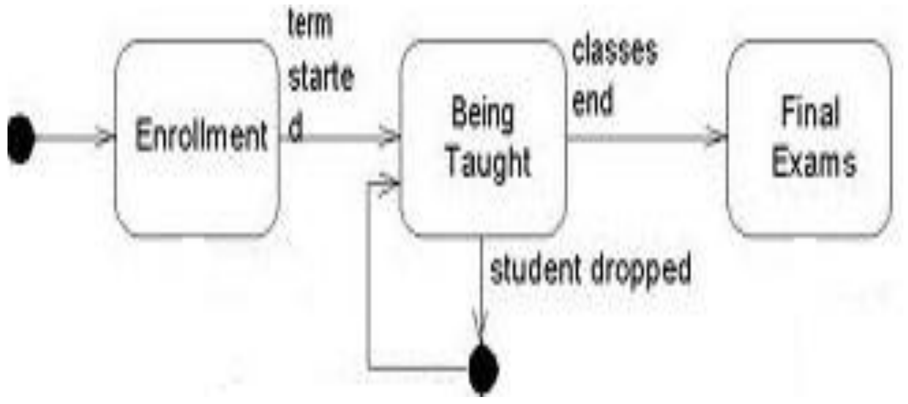


Fig 6.7: States and Relationships

Component Diagram

In the Unified Modeling Language, a component diagram depicts how components are wired together to form larger components and or software systems. They are used to illustrate the structure of arbitrarily complex systems.

User Interface Layer (ASP.Net Pages)

Fig 6.8: Components

These are different diagrams are available in a UML. Each diagram will be used for specific purpose. All diagrams programmers mainly used class diagram in order to generate code this kind of technique is called as forward engineering. That means here we convert model into code. If you convert code into model that kind of process is called as backward engineering.

6.2 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 for 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.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

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

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

6.2.1 Use Case Diagram

A use case diagram in the Unified Modeling 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 main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

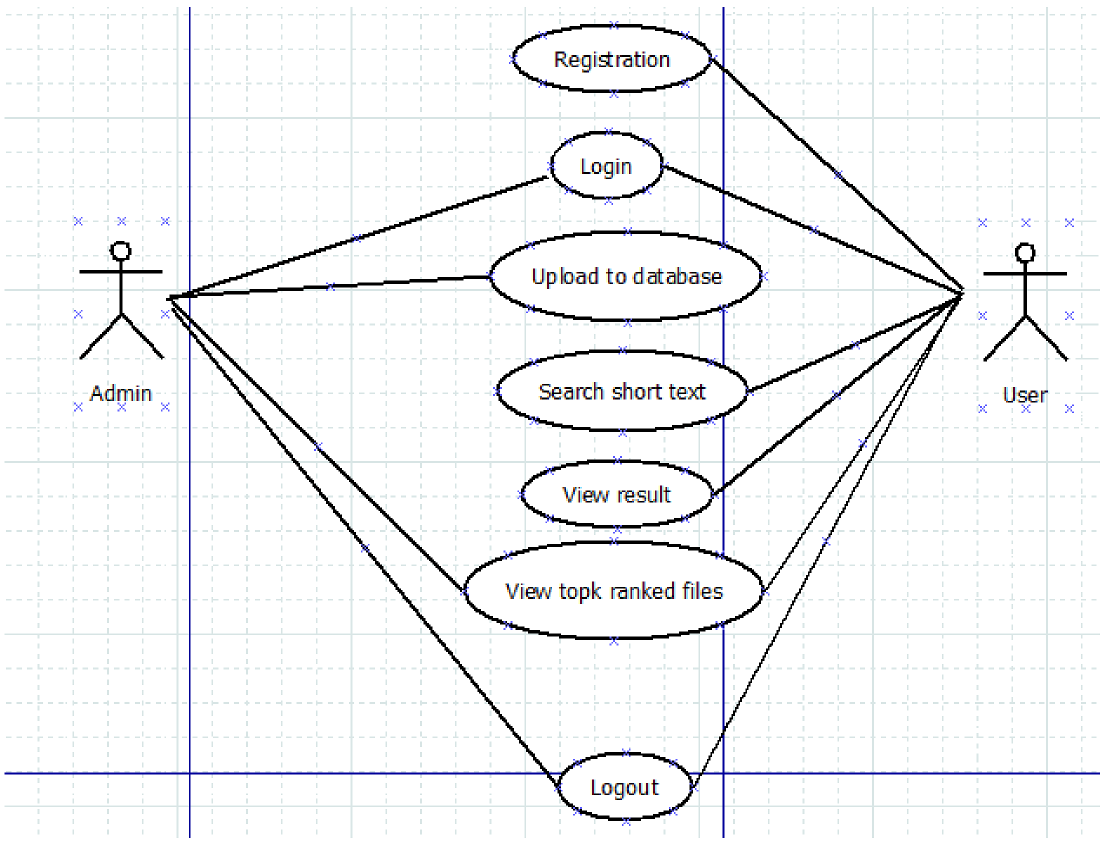


Fig 6.2.1.1: Use Case Diagram

6.2.2 Sequence Diagram

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

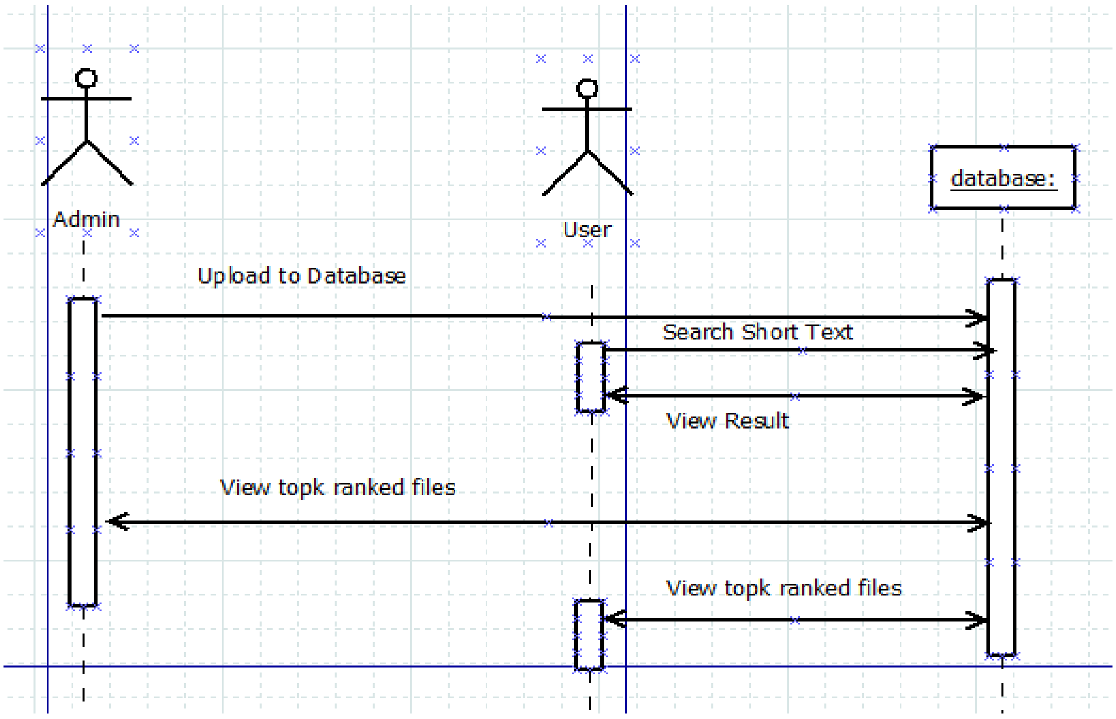


Fig 6.2.2.1: Sequence Diagram

6.2.3 Activity Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

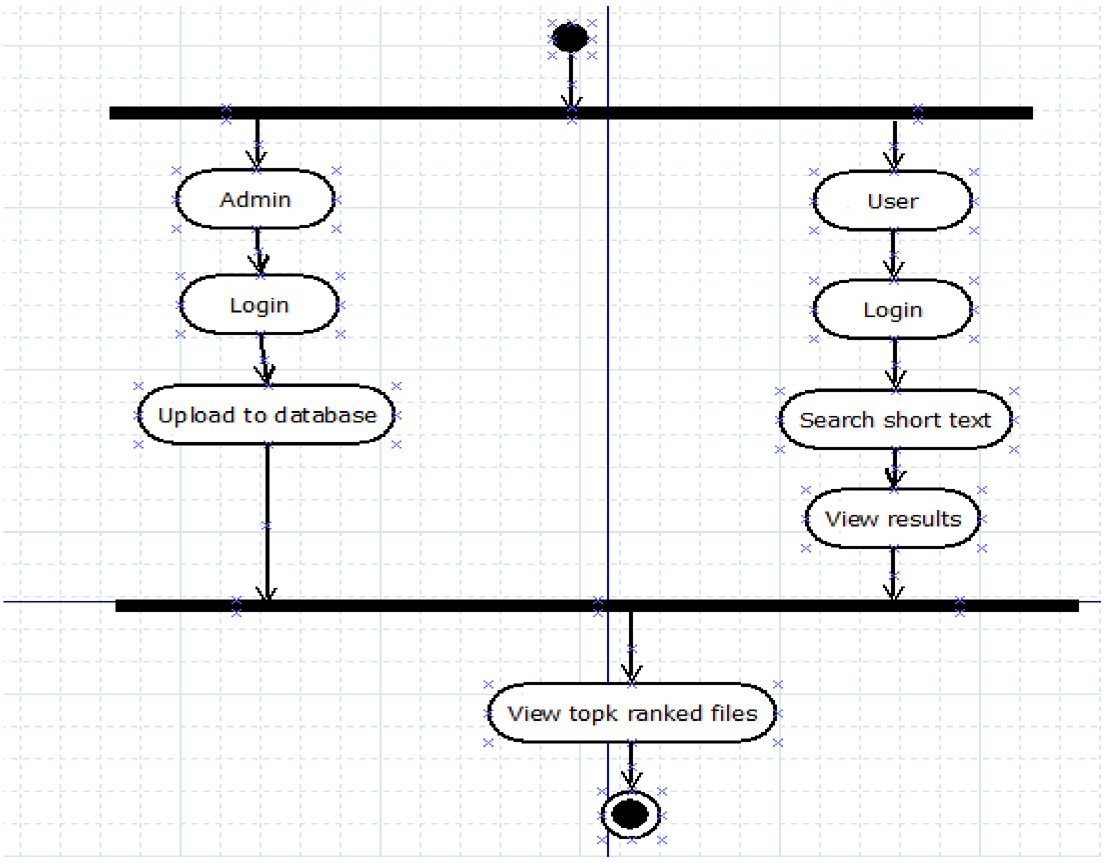


Fig 6.2.3.1: Activity Diagram

CHAPTER: 7
IMPLEMENTATION AND RESULTS

Implementation is the carrying out, execution, or practice of a plan, a method, or any design, idea, model, specification, standard or policy for doing something. As such, implementation is the action that must follow any preliminary thinking in order for something to actually happen.

7.1 Input and Output Design

7.1.1 Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

Objectives

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

7.1.2 Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

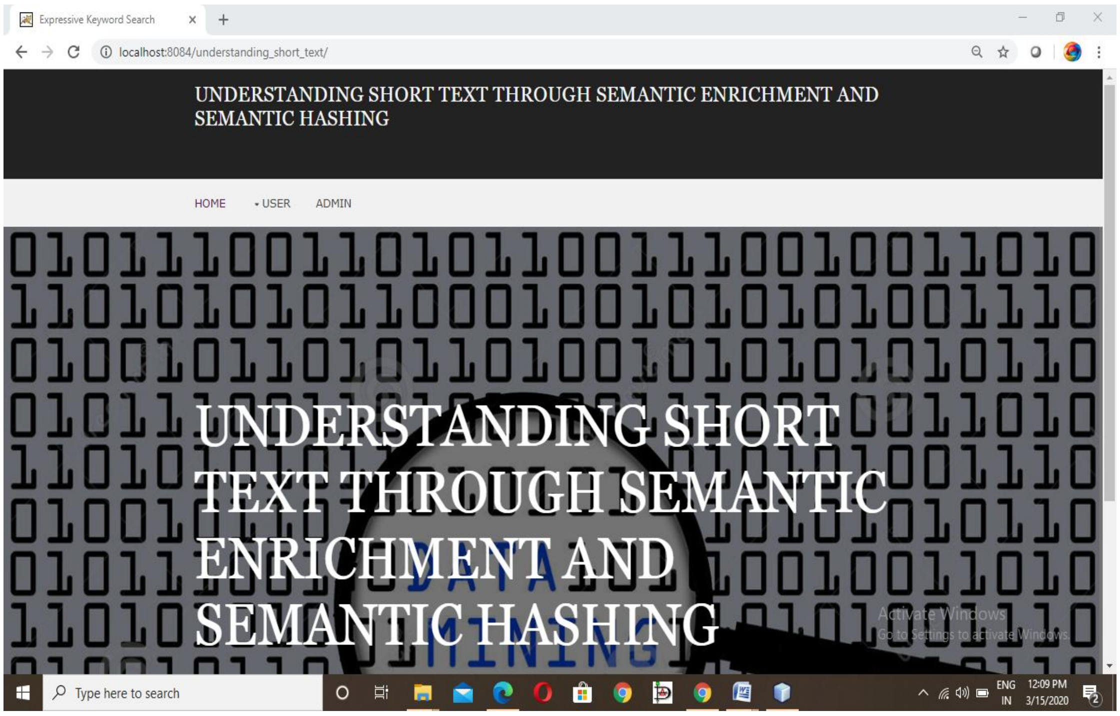
1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

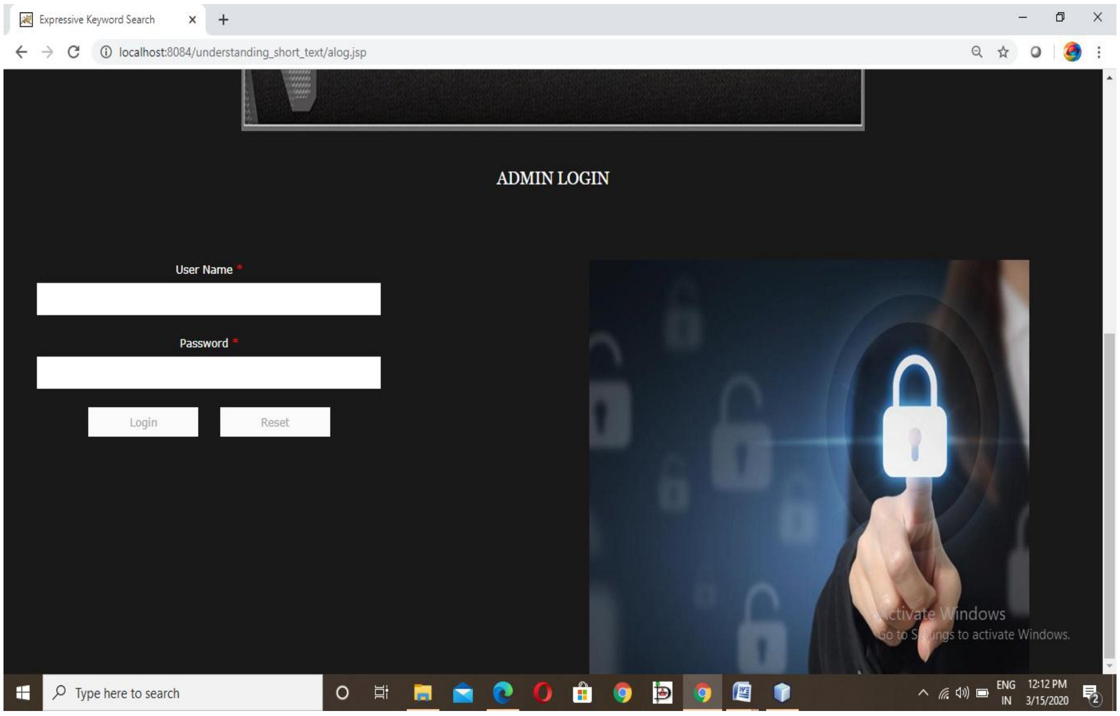
- Convey information about past activities, current status or projections of the future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.

- Confirm an action.

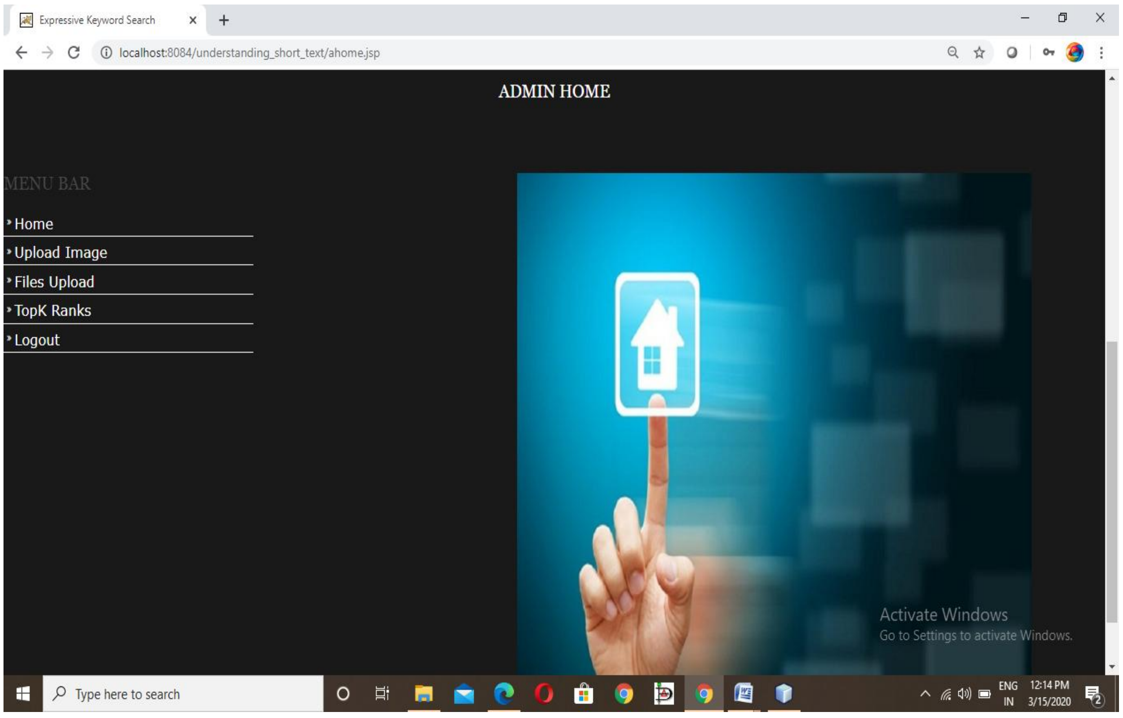
SCREEN 1: Home page


Expressive Keyword Search
localhost:8084/understanding_short_text/
UNDERSTANDING SHORT TEXT THROUGH SEMANTIC ENRICHMENT AND SEMANTIC HASHING
HOME
USER
ADMIN
UNDERSTANDING SHORT TEXT THROUGH SEMANTIC ENRICHMENT AND SEMANTIC HASHING
Type here to search

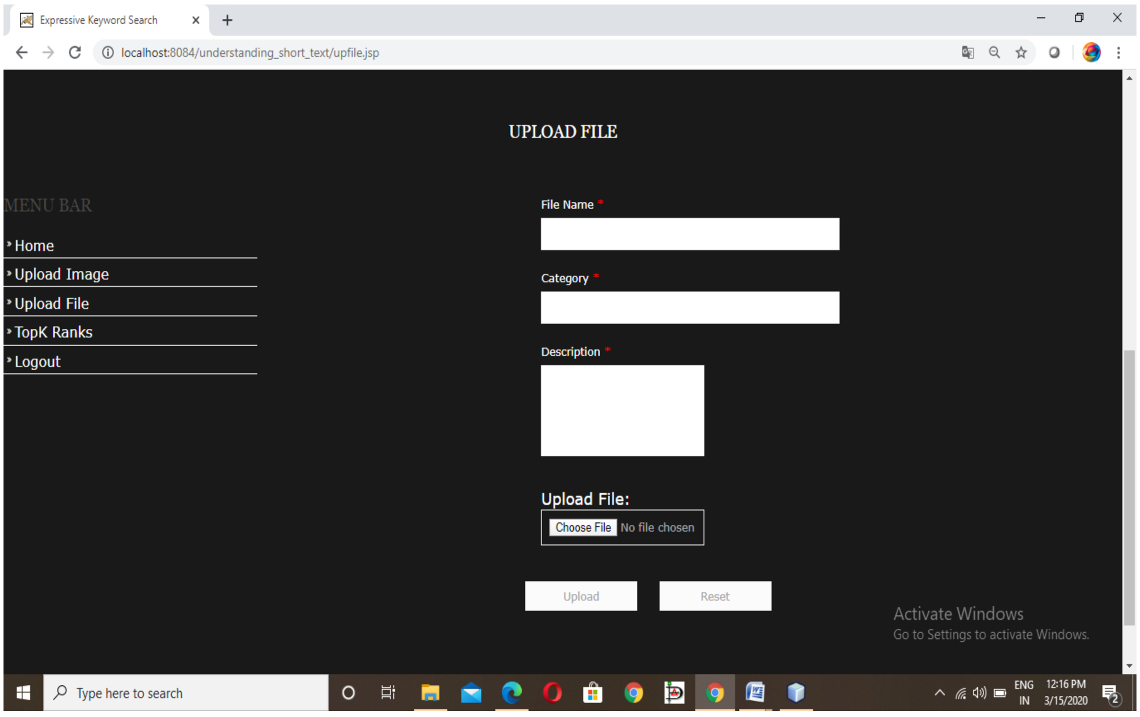
SCREEN 2: Admin Login Page


Expressive Keyword Search
localhost:8084/understanding_short_text/alog.jsp
ADMIN LOGIN
User Name *
Password *
Login
Reset
Type here to search

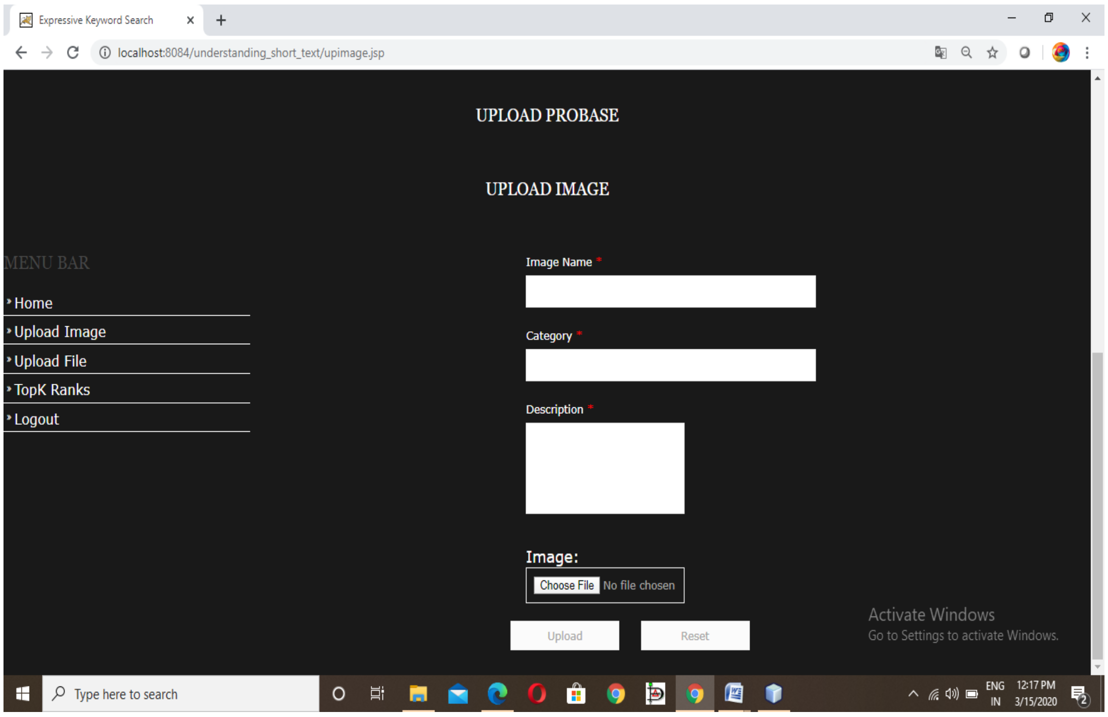
SCREEN 3: Admin home page



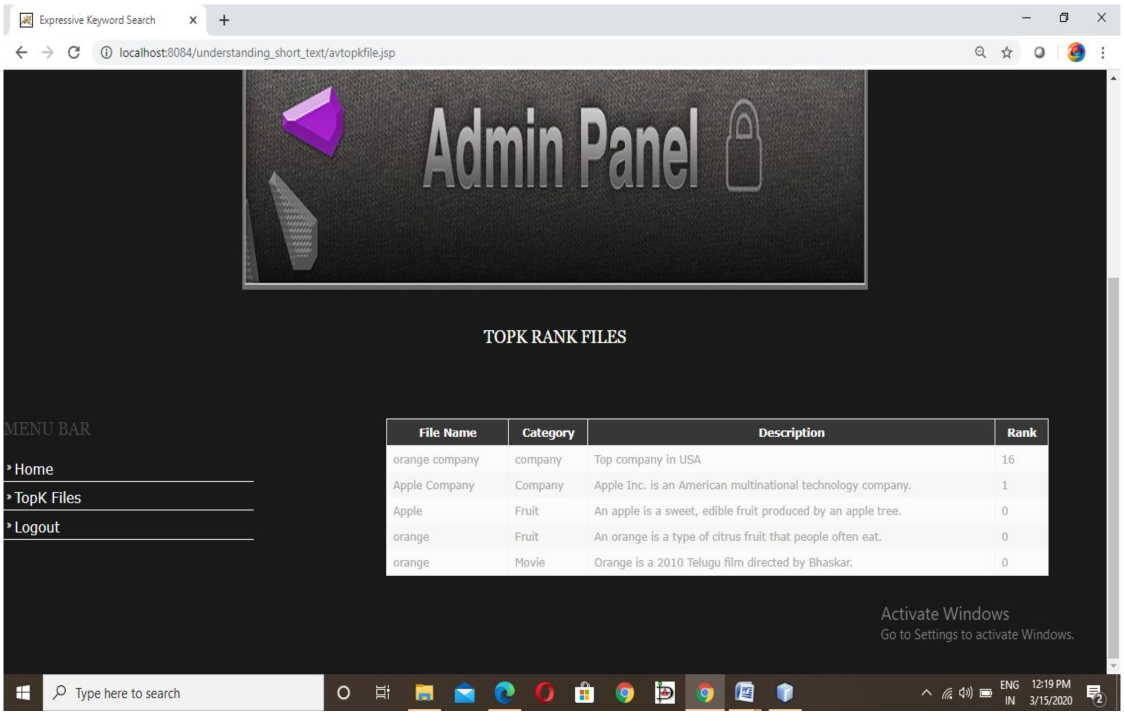
SCREEN 4: Upload files



SCREEN 5: Upload images

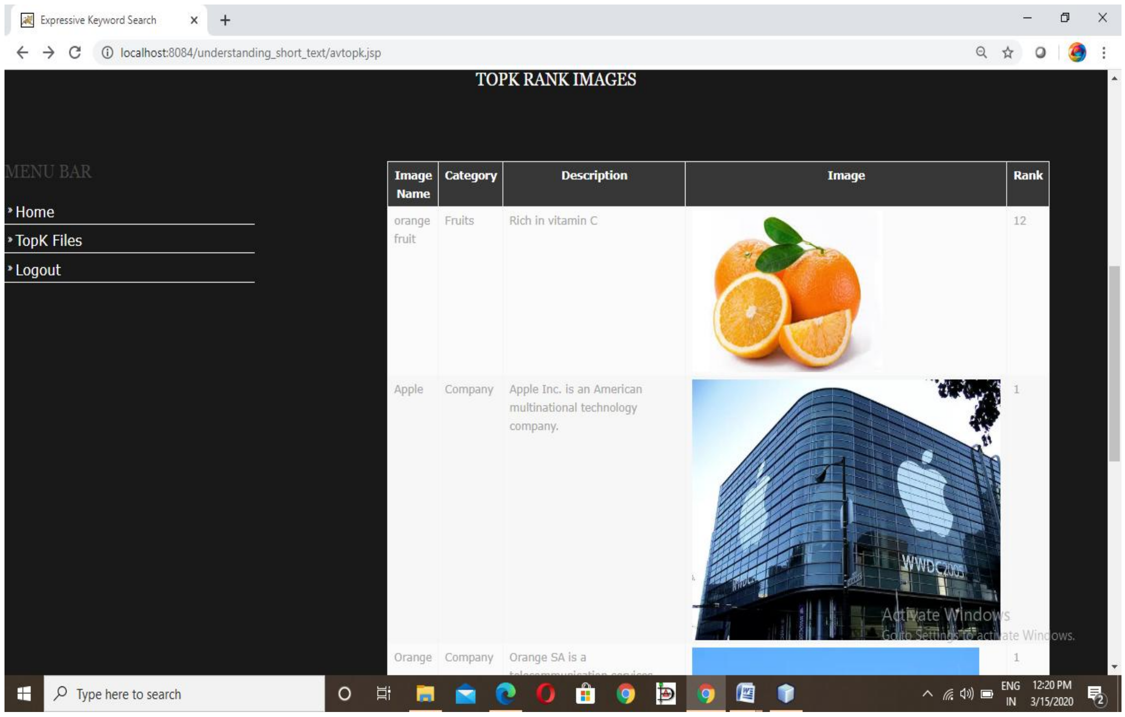


SCREEN 6: Top-k rank files


Expressive Keyword Search
localhost:8084/understanding_short_text/avtopkfile.jsp
Admin Panel
TOPK RANK FILES
MENU BAR
Home
TopK Files
Logout
Activate Windows
Go to Settings to activate Windows.
Type here to search
ENG IN
12:19 PM
3/15/2020

File Name	Category	Description	Rank
orange company	company	Top company in USA	16
Apple Company	Company	Apple Inc. is an American multinational technology company.	1
Apple	Fruit	An apple is a sweet, edible fruit produced by an apple tree.	0
orange	Fruit	An orange is a type of citrus fruit that people often eat.	0
orange	Movie	Orange is a 2010 Telugu film directed by Bhaskar.	0

SCREEN 7: Top-k rank images



Expressive Keyword Search

localhost:8084/understanding_short_text/avtopk.jsp

TOPK RANK IMAGES

MENU BAR
- Home
- TopK Files
- Logout

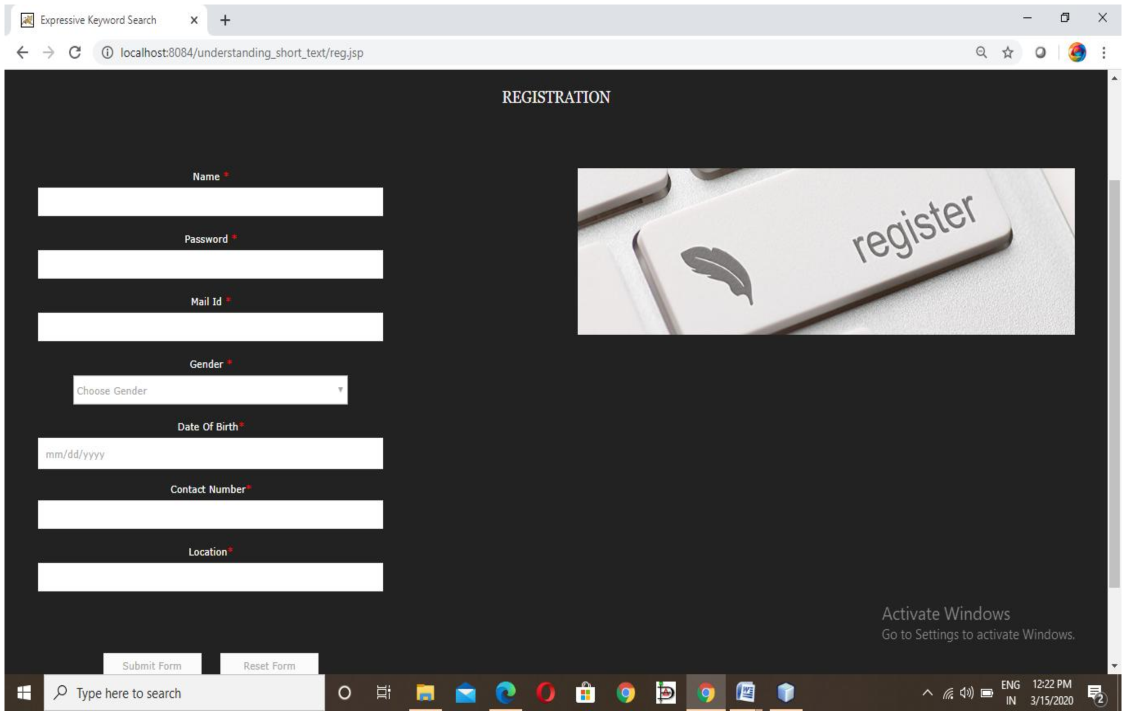
Image Name	Category	Description	Image	Rank
orange fruit	Fruits	Rich in vitamin C		12
Apple	Company	Apple Inc. is an American multinational technology company.		1
Orange	Company	Orange SA is a		1

Activate Windows
Go to Settings to activate Windows.

Type here to search

ENG IN 12:20 PM 3/15/2020

SCREEN 8: User registration



Expressive Keyword Search

localhost:8084/understanding_short_text/reg.jsp

REGISTRATION

Name *	
Password *	
Mail Id *	
Gender *	Choose Gender
Date Of Birth*	mm/dd/yyyy
Contact Number*	
Location*	

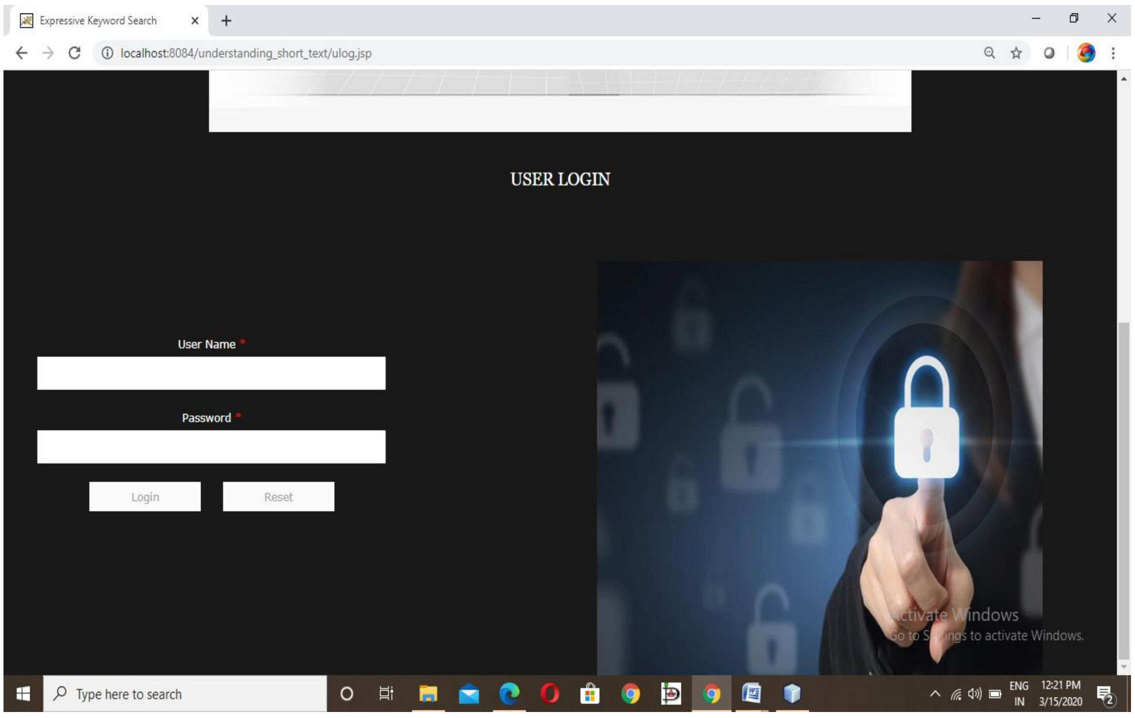
Submit Form Reset Form

Activate Windows
Go to Settings to activate Windows.

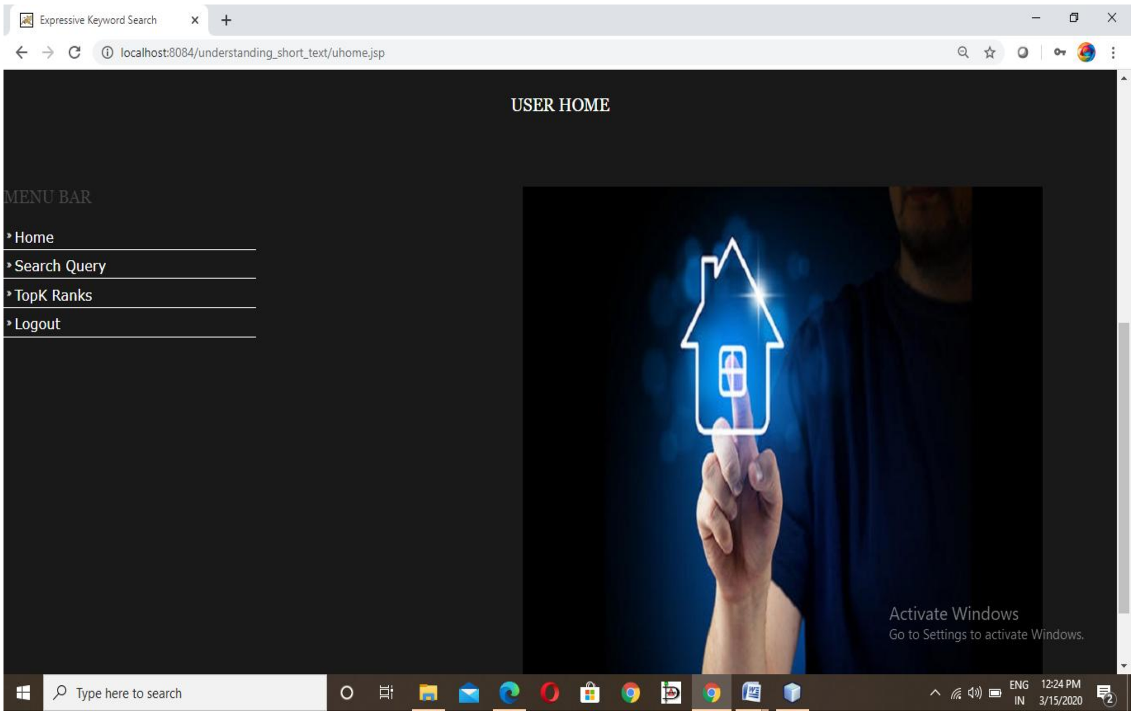
Type here to search

ENG IN 12:22 PM 3/15/2020

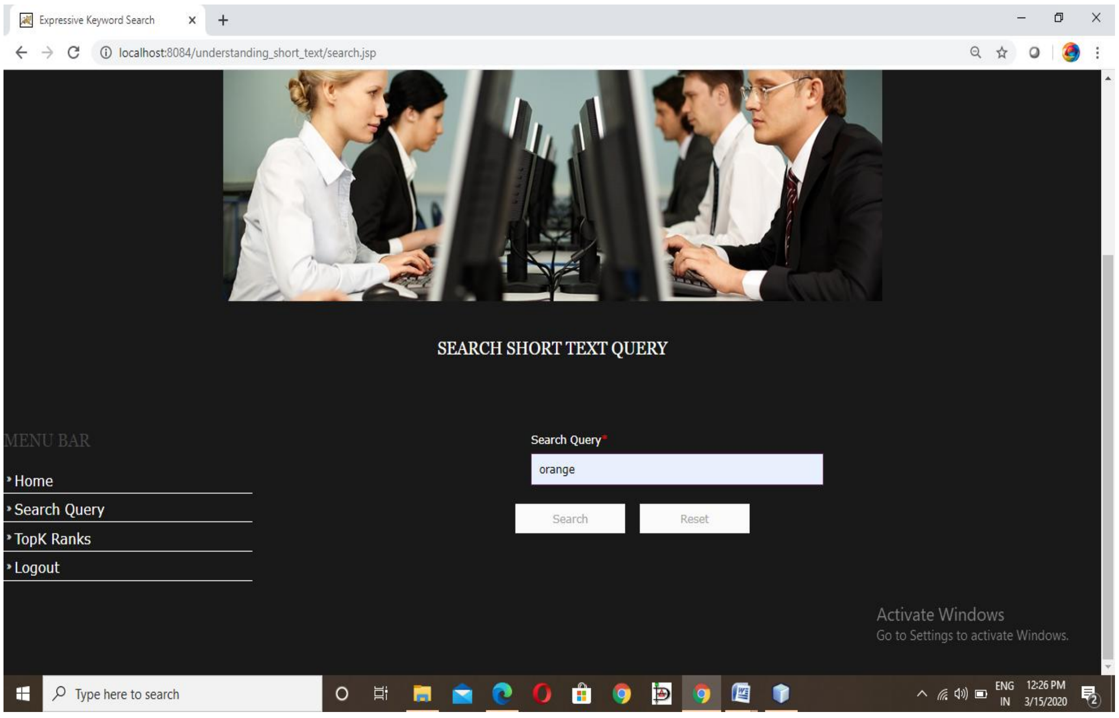
SCREEN 9: User login


Expressive Keyword Search
localhost:8084/understanding_short_text/ulog.jsp
USER LOGIN
User Name *
Password *
Login
Reset
Activate Windows
Go to Settings to activate Windows.
Type here to search
12:21 PM
3/15/2020

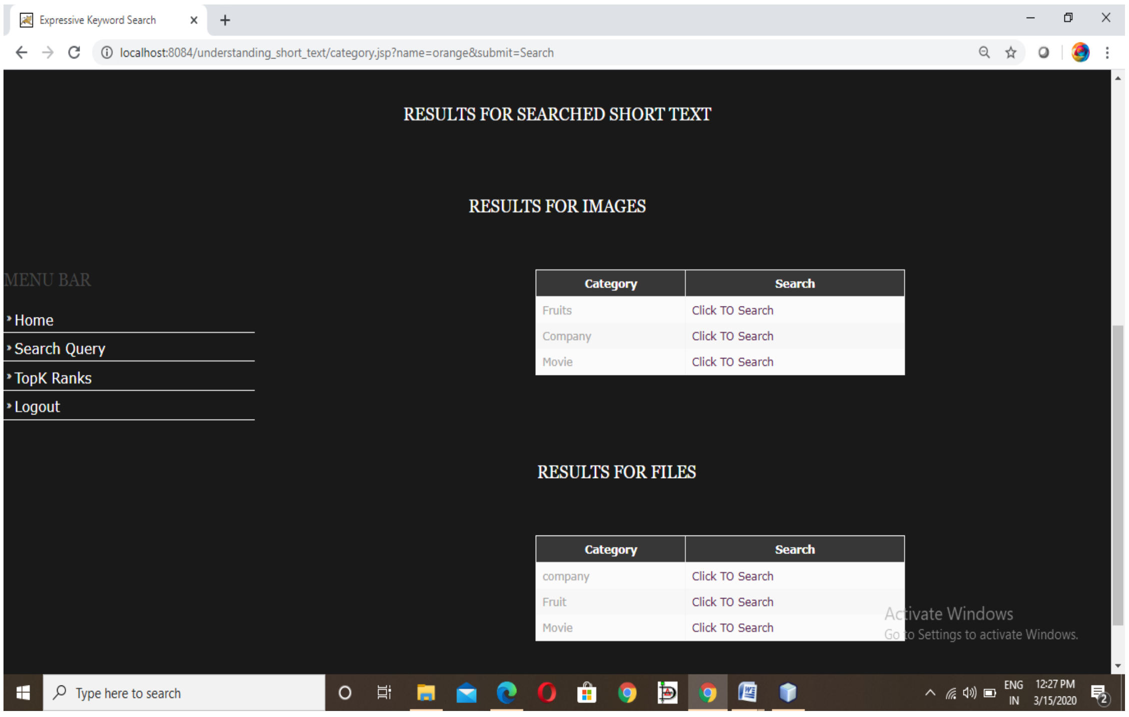
SCREEN 10: User home page


Expressive Keyword Search
localhost:8084/understanding_short_text/uhome.jsp
USER HOME
MENU BAR
Home
Search Query
TopK Ranks
Logout
Activate Windows
Go to Settings to activate Windows.
Type here to search
12:24 PM
3/15/2020

SCREEN 11: Search short text query



SCREEN 12: Results for searched short text


Expressive Keyword Search
localhost:8084/understanding_short_text/category.jsp?name=orange&submit=Search
MENU BAR
Home
Search Query
TopK Ranks
Logout
Activate Windows
Go to Settings to activate Windows.
Type here to search
12:27 PM
3/15/2020

RESULTS FOR SEARCHED SHORT TEXT

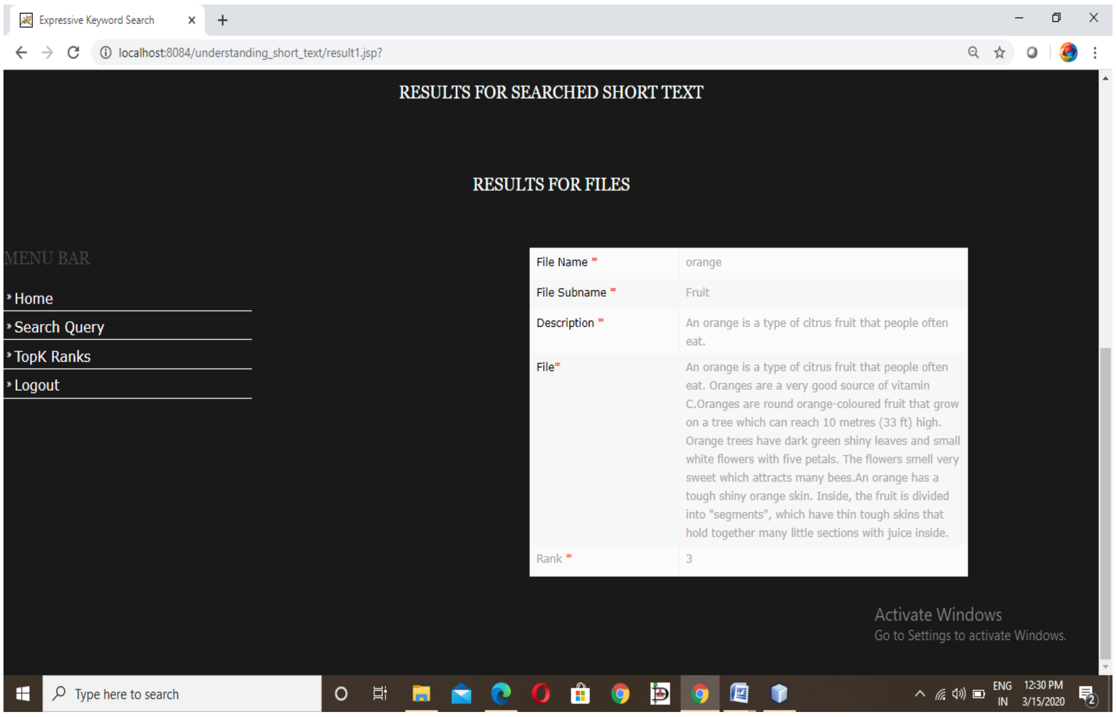
RESULTS FOR IMAGES

Category	Search
Fruits	Click TO Search
Company	Click TO Search
Movie	Click TO Search

RESULTS FOR FILES

Category	Search
company	Click TO Search
Fruit	Click TO Search
Movie	Click TO Search

SCREEN 13: Results for files

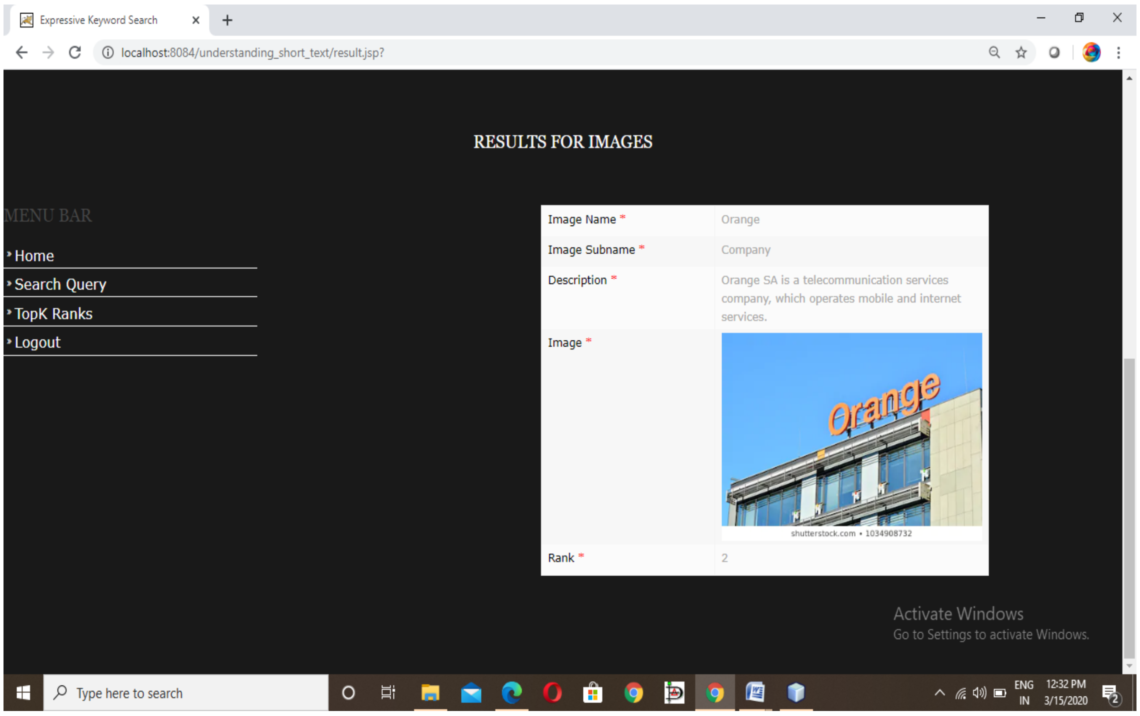

Expressive Keyword Search
localhost:8084/understanding_short_text/result1.jsp?
MENU BAR
Home
Search Query
TopK Ranks
Logout
Activate Windows
Go to Settings to activate Windows.
Type here to search
ENG IN
12:30 PM
3/15/2020

RESULTS FOR SEARCHED SHORT TEXT

RESULTS FOR FILES

File Name *	orange
File Subname *	Fruit
Description *	An orange is a type of citrus fruit that people often eat.
File*	An orange is a type of citrus fruit that people often eat. Oranges are a very good source of vitamin C.Oranges are round orange-coloured fruit that grow on a tree which can reach 10 metres (33 ft) high. Orange trees have dark green shiny leaves and small white flowers with five petals. The flowers smell very sweet which attracts many bees.An orange has a tough shiny orange skin. Inside, the fruit is divided into "segments", which have thin tough skins that hold together many little sections with juice inside.
Rank *	3

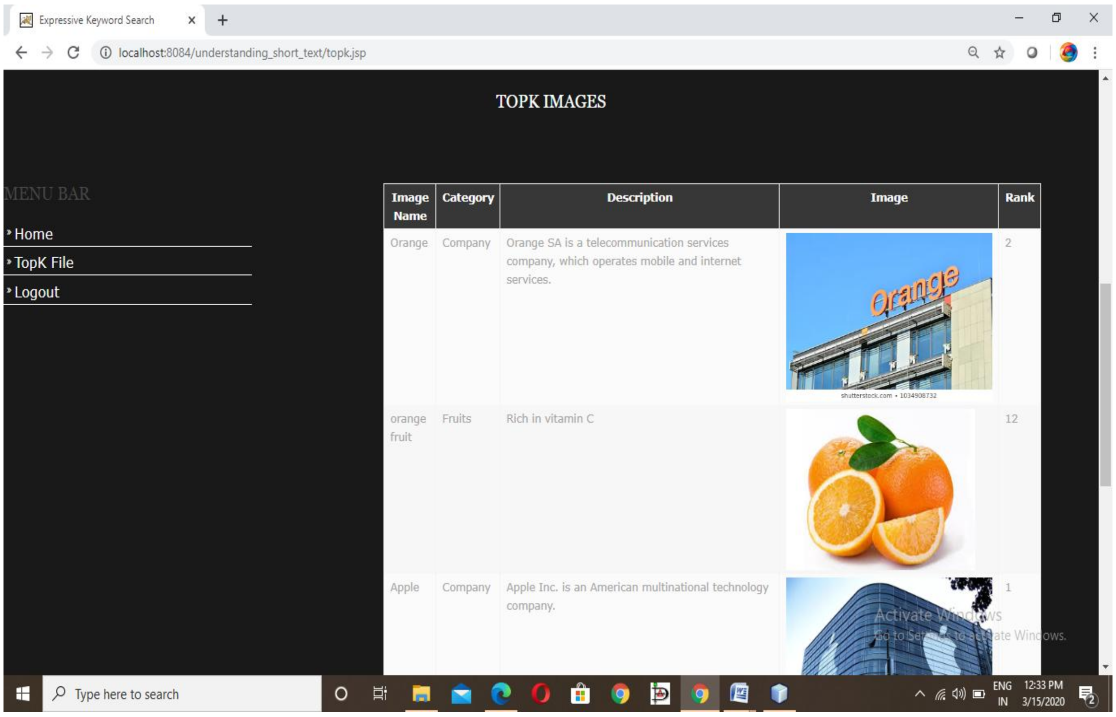
SCREEN 14: Results for images


Expressive Keyword Search
localhost:8084/understanding_short_text/result.jsp?
MENU BAR
Home
Search Query
TopK Ranks
Logout
Orange
shutterstock.com • 1034908732
Activate Windows
Go to Settings to activate Windows.
Type here to search
ENG IN
12:32 PM
3/15/2020

RESULTS FOR IMAGES

Image Name *	Orange
Image Subname *	Company
Description *	Orange SA is a telecommunication services company, which operates mobile and internet services.
Image *	
Rank *	2

SCREEN 15: Top-k images

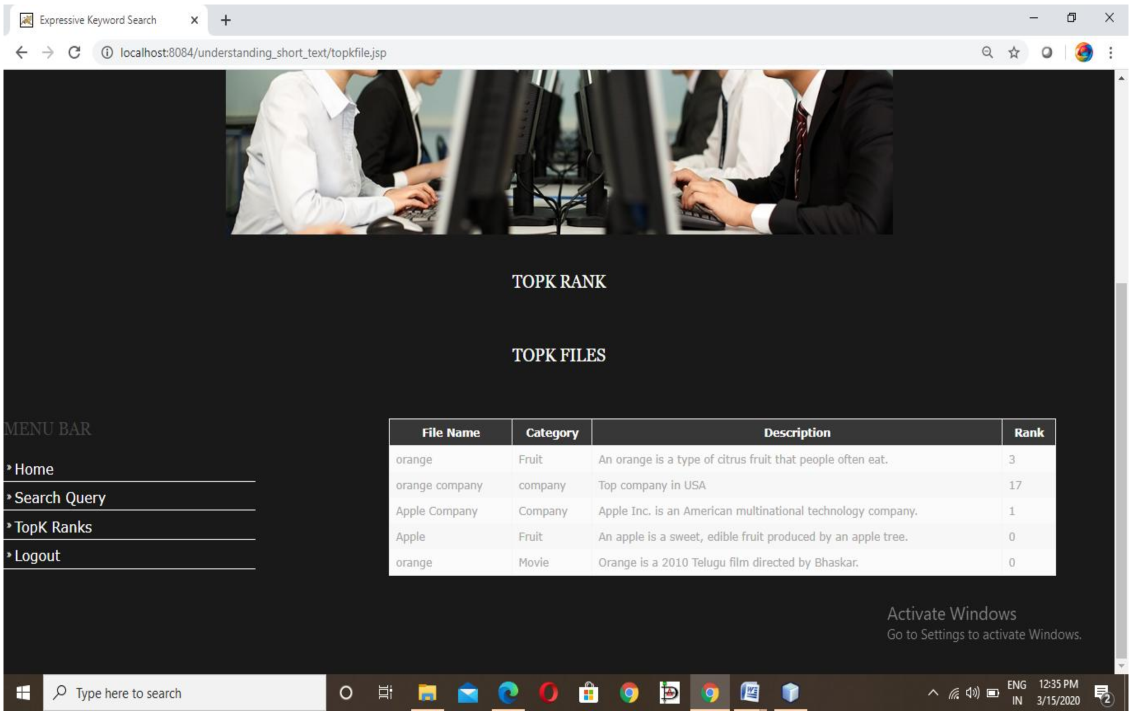


Expressive Keyword Search
localhost:8084/understanding_short_text/topk.jsp
MENU BAR
Home
TopK File
Logout
Type here to search

TOPK IMAGES

Image Name	Category	Description	Image	Rank
Orange	Company	Orange SA is a telecommunication services company, which operates mobile and internet services.		2
orange fruit	Fruits	Rich in vitamin C		12
Apple	Company	Apple Inc. is an American multinational technology company.		1

SCREEN 16: Top-k files



Expressive Keyword Search
localhost:8084/understanding_short_text/topkfile.jsp
MENU BAR
Home
Search Query
TopK Ranks
Logout
Type here to search

TOPK RANK

TOPK FILES

File Name	Category	Description	Rank
orange	Fruit	An orange is a type of citrus fruit that people often eat.	3
orange company	company	Top company in USA	17
Apple Company	Company	Apple Inc. is an American multinational technology company.	1
Apple	Fruit	An apple is a sweet, edible fruit produced by an apple tree.	0
orange	Movie	Orange is a 2010 Telugu film directed by Bhaskar.	0

CHAPTER: 8
TESTING AND VALIDATION

8.1 Introduction

Testing is an investigation conducted to provide stakeholders with information about the quality of the product or service under test. Software testing also provides an objective, independent view of the software to allow the business to appreciate and understand the risks of software implementation. Test techniques include, but are not limited to, the process of executing a program or application with the intent of finding software bugs.

8.2 Methodology used for Testing

The completion of a system will be achieved only after it has been thoroughly tested. Though this gives a feel the project is completed, there cannot be any project without going through this stage. Hence in this stage it is decided whether the project can undergo the real time environment execution without any break downs, therefore a package can be rejected even at this stage.

- **Testing methods** Software testing methods are traditionally divided into black box testing and white box testing. These two approaches are used to describe the point of view that a test engineer takes when designing test cases.
- **Black box testing** - Black box testing treats the software as a "black box," without any knowledge of internal implementation. Black box testing methods include: equivalence partitioning, boundary value analysis, all-pairs testing, fuzz testing, model-based testing, traceability matrix, exploratory testing and specification-based testing.
- **White box testing** - White box testing, by contrast to black box testing, is when the tester has access to the internal data structures and algorithms (and the code that implement these).White box testing methods can also be used to evaluate the completeness of a test suite that was created with black box testing methods. This allows the software team to examine parts of a system that are rarely tested and ensures that the most important function points have been tested.

- **Grey Box Testing:** Grey box testing involves having access to internal data structures and algorithms for purposes of designing the test cases, but testing at the user, or black-box level. Manipulating input data and formatting output do not qualify as "grey box," because the input and output are clearly outside of the "black-box" that we are calling the system under test. This distinction is particularly important when conducting integration testing between two modules of code written by two different developers, where only the interfaces are exposed for test. Grey box testing may also include reverse engineering to determine, for instance, boundary values or error messages.
- **Acceptance testing:** Acceptance testing can mean one of two things:

  A smoke test is used as an acceptance test prior to introducing a build to the main testing process.
  Acceptance testing performed by the customer is known as user acceptance testing.
- **Regression Testing:** Regression testing is any type of software testing that seeks to uncover software regressions. Such regression occurs whenever software functionality that was previously working correctly stops working as intended. Typically regressions occur as an unintended consequence of program changes. Common methods of regression testing include re-running previously run tests and checking whether previously fixed faults have re-emerged.
- **Non Functional Software Testing :** Special methods exist to test non-functional aspects of software.

Performance testing checks to see if the software can handle large quantities of data or users. This is generally referred to as software scalability. This activity of Non Functional Software Testing is often times referred to as Load Testing. Stability testing checks to see if the software can continuously function well in or above an acceptable period. This activity of Non Functional Software Testing is often times referred to as indurations test. Usability testing is needed to check if the user interface is easy to use and understand. Security testing is essential for software which processes confidential data and to prevent system intrusion by hackers. Internationalization and localization is needed to test these aspects of software, for which a pseudo localization method can be used.

Software Testing Strategies

A software testing strategy provides a road map for the software developer. Testing is a set of activities that can be planned in advance and conducted systematically. For this reason a template for software testing a set of steps into which we can place specific test case design methods should be defined for software engineering process. Any software testing strategy should have the following characteristics:

- Testing begins at the module level and works "outward" toward
  the integration of the entire computer based system.
- Different testing techniques are appropriate at different points in time.
- The developer of the software and an independent test group conducts testing.
- Testing and Debugging are different activities but debugging must be accommodated in any testing strategy.

Unit Testing

Unit testing focuses verification efforts in smallest unit of software design (module).

- Unit test considerations
- Unit test procedures

Integration Testing

Integration testing is a systematic technique for constructing the program structure while conducting tests to uncover errors associated with interfacing. There are two types of integration testing:

- Top-Down Integration: Top down integration is an incremental approach to construction of program structures. Modules are integrated by moving down wards throw the control hierarchy beginning with the main control module.
- Bottom-Up Integration: Bottom up integration as its name

implies, begins construction and testing with automatic modules.

- Regression Testing: In this contest of an integration test strategy, regression testing is the re execution of some subset of test that have already been conducted to ensure that changes have not propagate unintended side effects.

8.3 Validation

At the culmination of integration testing, software is completely assembled as a package; interfacing errors have been uncovered and corrected, and a final series of software tests – validation testing may begin. Validation can be fined in many ways, but a simple definition is that validation succeeds when software functions in a manner that can be reasonably expected by the customer.

Reasonable expectation is defined in the software requirement specification – a document that describes all user-visible attributes of the software. The specification contains a section titled “Validation Criteria”. Information contained in that section forms the basis for a validation testing approach.

8.4 Summary

This software has been computed successfully and was also tested successfully by taking “test cases”. It is user friendly, and has required options, which can be utilized by the user to perform the desired operations. Application software meets the information requirements specified to a great extent. The system has been designed keeping in view the present and future requirements in mind and made very flexible. The goals that are achieved by the software are instant access, improved productivity, Optimum utilization of resources, and efficient management of recur.

CONCLUSION

Many approaches have been proposed to facilitate short text understanding by enriching the short text representations. In this paper, we propose a novel approach for understanding short texts. First, we introduce a mechanism to enrich short texts with concepts and co-occurring terms that are extracted from a Probase.

In this project we treat a short text as a query, we enrich it with search results (e.g., webpage titles and snippets) returned by a search engine. We cluster the short text by their meaning. To cluster short texts by their meanings, we propose to add more semantics to short texts. For each term in a short text, we obtain its concepts and co-occurring terms from a database to enrich the short text. The output of the project is the search results clustered by their meanings and the user can retrieve the information as per his/her interest.

BIBLIOGRAPHY

Base Papers:

[1] M. Sahami and T. D. Heilman, “A web-based kernel function for measuring the similarity of short text snippets,” in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 377–386.

[2] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang, “Query enrichment for web-query classification,” ACM Trans. Inf. Syst., vol. 24, no. 3, pp. 320–352, 2006.

[3] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang, “Query enrichment for web-query classification,” ACM Trans. Inf. Syst., vol. 24, no. 3, pp. 320–352, 2006.

[4] S. Banerjee, K. Ramanathan, and A. Gupta, “Clustering short texts using wikipedia,” in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 787–788.

[5] Xuan-Hieu Phan, Le-Minh Nguyen,et.al.,” Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections”, April 21-25, 2008 · Beijing, China

[6] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, “Exploiting internal and external semantics for the clustering of short texts using world knowledge,” in Proc. 18th ACM Conf. Inf. Knowl. Manage., 2009, pp. 919–928.

[7] Y Krishna Priya, Dr.K.Venkata Ramana,” Semantic Enrichment of Short Texts through Conducive Formula”,Volume no:3,Issue no:6(November-2017).

Journals:

[8] Justin Zhan, Binay Dahal,” Using deep learning for short text understanding” , published in the Journal of Big Data (Springer Paper) DOI 10.1186/s40537-017-0095-2 in 2017.

[9] W. Hua, Z. Wang, H. Wang, K. Zheng and X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge" in IEEE Transactions on Knowledge & Data Engineering, vol. 29, no. 03, pp. 499-512, 2017.

Websites:

[10]https://academic.microsoft.com/search?q=Understanding+Short+Texts+through+Semantic+Enrichment+and+Hashing.

[11]https://www.semanticscholar.org/paper/Understanding-short-texts-through-semantic-and-Yu-Wang/15d9ab23a149a655efe707e3c2e31d7d8a0ae097

[12]http://www.okokprojects.com/understanding-short-texts-through-semantic-enrichment-and-hashing

[13]http://www.letscode.in/yaf_postst26864_Understanding-Short-Texts-through-Semantic-Enrichment-and-Hashing.aspx