**CHAPTER 1:**

**INTRODUCTION**

**1.1 Overview**

The widespread use of the internet has revolutionized the way people access to information. The advent of search engines has made these happenings which have proliferated tremendously over the years. There are approximately 300 million Google users and 2 billion log-in searches per day and few are valuable. These figures are so stupendous which could easily lead to information overloading and vocabulary differences.

The core of the searching capabilities is the availability of search engines which represent the user interface needed to permit users to query for information. A basic search engine consists of a crawler, indexer and a searcher or query processor which will finally rank the results according to the required output. A crawler as the name implies, crawl the website and follows all the links that are available. Developing and running a web crawler is a challenging task as it involves performance, reliability and social issues. It also consumes resources belonging to other organizations as well.

An acquittal signifies that a prosecutor failed to prove his or her case beyond a reasonable doubt, not that a defendant is innocent. Thus, a person may be acquitted of a crime but found civilly liable in a civil case regarding that same crime, e.g. O.J. Simpson, because civil cases have a lower burden of proof than criminal cases.

In several cases throughout the years, people who have been declared guilty are mostly those who are not able to prove their innocence and don’t have the resources to know what they can do. In these cases, people are proved guilty because they cannot prove themselves innocent.

So, Acquittal Search, i.e. our search engine will provide them with the means to search for various cases that can relate to them and resources to find means to contact lawyers relating to these cases. It's also a simple search engine for Jurist and Law practitioner for their advancing career.

**1.2 Background**

There are more than 1000 crimes reported each day in India which include criminal offences as well as domestic offences. Suppose a lawyer is working on a case and he needs to refer to previous cases which are similar to the one he is working on. He will have to refer to some of those cases to frame an appropriate argument.

Many of the cases don’t have a base and even if they have the defendant or their opponent does not know how to make sense out of it. This phenomenon can be due to lack of knowledge in law and crimes that are being executed over the time by either of the parties or both. Many of us don’t know how to make simple case file or process for getting justice through our justice system which is built by our own society over the years. Even after the understanding of the process, one may not be able to get certain resources even the person capable of due to the lack of knowledge and social skills. Many of us also don’t know our rights when being under legal process by our/local legal system. Even the senior advocate don’t have all the info due to the large and complicated legal and constitution documents that are amended to adjust the society’s fashion and way of living. So, most of the advocates rely on retained knowledge and instincts based on experiences gained during their advancing career.

Knowledge acquisition in the 21st century is not a problem but correct and required is a problem. There are many sources to acquire the knowledge but which source to trust and make out through their references is difficult task and time consuming, and time is crucial for solving the case for faster and efficient result. So, data should be easily available to be made out the most of the case and should be verified.

**1.3 Problem Statement**

The legal process in a country is complicated and large at most and this is a problem with government and the people who comes under the legal process. Most of the time government also don’t get to know how to solve the case due to lack of knowledge of previous cases and it also applies to common citizen who don’t even know how to proceed in quest of justice. Searching through for answer is a tough one when it comes to non-popular domain like law, metrological area maps which are present but have fewer citations and not searched more for higher indexing.

The knowledge acquisition is a problem in many areas and it applies to law, legal cases, etc. There is plenty of data around but searching through is a tough part due to diversity of data to build a common search engine. So, to do there should be a domain specific search engine.

The search engine should serve all the people who are having trouble understanding and proceeding in a legal process. It should give what happed in past, how it happened, who were involved in that process, when it was recorded, where the process began and end, etc. to a bigger picture about the situation that is being dealt with.

**1.4 Methodology**

To tackle a problem there should be a method or set of methods to approach a problem. So, at early stage of the project our team tried to think through multiple possibilities to solve the problem that we mentioned in section 1.2. From the multiple possibilities many of them were not feasible to carry either practically or theoretically.

* To make the roadmap of our project we started to go through the domain of our project to find out how the legal system works.
  + What are the hurdles faced by the people to access.
  + How the legal system change to adapt the society and trend.
  + How its affect our lives.
  + How to go through the legal process.
  + How laws are formed.
  + And many other subtle questions to clarify our view and working of the system.
* Then we selected a country which was India. Selection was important because selecting multiple countries would make our project complex and also the time to make project would be increased by multiple folds which was not feasible economically and under time constraint.

So, by narrowing those possibilities we came to conclusion to carry our project in phases and divide the work among us for faster execution of project work.

**1.4.1 Solution**

More over our team created the simple abstract model to work out the project in small and impactful chunks of work.

1. Acquire certain knowledge about the law domain to proceed with our project.
2. Find sources related to laws, legal cases and constitution for a specific country (in our case we chose India).
3. Create the database and consolidate the data from the sources.
4. Create algorithm for search engine
   1. Create machine learning algorithm for efficient search.
   2. Create NLP (Natural Language Processing) system to understand the language written in the search bar and classify the problem in the statement.
5. User Interface for the user to interact with backend system.

A solution to this problem is that a system has to be developed which would help the lawyer to search and go through all the similar cases that have occurred in the past. This will reduce the time taken by a lawyer to frame an argument and present it in the court. The benefit of this system is that justice would be provided much earlier and faster manner. This will indirectly improve the efficiency of the courts. Developing a search engine which can handle the queries related to crimes could act as a platform for the lawyers and their clients.

**1.5 Literature Review**

**1.5.1 Law**

Law is a system of rules that are created and enforced through social or governmental institutions to regulate behaviour. Law as a system helps regulate and ensure that a community show respect, and equality amongst themselves. State-enforced laws can be made by a collective legislature or by a single legislator, resulting in statutes, by the executive through decrees and regulations, or established by judges through precedent, normally in common law jurisdictions. Private individuals can create legally binding contracts, including arbitration agreements that may elect to accept alternative arbitration to the normal court process. The formation of laws themselves may be influenced by a constitution, written or tacit, and the rights encoded therein. The law shapes politics, economics, history and society in various ways and serves as a mediator of relations between people.

A general distinction can be made between (a) civil law jurisdictions, in which a legislature or other central body codifies and consolidates their laws, and (b) common law systems, where judge-made precedent is accepted as binding law. Historically, religious laws played a significant role even in settling of secular matters, and are still used in some religious communities. Islamic Sharia law is the world's most widely used religious law, and is used as the primary legal system in some countries, such as Iran and Saudi Arabia.

The adjudication of the law is generally divided into two main areas. Criminal law deals with conduct that is considered harmful to social order and in which the guilty party may be imprisoned or fined. Civil law (not to be confused with civil law jurisdictions above) deals with the resolution of lawsuits (disputes) between individuals or organizations.

Law provides a source of scholarly inquiry into legal history, philosophy, economic analysis and sociology. Law also raises important and complex issues concerning equality, fairness, and justice.

**1.5.2 Amendments**

An amendment is a formal or official change made to a law, contract, constitution, or other legal document. It is based on the verb to amend, which means to change. Amendments can add, remove, or update parts of these agreements. They are often used when it is better to change the document than to write a new one.

Amending the Constitution of India is the process of making changes to the nation's fundamental law or supreme law. The procedure of amendment in the constitution is laid down in Part XX (Article 368) of the Constitution of India. This procedure ensures the sanctity of the Constitution of India and keeps a check on arbitrary power of the Parliament of India.

However, there is another limitation imposed on the amending power of the constitution of India, which developed during conflicts between the Supreme Court and Parliament, where Parliament wants to exercise discretionary use of power to amend the constitution while the Supreme Court wants to restrict that power. This has led to the laying down of various doctrines or rules in regard to checking the validity/legality of an amendment, the most famous among them is the Basic structure doctrine as laid down by the Supreme Court in the case of Kesavananda Bharati v. State of Kerala.

The framers of the Constitution were neither in favour of the traditional theory of federalism, which entrusts the task of constitutional amendment to a body other than the Legislature, nor did they favour a rigid special procedure for such amendments. They also never wanted to have a British-style system where Parliament is supreme. The framers, instead, adopted a combination of the "theory of fundamental law", which underlies the written Constitution of the United States with the "theory of parliamentary sovereignty" as existing in the United Kingdom. The Constitution of India vests constituent power upon the Parliament subject to the special procedure lay down therein.

**1.5.3 Issues with Judicial System**

* **Corruption in judiciary:**

Like any other institution of the Government, the Indian judicial system is equally corrupted. The various recent scams like the CWG scam, 2G scam, Adarsh Society scam, including rapes and other atrocities in the society etc. have emphasised both the conduct of politicians and public dignitaries, including the common man, and also on the drawbacks in the functioning of Indian judiciary. There is no system of accountability. The media also do not give a clear picture on account of the fear of contempt. There is no provision for registering an FIR against a judge taking bribes without taking the permission of the Chief Justice of India.

* **Backlog of pending cases:**

India’s legal system has the largest backlog of pending cases in the world – as many as 30 million pending cases. Of them, over four million are High Court cases, 65,000 Supreme Court cases. This number is continuously increasing and this itself shows the inadequacy of the legal system. It has always been discussed to increase the number of judges, creating more courts, but implementation is always late or inadequate. The victims are the ordinary or poor people, while the rich can afford expensive lawyers and change the course of dispensation of the law in their favour. This also creates a big blockade for international investors and corporations to do business in India. And also due to this backlog, most of the prisoners in India’s prisons are detainees awaiting trial. It is also reported that in Mumbai, India’s financial hub, the courts are burdened with age-old land disputes, which act as a hurdle in the city’s industrial development.

* **Lack of transparency:**

Another problem facing the Indian judicial system is the lack of transparency. It is seen that the Right to Information (RTI) Act is totally out of the ambit of the legal system. Thus, in the functioning of the judiciary, the substantial issues like the quality of justice and accountability are not known properly.

* **Hardships of the under trials:**

In Indian jails, most of the prisoners are under trials, and who are confined to the jails till their case comes to a definite conclusion. In most of the cases, they end up spending more time in the jail than the actual term that might have had been awarded to them had the case been decided on time and, assuming, against them. Plus, the expenses and pain and agony of defending themselves in courts are worse than serving the actual sentence. Under trials are not guilty till convicted. On the other hand, the rich and powerful people can bring the police to their sides, and the police can harass or silence inconvenient and poor persons, during the long ordeals in the courts.

* **No interaction with the society:**

It is very essential that the judiciary of any country should be an integral part of the society and its interactions with the society must be made regular and relevant. It is also seen that there is involvement of common citizens in judicial decision-making in several countries. However, in India, the Indian judicial system has no connection with the society, something which it had inherited from the British judicial set-up. But, things should have changed over the last 60 years. Even today, the law officers have not been able to come closer to the ground to meet the common people.

**1.5.4 Search Engine**

A search engine maintains the following processes in near real time:

* Web crawling
* Indexing
* Searching

Web search engines get their information by web crawling from site to site. The "spider" checks for the standard filename robots.txt, addressed to it, before sending certain information back to be indexed depending on many factors, such as the titles, page content, JavaScript, Cascading Style Sheets (CSS), headings, as evidenced by the standard HTML mark up of the informational content, or its metadata in HTML meta tags.

Indexing means associating words and other definable tokens found on web pages to their domain names and HTML-based fields. The associations are made in a public database, made available for web search queries. A query from a user can be a single word. The index helps find information relating to the query as quickly as possible.

Some of the techniques for indexing and caching are trade secrets, whereas web crawling is a straightforward process of visiting all sites on a systematic basis.

Between visits by the spider, the cached version of page (some or all the content needed to render it) stored in the search engine working memory is quickly sent to an inquirer. If a visit is overdue, the search engine can just act as a web proxy instead. In this case the page may differ from the search terms indexed. The cached page holds the appearance of the version whose words were indexed, so a cached version of a page can be useful to the web site when the actual page has been lost, but this problem is also considered a mild form of link rot.

Typically when a user enters a query into a search engine it is a few keywords. The index already has the names of the sites containing the keywords, and these are instantly obtained from the index. The real processing load is in generating the web pages that are the search results list: Every page in the entire list must be weighted according to information in the indexes. Then the top search result item requires the lookup, reconstruction, and mark up of the snippets showing the context of the keywords matched. These are only part of the processing each search results web page requires, and further pages (next to the top) require more of this post processing.

Beyond simple keyword lookups, search engines offer their own GUI- or command-driven operators and search parameters to refine the search results. These provide the necessary controls for the user engaged in the feedback loop users create by filtering and weighting while refining the search results, given the initial pages of the first search results. For example, from 2007 the Google.com search engine has allowed one to filter by date by clicking "Show search tools" in the leftmost column of the initial search results page, and then selecting the desired date range. It's also possible to weight by date because each page has a modification time. Most search engines support the use of the Boolean operators AND, OR and NOT to help end users refine the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search, which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human.

The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This first form relies much more heavily on the computer itself to do the bulk of the work.

**1.6 Intended Users**

* An Advocate or Lawyer.
* Academic person studying law.
* Person researching on laws and constitution rights.
* A Judge.
* Defendant
* Plaintiff
* Any person who has been legally charged.
* Any who want to know about previous cases of their like.

**CHAPTER 2:**

**SYSTEM ANALYSIS**

**2.1 General**

Acquittal Search is a big project which involves various phases which are time consuming, the project modules in itself takes much time to build and work. So, there is much more to do technically and analyse the potential and its requirement to understand the project and proceed with it.

We had done research to accomplish the project goal. Most of the research was carried out on internet for faster and efficient results. We had to see that our project can be made with good results and check the feasibility of our project in multiple domains.

* Theoretically
* Technically
* Economically
* Legally
* Data Acquisition
* Completion in time limit.

This section will provide with simple and effective information that is basis for the project that is built. It will also give an abstract idea of our project to understand the main idea behind the project.

**2.2 Preliminary Analysis**

We collected a number of requirements for project from our primitive research, website visits, and interview to the concerned personnel and their experiences regarding the concepts of its development. To make the project we understood the similar projects that were operational, most of them were using third-party tools to accomplish their tasks and made their project easier to work with. Some of those projects made their own database to solve the problem.

We made the multiple scenarios to build the project that include:

* What technology to use for:
  + Database
  + User Interface
  + Machine Learning
  + NLP
  + Web Crawling
  + Web Server
  + Web Backend programming
  + Operating system.
* Where to host the server
* Accessibility
* Ease of use
* Scalability
* And many other micro features.

**1.3 Feasibility**

Feasibility study is a test of a system proposal according to its workability, impact on the organization, ability to meet user needs, and effective use of resources. It focuses on 3 major questions: -

1. What are the user’s demonstrable needs &how does a candidate system meet them?
2. What resources are available for given candidate systems? Is the problem worth solving?
3. What are the likely impacts of the candidate system on the organization? How well does it fit within the organization’s?

The objective of a feasibility study is not to solve the problem but to acquire a sense of its scope. During the study, the problem definition is crystallized and aspects of the problem to be included in the system are determined. Consequently, costs and benefits are estimated with greater accuracy at this stage.

The result of the feasibility is a formal proposal. It consists of the following: -

* Statement of the problem
  + A carefully worded statement of the problem that led to analysis.
* Summary of findings & recommendations
  + It is ideal for user who requires quick access to the results of the analysis of the system under study. Conclusions are stated, followed by a list of the recommendation justification for them.
* Details of findings
  + An outline of the methods & procedures undertaken by the existing system, followed by coverage of the objective and procedures of the candidate system. It also includes output reports, file structures& cost and benefits of candidate system.
* Recommendations and conclusions
  + Specific Recommendations regarding the candidate system, including personnel assignments, costs, project schedule, and target dates.

**1.3.1 Technical Feasibility**

A project is technical feasible if it is able to make the project with accessible technology. Our project utilises technology accessible and mostly open source software and flexible and scalable hardware.

**1.3.2 Economical Feasibility**

It is used for evaluating the effectiveness of a candidate system. More commonly known as cost/benefit analysis, the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with costs.

Our system is reasonably economical to be made and maintain due to the common expertise used to make and maintain the system.

**1.3.3 Operational Feasibility**

A project is operationally feasible when it is able to operate with ease and without any overhead other than what is reasonable required.

Our project is operationally feasible as it is ease to operate as it only requires discrete maintenance by the developer. And it is easy to operate by the user through web application.

**1.3.4 Behavioural Feasibility**

People are inherently resistant to change & computers have been known to facilitate change. An estimate should be made of how strong a reaction the user stay is likely to have toward the development of a computerized system. It is common knowledge that computer installations have something to do with turnover, transfers & changes in employee job status.

Therefore, it is understandable that the introduction of a candidate system requires special effort to educate, sell & train the staff on new ways of conducting business.

**1.4 Software and Hardware specifications**

**1.4.1 Hardware Requirements**

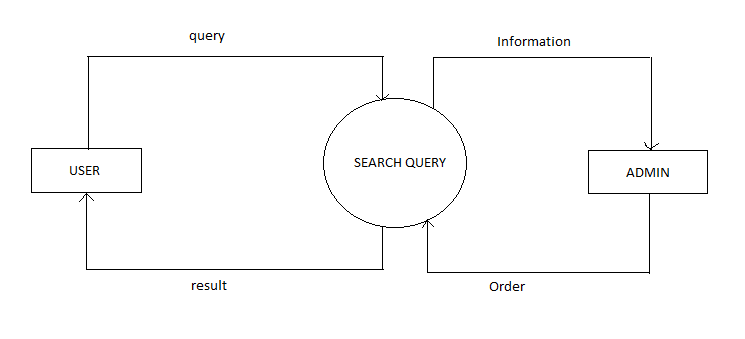
* **Server Side**
  + ***Processor:*** min. Duel Core 1.5 GHz
  + ***RAM:*** min. 4 GB
  + ***Storage Space:*** min. 80 GB
  + ***Database Space:*** min. 20GB
* **Client Side**
  + ***Processor:*** min. Duel Core 1.2 GHz
  + ***RAM:*** min. 1 GB
  + ***Storage Space:*** not applicable.

**1.4.2 Software Requirements**

* **Server Side**
  + ***Operating System:*** Any Linux Distribution with kernel 4.4 or higher.
  + ***Database:*** MySQL Server
  + ***Web Server:*** Apache HTTP server v2
  + ***Server Side Scripting language:*** PHP 5.5
  + ***Web Scraping:*** Python v3.5 or higher.
  + ***Server Connection:*** SSH server
  + ***Version control:*** Git (latest)
* **Developer Side**
  + ***Operating System:*** Any OS with Network connection
  + ***Database Client:*** MySQL client
  + ***Server Connection software:*** SSH client
  + ***SFTP Client:*** SSH data transfer client (like WinSCP)
  + ***Version control:*** Git
* **Client Side**
  + ***Operating System:*** Any OS with latest browser support.
  + ***Web Browser:*** Any Web Browser with HTML5 or higher version support.

**1.5 Data Flow Diagram**

This is a 0-level data flow diagram of a search engine system. It is the lowest level representation of a search engine where it shows that the user provides a query which will be processed by query processing methods and the search engine will return an appropriate result.



**CHAPTER 3:**

**SYSTEM DESIGN**

**3.1 Software Design**

Software design is the process by which an agent creates a specification of a software artefact, intended to accomplish goals, using a set of primitive components and subject to constraints. Software design may refer to either "all the activity involved in conceptualizing, framing, implementing, commissioning, and ultimately modifying complex systems" or "the activity following requirements specification and before programming, as ... [in] a stylized software engineering process."

Software design usually involves problem solving and planning a software solution. This includes both a low-level component and algorithm design and a high-level, architecture design.

**3.1.1 Use Case Diagram**

Use case modelling plays a great role in capturing requirements, as well as presenting the behaviour of the system. It can be used in understanding the current system, whether it is automated or manual, and to model a new system.

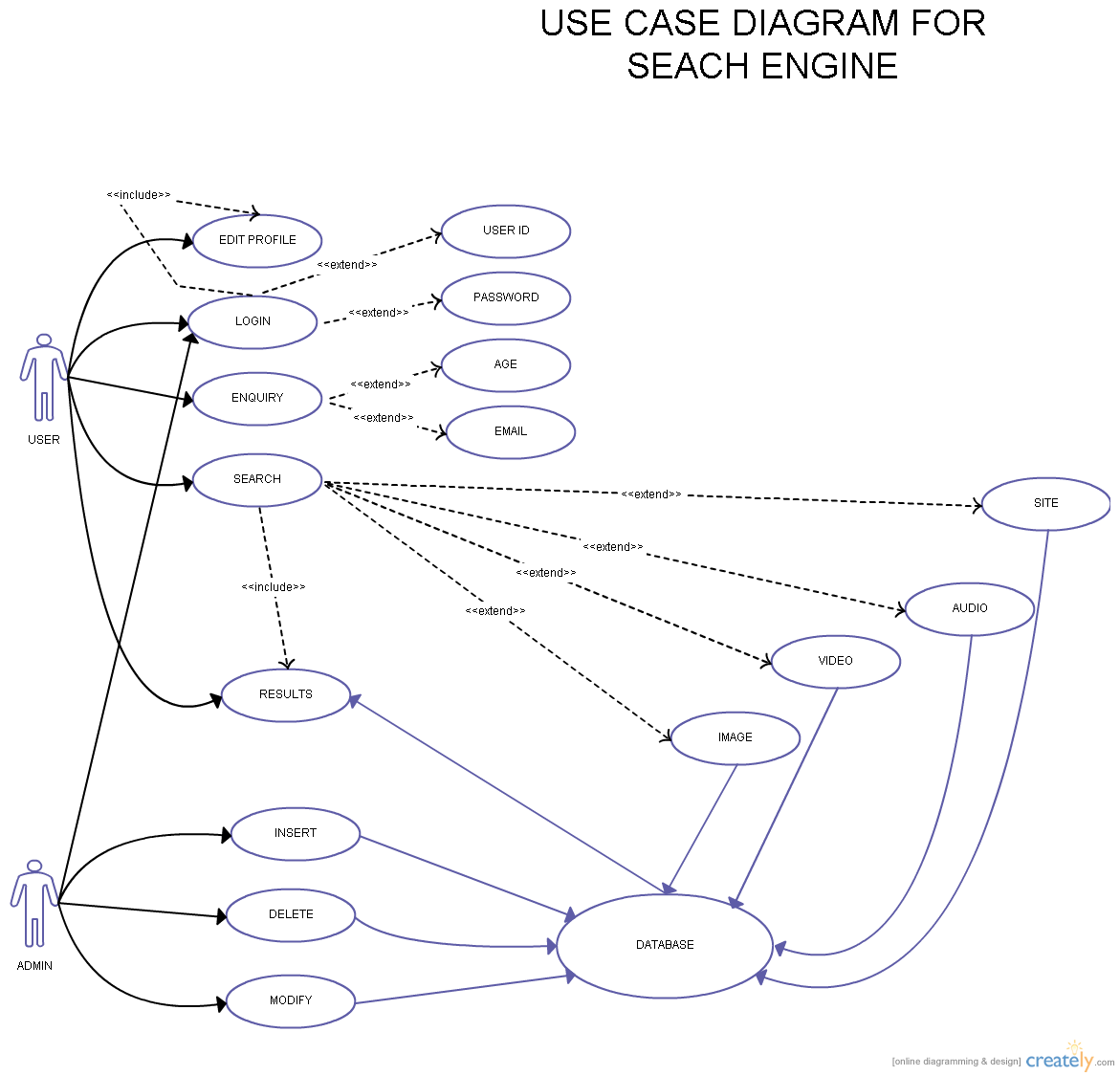
A use case simply describes how a user gets something done. It is an event that describes behaviours from a user’s perspective. Together, a set of use cases define all the ways to use your system.

The first places to look for use cases are from the goals belonging to the actors we’ve identified. Those are a great start, but there are other places to look for use cases. For example, ask yourself how information is going to get into the system. What events happen in the real world that would make an actor want to do something? What output does the system need to produce? What would happen outside the system that would cause it to make something?

We can define the scope of a use case through preconditions and success guarantees. Preconditions describe what needs to be true before a use case can start. A good precondition is something that your system can detect. Success guarantees define what must be true after a use case has successfully completed. It should also be written as something the system can detect.

It describes the interaction between the actor and the system as the actor completes the use case.

The user will interact with the search interface and will type his query and will get the result in response to that query, it could be text, image, video or audio.

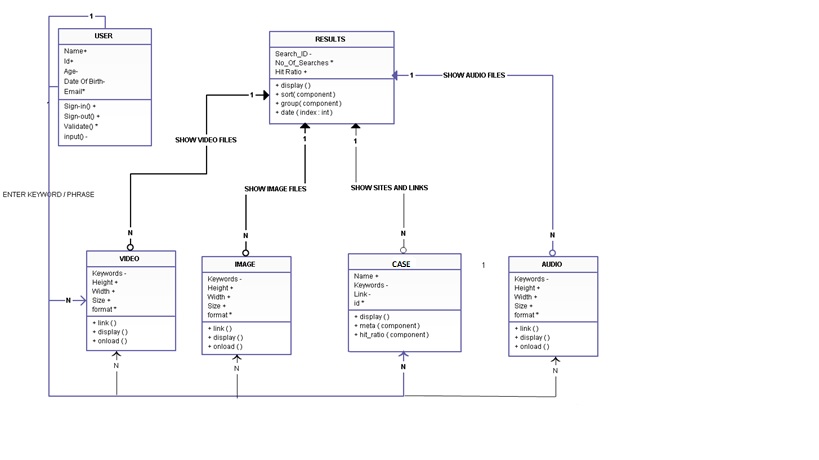


**3.1.2 Class Diagram**

Class diagram is a static diagram. It represents the static view of an application.

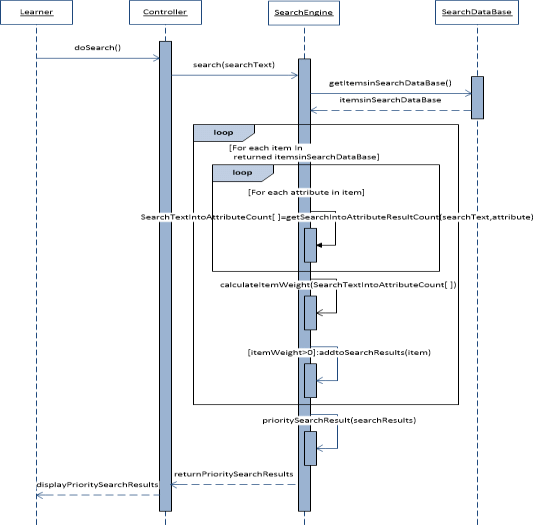
The class shape itself consists of a rectangle with three rows. The top row contains the name of the class, the middle row has the attributes of the class, and the bottom section expresses the methods or operations that the class may utilize.

Here the USER class has a bidirectional relationship with all of the VIDEO, IMAGE, CASE AND AUDIO Class with a multiplicity value of 1 meaning it can have N instances of VIDEO, IMAGE, SITES and AUDIO associated with it. All of the classes can have only one instance of RESULTS, but the RESULTS can have N instance of each of them associated with it.



**3.1.3 Sequence Diagram**

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios.



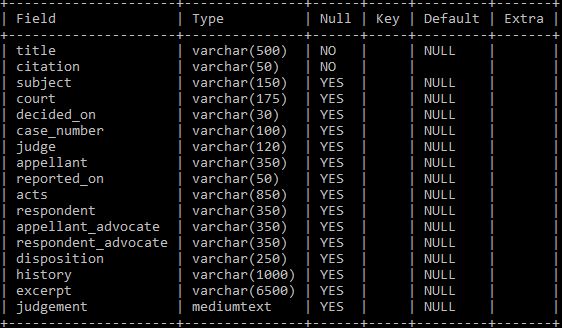
**3.2 User Interface Design**

**3.3 Database Design**

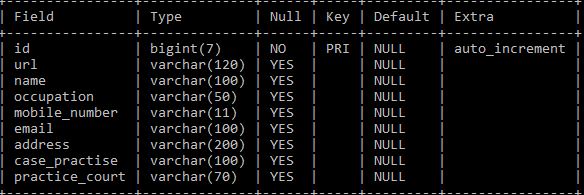
Database design is the process of producing a detailed data model of database. This data model contains all the needed logical and physical design choices and physical storage parameters needed to generate a design in a data definition language, which can then be used to create a database. A fully attributed data model contains detailed attributes for each entity.

The term database design can be used to describe many different parts of the design of an overall database system. Principally, and most correctly, it can be thought of as the logical design of the base data structures used to store the data. In the relational model these are the tables and views. In an object database the entities and relationships map directly to object classes and named relationships. However, the term database design could also be used to apply to the overall process of designing, not just the base data structures, but also the forms and queries used as part of the overall database application within the database management system (DBMS).

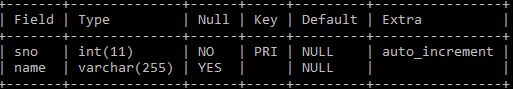
**3.1 Cases Database**



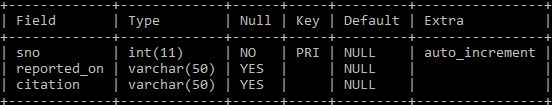
**3.2 Lawyers Database**



**3.3 Acts Database**



**3.4 Reported On Database**



**CHAPTER 4:**

**IMPLEMENTATION**

**4.1 Resources Utilised**

In project management terminology, resources are required to carry out the project tasks. They can be people, equipment, facilities, funding, or anything else capable of definition (usually other than labour) required for the completion of a project activity. The lack of a resource will therefore be a constraint on the completion of the project activity. Resources may be storable or non storable. Storable resources remain available unless depleted by usage, and may be replenished by project tasks which produce them. Non-storable resources must be renewed for each time period, even if not utilised in previous time periods.

**4.1.1 Hardware Resources**

* **Server Side**
  + ***Server Used:*** Amazon AWS
    - **Amazon Web Services** (AWS) is a subsidiary of Amazon.com that provides on-demand cloud computing platforms to individuals, companies and governments, on a paid subscription basis with a free-tier option available for 12 months. The technology allows subscribers to have at their disposal a full-fledged virtual cluster of computers, available all the time, through the internet. AWS's version of virtual computers have most of the attributes of a real computer including hardware (CPU(s) & GPU(s) for processing, local/RAM memory, hard-disk/SSD storage); a choice of operating systems; networking; and pre-loaded application software such as web servers, databases, CRM, etc. Each AWS system also virtualizes its console I/O (keyboard, display, and mouse), allowing AWS subscribers to connect to their AWS system using a modern browser. The browser acts as a window into the virtual computer, letting subscribers log-in, configure and use their virtual systems just as they would a real physical computer. They can choose to deploy their AWS systems to provide internet-based services for their own and their customers' benefit.
  + ***Processor:*** Duel Core 1.5 GHz
  + ***RAM:*** 4 GB
  + ***Storage Space:*** 80 GB
  + ***Database Space:*** min. 20GB
* **Client Side**
  + ***Processor:*** Duel Core 1.2 GHz equivalent or more.
  + ***RAM:*** 1 GB or more
  + ***Storage Space:*** not applicable.

**4.1.2 Software Resources**

* **Server Side**
  + ***Operating System:*** Any Linux Distribution with kernel 4.4 or higher.
    - An **operating system** (OS) is system software that manages computer hardware and software resources and provides common services for computer programs. All computer programs, excluding firmware, require an operating system to function.
    - **Linux** is a name which broadly denotes a family of free and open-source software operating systems (OS) built around the Linux kernel. A Linux-based system is a modular Unix-like operating system, deriving much of its basic design from principles established in UNIX during the 1970s and 1980s. Such a system uses a monolithic kernel, the Linux kernel, which handles process control, networking, access to the peripherals, and file systems. Device drivers are either integrated directly with the kernel, or added as modules that are loaded while the system is running.
    - A **Linux distribution** (often abbreviated as distro) is an operating system made from a software collection, which is based upon the Linux kernel and, often, a package management system. Linux users usually obtain their operating system by downloading one of the Linux distributions, which are available for a wide variety of systems ranging from embedded devices (for example, OpenWrt) and personal computers (for example, Linux Mint) to powerful supercomputers (for example, Rocks Cluster Distribution). A typical Linux distribution comprises a Linux kernel, GNU tools and libraries, additional software, documentation, a window system (the most common being the X Window System), a window manager, and a desktop environment. Most of the included software is free and open-source software made available both as compiled binaries and in source code form, allowing modifications to the original software.
  + ***Database:*** MySQL Server
    - A **database** is an organized collection of data.
    - **MySQL** is an open-source relational database management system (RDBMS). Its name is combination of “My”, the name of co-founder Michael Widenius's daughter, and "SQL", the abbreviation for Structured Query Language. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation. For proprietary use, several paid editions are available, and offer additional functionality.
  + ***Web Server:*** Apache HTTP server
    - A **web server** is a computer system that processes requests via HTTP, the basic network protocol used to distribute information on the World Wide Web. The term can refer to the entire system, or specifically to the software that accepts and supervises the HTTP requests.
    - The **Apache HTTP** Server, colloquially called Apache, is free and open-source cross-platform web server software, released under the terms of Apache License 2.0. Apache is developed and maintained by an open community of developers under the auspices of the Apache Software Foundation.
  + ***Server Side Scripting language:*** PHP 5.5
    - **Server-side scripting** is a technique used in web development which involves employing scripts on a web server which produce a response customized for each user's (client's) request to the website.
    - **PHP** is a server-side scripting language designed primarily for web development but also used as a general-purpose programming language. Originally created by Rasmus Lerdorf in 1994.
  + ***Web Scraping:*** Python v3
    - **Web scraping** (web harvesting or web data extraction) is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.
    - **Python** is a widely used high-level programming language for general-purpose programming, created by Guido van Rossum and first released in 1991. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++ or Java. It provides constructs that enable clear programming on both small and large scales.
  + ***Server Connection:*** OpenSSH server
    - **Secure Shell** (SSH) is a cryptographic network protocol for operating network services securely over an unsecured network. The best known example application is for remote login to computer systems by users. SSH provides a secure channel over an unsecured network in a client-server architecture, connecting an SSH client application with an SSH server. Common applications include remote command-line login and remote command execution, but any network service can be secured with SSH. The protocol specification distinguishes between two major versions, referred to as SSH-1 and SSH-2.
    - **OpenSSH** is the premier connectivity tool for remote login with the SSH protocol. It encrypts all traffic to eliminate eavesdropping, connection hijacking, and other attacks. In addition, OpenSSH provides a large suite of secure tunnelling capabilities, several authentication methods, and sophisticated configuration options.
  + ***Version control:*** Git
    - **Version control** is a system that records changes to a file or set of files over time so that you can recall specific versions later.
    - **Git** is a version control system for tracking changes in computer files and coordinating work on those files among multiple people. It is primarily used for source code management in software development, but it can be used to keep track of changes in any set of files. As a distributed revision control system it is aimed at speed, data integrity, and support for distributed, non-linear workflows.
* **Developer Side**
  + ***Operating System:*** Any OS with Network connection
  + ***Database Client:*** MySQL client
  + ***Server Connection software:*** SSH client
  + ***SFTP Client:*** SSH data transfer client (like WinSCP)
    - In computing, the **SSH File Transfer Protocol** (also Secure File Transfer Protocol or SFTP) is a network protocol that provides file access, file transfer, and file management over any reliable data stream. It was designed by the Internet Engineering Task Force (IETF) as an extension of the Secure Shell protocol (SSH) version 2.0 to provide secure file transfer capabilities.
    - **WinSCP** (Windows Secure Copy) is a free and open-source SFTP, FTP, WebDAV and SCP client for Microsoft Windows. Its main function is secure file transfer between a local and a remote computer. Beyond this, WinSCP offers basic file manager and file synchronization functionality. For secure transfers, it uses Secure Shell (SSH) and supports the SCP protocol in addition to SFTP.
  + ***Version control:*** Git
* **Client Side**
  + ***Operating System:*** Any OS with latest browser support.
  + ***Web Browser:*** Any Browser with HTML5 or higher version support.
    - A **web browser** (commonly referred to as a browser) is a software application for retrieving, presenting and traversing information resources on the World Wide Web. An information resource is identified by a Uniform Resource Identifier (URI/URL) that may be a web page, image, video or other piece of content. Hyperlinks present in resources enable users easily to navigate their browsers to related resources.

**4.2 Work Done**

* Configured the AWS ES2 server
* Configured the database server
* Configured the HTTP server
* Installed and Configured
  + Python version 3
  + SSH server
  + Git
* Scraped the data from the website [www.legalcrystal.com](http://www.legalcrystal.com/) and stored in the MySQL database. With details like:
  + Citation
  + Court
  + Decided On
  + Judge
  + Appellant
  + Report
  + Acts
  + History
  + Excerpt
  + Judgment
* Scraped the Lawyers data
* Scraped the Acts data
* Scraped the cases records data
* More the 0.1 million case details have been collected.
* Code of the project is available on
  + <https://github.com/vidit-agarwal/Legal-Crystal-Scrapping>
  + <https://github.com/vidit-agarwal/Acquittal-Search>

**CHAPTER 5:**

**TESTING**

**5.1 Overview**

Testing is the major control measure used during software development. Its basic function is to detect errors in the software. During requirement analysis and design, the output is a document that is usually textual and no executable. After the coding phase, computer programs are available that can be executed for testing purpose. This implies that testing not only has to uncover errors introduced during coding, but also errors introduced during previous phase. Thus the goal of testing is to uncover the requirements, design and coding errors in the programs.

1. **Objectives of testing**

* Testing as a process of executing a program with the intent of finding errors.
* To perform testing, test cases are designed. A test case is a particular made up of artificial situation upon which a program is exposed so as to find errors. So a good test case is one that finds undiscovered errors.
* If testing is done properly, it uncovers errors and after fixing those errors we have software that is being developed according to specifications.

The above objective implies a dramatic change in viewpoint .The move counter to the commonly held view than a successful test is one in which no errors are found. In fact, our objective is to design tests that a systematically uncover different classes of errors and do so with a minimum amount of time and effort.

**b) Testing principles**

Before applying methods to design effective test cases, software engineer must understand the basic principles that guide the software testing process. Some of the most commonly followed principles are:

All test should be traceable to user requirements as the objective of testing is to uncover errors, it follows that the most severe defects (from the users point of view) are those that causes the program to fail to meet its requirements.

Tests should be planned long before the testing begins. Test planning can begin as soon as the requirement model is complete. Detailed definition of test cases can begin as soon as the design model has been sol dated. Therefore, all tests can be planned and designed before any code can be generated.

The Pareto principle applies to software testing stated simply the Pareto principle implies that 80 percent of all errors uncovered during testing will likely be traceable to 20 percent of all program modules. The problem of course, is to isolate these suspects’ modules and to thoroughly test them.

Testing should begin “in the small” and progress towards testing “in large”. The tests planned and executed generally focus on individual modules. As testing progresses, testing shifts focus in an attempt to find errors in integrated clusters of modules and ultimately in the entire system.

Exhaustive testing is not possible. The number of paths permutations for impossible to execute every combination of paths during testing. It is possible however to adequately cover program logic and to ensure that all conditions in the procedural design have been exercised.

To be most effective, an independent third party should conduct testing. By “most effective”, we mean testing that has the highest probability of finding errors (the primary objective of testing).

**c) Test information flow**

Testing is a complete process. For testing we need two types of inputs:

* Software configuration
  + It includes software requirement specification, design specification and source code of program. Software configuration is required so that testers know what is to be expected and tested.
* Test configuration
  + It is basically test plan and procedure. Test configuration is testing plan that is, the way how the testing will be conducted on the system. It specifies the test cases and their expected value. It also specifies if any tools for testing are to be used.

Test cases are required to know what specific situations need to be tested. When tests are evaluated, test results are compared with actual results and if there is some error, then debugging is done to correct the error. Testing is a way to know about quality and reliability. Error rate that is the occurrence of errors is evaluated. This data can be used to predict the occurrence of errors in future.

A software product can be tested in two ways:-

In first approach, only overall functioning of the product is tested. Inputs are given and outputs are checked. This approach is called black box testing. It does not care about the internal functioning of the product.

The other approach is called white box testing. Here the internal functioning of the product is tested. Each procedure is tested for its accuracy. It is more intensive than black box testing. But for the overall product both these techniques are crucial. There should be sufficient number of tests in both categories to test the overall product.

**d) Basic methods of Testing**

**White box testing**

White box testing is performed to reveal problems with the internal structure of a program. This requires the tester to have detailed knowledge of the internal structure. A common goal of white box testing is to ensure a test case exercises every path through a program. A fundamental strength that all white box strategies share is that the entire software implementation is taken into account during testing, which facilitates error detection even when software specification is vague or incomplete. The effectiveness or thoroughness of white box testing is commonly expressed in terms of test or code coverage metrics, which measure the fraction of code exercised by test cases.

**Basic Path Testing**

It is a white box technique. It was proposed by Tom McCabe. These tests guarantee to execute every statement in the program at least one time during testing. Basic set is the set of all execution paths of a procedure.

**Black Box Testing**

Black box tests are performed to access how well a program meets its requirements, looking for incorrect or missing functionality. Functional tests typically exercise code with valid or nearly valid input for which the expected output is known. This includes concepts such as ‘boundary values’.

Performance tests evaluate response time, memory usage, throughput, device utilization and execution time. Stress tests push the system to or beyond its specified limits to evaluate its robustness and error handling capabilities. Reliability tests monitor system response to representative user input, counting failures over time to measure or certify reliability.

Black box testing uncovers the following types of errors

1. Incorrect or missing functions

2. Interface errors

3. External database access

4. Performance errors

5. Initialization and termination errors

**e) What is a Black Box Testing Strategy?**

Black Box Testing is not a type of testing; it instead is a testing strategy, which does not need any knowledge of internal design or code etc. As the name "black box" suggests, no knowledge of internal logic or code structure is required. The types of testing under this strategy are totally based/focused on the testing for requirements and functionality of the work product/software application. Black box testing is sometimes also called as "Opaque Testing", "Functional/Behavioural Testing" and "Closed Box Testing".

The base of the Black box testing strategy lies in the selection of appropriate data as per functionality and testing it against the functional specifications in order to check for normal and abnormal behaviour of the system. Now a days, it is becoming common to route the Testing work to a third party as the developer of the system knows too much of the internal logic and coding of the system, which makes it unfit to test the application by the developer.

In order to implement Black Box Testing Strategy, the tester is needed to be thorough with the requirement specifications of the system and as a user, should know, how the system should behave in response to the particular action.

Various testing types that fall under the Black Box Testing strategy are: functional testing, stress testing, recovery testing, volume testing, User Acceptance Testing (also known as UAT), system testing, Sanity or Smoke testing, load testing, Usability testing, Exploratory testing, ad-hoc testing, alpha testing, beta testing etc.

These testing types are again divided in two groups:

a) Testing in which user plays a role of tester

b) User is not required.

Testing method where user is not required:-

**Functional Testing**:-

In this type of testing, the software is tested for the functional requirements. The tests are written in order to check if the application behaves as expected.

**Stress Testing**:-

The application is tested against heavy load such as complex numerical values, large number of inputs, large number of queries etc. which checks for the stress/load the applications can withstand.

**Load Testing**:-

The application is tested against heavy loads or inputs such as testing of web sites in order to find out at what point the web-site/application fails or at what point its performance degrades.

**Ad-hoc Testing**:-

This type of testing is done without any formal Test Plan or Test Case creation. Ad-hoc testing helps in deciding the scope and duration of the various other testing and it also helps testers in learning the application prior starting with any other testing.

**Exploratory Testing**:-

This testing is similar to the ad-hoc testing and is done in order to learn/explore the application.

**Usability Testing**:-

This testing is also called as ‘Testing for User-Friendliness’. This testing is done if User Interface of the application stands an important consideration and needs to be specific for the specific type of user.

**Smoke Testing**:-

This type of testing is also called sanity testing and is done in order to check if the application is ready for further major testing and is working properly without failing up to least expected level.

**Recovery Testing**:-

Recovery testing is basically done in order to check how fast and better the application can recover against any type of crash or hardware failure etc. Type or extent of recovery is specified in the requirement specifications.

**Volume Testing**:-

Volume testing is done against the efficiency of the application. Huge amount of data is processed through the application (which is being tested) in order to check the extreme limitations of the system.

Testing where user plays a role/user is required:

**User Acceptance Testing**:-

In this type of testing, the software is handed over to the user in order to find out if the software meets the user expectations and works as it is expected to.

**Alpha Testing**:-

In this type of testing, the users are invited at the development centre where they use the application and the developers note every particular input or action carried out by the user. Any type of abnormal behaviour of the system is noted and rectified by the developers.

**Beta Testing**:-

In this type of testing, the software is distributed as a beta version to the users and users test the application at their sites. As the users explore the software, in case if any exception/defect occurs that is reported to the developers.

**f) Advantages of Black Box Testing**

* More effective on larger units of code than glass box testing
* Tester needs no knowledge of implementation, including specific programming languages
* Tester and programmer are independent of each other
* Tests are done from a user's point of view
* will help to expose any ambiguities or inconsistencies in the specifications
* Test cases can be designed as soon as the specifications are complete

**g) Disadvantages of Black Box Testing**

* Only a small number of possible inputs can actually be tested, to test every possible input stream would take nearly forever
* Without clear and concise specifications, test cases are hard to design
* There may be unnecessary repetition of test inputs if the tester is not informed of test cases a programmer has already tried
* May leave many program paths untested
* Cannot be directed toward specific segments of code which may be very complex (and therefore more error prone).

**CHAPTER 6:**

**CONCLUSION AND**

**FUTURE SCOPE**

The query analysis remains a crucial task that still remains to be completed. After query analysis is completed the next step is to index the data that has been collected. This is the next challenging task. After the pages are indexed appropriately they will be ranked accordingly.

To increase more functionality and accuracy of the search results we will use machine learning along with natural language processing capabilities.

Search engine requires immense amount of data so that important information can be retrieved when query is provided. For collecting data we have used web crawler. Web crawler , sometimes called as spider is an internet bot that systematically browses the world wide web, typically for the purpose of web indexing. Crawler copies all the pages it visits for latter processing by the search engine. The web crawler retrieves the data from all the pages it has crawled and we can store the data in a server.

The next major step is query processing, this requires a lot of knowledge on compiler development. Another method of query processing is through Natural Language Processing (NLP). We can implement NLP through python by using its NLTK library. NLTK is a leading platform for building python programs to work with human language data. It provides easy to use interface to over 50 corpora and lexical resources, along with a suite of text processing libraries for classification, tokenizing, stemming, tagging, parsing and semantic reasoning.

A system is developed to search the criminal case procedings in the court. Machine Learing will be implemented for query analysis and accurate search results in future as a part of additional fucnctionality.

**CHAPTER 7:**

**APPENDIX**

**Symbols used in Use Case**

**Actor:**



**Entity:**



**Symbols used in Sequence Diagram**

**Actor:**



**Class:**

**Symbols used in E-R Diagram**

**Entity:**

**Attribute:**

**Relationship:**

**BIBILOGRAPHY**

1. “Python”, <https://www.python.org/>, 2017.
2. “Selenium WebDriver”, <http://www.seleniumhq.org/projects/webdriver/>, 2017
3. “pymysql”, <http://pymysql.readthedocs.io/en/latest/user/examples.html>, 2017
4. “Ubuntu”, <https://www.ubuntu.com/>, 2017
5. “Amazon Web Services”, <https://aws.amazon.com/>, 2017
6. “MySQL”, <https://www.mysql.com/>, 2017
7. “Apache HTTP server”, <https://httpd.apache.org/>, 2017
8. “OpenSSH”, <https://www.openssh.com/>, 2017
9. “WinSCP”, <https://winscp.net/eng/index.php>, 2017
10. “PuTTY”, <http://www.putty.org/>, 2017
11. “GitHUB”, <https://www.github.com/>, 2017
12. “GitHUB Desktop”, <https://desktop.github.com/>, 2017
13. “Database Design”, <https://en.wikipedia.org/wiki/Database_design>, 2017
14. “Web Search Engine”, <https://en.wikipedia.org/wiki/Web_search_engine>, 2017
15. “Problems in Judical System”, <https://www.mapsofindia.com/my-india/society/five-major-problems-the-indian-judicial-system-is-facing>, 2017
16. “SFTP”, <https://en.wikipedia.org/wiki/SSH_File_Transfer_Protocol>, 2017
17. “Law”, <https://en.wikipedia.org/wiki/Law>, 2017
18. “Legal Crystal”, <https://www.legalcrystal.com/>, 2017
19. HTML and CSS: Design and Build Websites: Jon Duckett