International Conference on Latest Advancements & Future Trends In Engineering, Science & Management
Sambhram Institute of Technology, Bangalore (India)
03rd - 04th May 2019 www.conferenceworld.in
SAMBHRAM INSTITUTE OF TECHNOLOGY
ISBN : 978-93-87793-86-6

Data Extraction Using Text Mining

Akshitha Shetty P[1], Sumana[2], Sushmitha K Shetty[3], Ujwala B M[4].

[1]*Information Science and Engineering, Canara Engineering College, Mangalore-574219 (India)*

[2]*Information Science and Engineering, Canara Engineering College, Mangalore-574219 (India)*

[3]*Information Science and Engineering, Canara Engineering College, Mangalore-574219 (India)*

[4]*Information Science and Engineering, Canara Engineering College, Mangalore-574219 (India)*

ABSTRACT

This paper is focused on data extraction. Nowadays there is an increasing trend in the usage of computers for storing documents. As a result of it substantial volume of data is stored in the computers in the form of documents. The documents can be of any form such as structured documents, semi-structured documents and unstructured documents. The tedious task is retrieving useful information from huge volume of documents. Text mining is an inspiring research area as it tries to discover knowledge from unstructured text. This paper gives an overview of applications, issues, concepts, and tools used for text mining. In the last decades explosion of information and communication technologies has led to a whole new scenario concerning people's accessibility to new job opportunities and company's options for employing the right person for the right job. In this work, we present a set of techniques that makes the whole recruitment process more effective. Finally, it presents the results to the recruiter based on their criteria. In this case today's technology is used which basically helps the recruiter to get only the required resumes from the huge slot. The aim is to obtain a resumes from bulk of the resume pool and to extract the min data on particular user defined condition.

Keywords: Information Extraction, Knowledge Discovery from Databases, Text Mining.

I. INTRODUCTION

As we know these days, there is a huge Imbalance in recruitment of employees and the number of candidates that apply. There are many number of resumes received each day by the company for vacancy. So it is a highly tedious job to sort these out without missing on a prospective candidate. There is a chance of missing some of the resumes by checking them manually. For example, consider website naukri.com in which many people will be applying for various job description. So in this paper Text Mining Technique is implemented. In on-line recruitment systems, here typically candidates upload their resumes in the form of a document in a unstructured form, which must be considered by an expert recruiter. This incorporates a great asymmetry of resources required from candidates and recruiters, resulting in candidates uploading the same resumes in numerous HR agencies that become overwhelmed with thousands of resumes. In this work, we follow a different approach in

the resumes submission process. In the proposed system, we mandate that applicants submit their CVs in a structured way. This examines the candidate's professional qualifications and his personality.

II. LITERATURE SURVEY

In the next sections the methodology, research process and findings of the literature review are presented.

[1]Raymond J. Mooney and Un Yong Nahm, suggested a new framework for text mining based on the integration of *Information Extraction* (IE) and Knowledge Discovery from Databases (KDD), a.k.a. *data mining*. KDD and IE are both topics of significant recent interest. KDD considers the application of statistical and machine-learning methods to discover novel relationships in large relational databases. IE concerns locating specific pieces of data in natural-language documents, thereby extracting structured information from free text. However, there has been little if any research exploring the interaction between these two important areas. In this paper, we explore the mutual benefit that the integration of IE and KDD for text mining can provide

[2]Evanthia Faliagka, Konstantinos Ramantas, Athanasios Tsakalidis, Manolis Viennas, discovered that rapid development of modern Information and Communication technologies (ICTs) in the past few years and their introduction into people's daily lives has led to new circumstances at all the environment(work,interpersonalrelations,entertainment, etc.). People have been steadily turning to the web for job seeking and career development, using web 2.0 services like LinkedIn and job search sites (Bizer, 2005). On the other hand, a lot of companies use online knowledge management systems to hire employees, exploiting the advantages of the World Wide Web. These are termed e-recruitment systems and automate the process of publishing positions and receiving CVs.

[3]Shaidah Jusoh and Hejab M. Alfawarehsaid that, In this modern culture, text is the most common vehicle for the formal exchange of information. Although extracting useful information from texts is not an easy task, it is a need of this modern life to have a business intelligent tool which is able to extract useful information as quick as possible and at a low cost. Text mining is a new and exciting research area that tries to take the challenge and produce the intelligence tool. The tool is a text mining system which has the capability to analyze large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information [1]. The aim of text mining tools is to be able to answer sophisticated questions and perform text searches with an element of intelligence

[4]According to K.L.Sumathy, M.Chidambaram, said that, Nowadays there is an increasing trend in the usage of computers for storing documents. As a result of it substantial volume of data is stored in the computers in the form of documents. The documents can be of any form such as structured documents, semi-structured documents and unstructured documents. Retrieving useful information from huge volume of documents is very tedious task. Text mining is an inspiring research area as it tries to discover knowledge from unstructured text.

[5]Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil, said that, Nowadays most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. A document may contain some largely unstructured text components like abstract

additionally few structured fields as title, name of authors, date of publication, category, and so on. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. The great deal of studies done on the modelling and implementation of semi structured data in recent database research. On the basis of these researches information retrieval techniques such as text indexing methods have been developed to handle unstructured documents. In traditional search the user is typically look for already known terms and has been written by someone else. The problem is in result as it is not relevant to users need. This is the goal of text mining to discover unknown information which is not known and yet not written down.

[6]Samiddha Mukherjee, Ravi Shaw, Nilanjan Halder, Satyasaran Changdar, said that, In layman terms Data-mining can be related to human cognitive mind where based on previous knowledge and experience we can relate things happening around us or sometimes even predict the future. Data mining is a process of searching data from a pool of data like database, web-servers, cloud based servers etc. and provide a pattern or relationships among those data to produce desired information. This paper conducts a formal review of the concept of data-mining, the standard tasks involve in data-mining, its applications in day to day field, techniques and methodology.

[7]According to author Ramzan Talib, Muhammad Kashif Hanify, Rapid progress in digital data acquisition techniques have led to huge volume of data. More than 80 percent of today's data is composed of unstructured or semi-structured data. The discovery of appropriate patterns and trends to analyse the text documents from massive volume of data is a big issue. Text mining is a process of extracting interesting and nontrivial patterns from huge amount of text documents. There exist different techniques and tools to mine the text and discover valuable information for future prediction and decision making process. The selection of right and appropriate text mining technique helps to enhance the speed and decreases the time and effort required to extract valuable information. This paper briefly discusses and analyze the text mining techniques and their applications in diverse fields of life. Moreover, the issues in the field of text mining that affect the accuracy and relevance of results are identified.

III. METHODLOGY

In the contemporary world the text is the most common means for exchanging information. The data stored in the computer can be in any one of the form (i) structured (ii) semi structured and (iii) unstructured. The data stored in databases is an example for structured datasets. The examples for semi structured and unstructured data sets include emails, full text documents and HTML files etc. Huge amount of data today are stored in text databases and not in structured databases. Text Mining is defined as the process of discovering hidden, useful and interesting pattern from unstructured text documents. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining. Approximately 80% percent of the corporate data is in unstructured format. The information retrieval from unstructured text is very complex as it contains massive information which requires specific processing methods and algorithms to extract useful patterns. As

the most likely form of storing information is text, text mining is considered to have a high value than that of data mining. Text mining is an interdisciplinary field which incorporates data mining, web mining, information retrieval, information extraction.

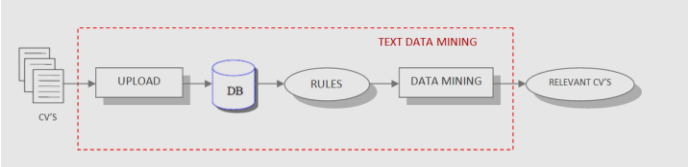


Figure1: Architecture of Text Data Mining

The CVs are sent by the candidates to the companies and these CVs are uploaded to company's Database. By following some specific requirements for the company CVs are filtered accordingly using text mining. Then the CVs are segregated using data mining and the relevant CVs are accessed by the admin as shown in Figure1.

IV. IMPLEMENTATION APPROACH

We implement front end using Java Server Pages (JSP) is a technology that helps software developers create dynamically generated web pages based on HTML, XML or other document types. It is released in 1999 by Sun Microsystems. JSP is similar to PHP and ASP, but it uses the java programming language. To deploy and run Java Server Pages, a compatible web server with a servlet container, such as Apache Tomcat or Jetty is required. Architecturally JSP may be viewed as a high-level abstraction of Java Servlets. JSPs are translated into servlets at run time, therefore JSP is a servlet. Each JSP servlet is cached and reused until the original JSP is modified. JSP can be used independently or as the view component of a server-side-model-view-controller design, normally with JavaBeans as the model and Java servlets as the controller. JSP allows java code and certain predefined actions to be interleaved with static web mark-up content, such as HTML, with the resulting page being compiled and executed on the server to deliver a document.

JSPs are usually used to deliver HTML and XML documents but the through the use of Output Stream they can deliver other type of data as well. JSP compiler is a program that parses JSPs and transforms them into executable Java Servlets. A program of this type is usually embedded into the application server and run automatically the first time a JSP is accessed, but pages may also be precompiled for better performance or compiled as a part of the build process to test for errors.

JSP is used in the project because it comes loaded with features like:

- Portable
- Robust
- More secure
- Easier to read data from user
- Easier to display server response
- Easier to connect to the database

V. SOFTWARE DESIGNAND IMPLEMENTATION

NETBEANS

The NetBeans Platform is a framework for simplifying the development of Java Swing desktop applications. The NetBeans IDE bundle for Java SE contains what is needed to start developing Net Beans plugins and NetBeans Platform based applications; no additional SDK is required. Applications can install modules dynamically. Any application can include the Update Center module to allow users of the application to download digitally signed upgrades and new features directly into the running application. Reinstalling an upgrade or a new release does not force users to download the entire application again.

The platform offers reusable services common to desktop applications, allowing developers to focus on the logic specific to their application. Among the features of the platform are:

- User interface management (e.g. menus and toolbars)
- User settings management
- Storage management (carries out efficient storage)
- Window management
- Wizard framework (supports step-by-step dialogs)
- NetBeans Visual Library
- Integrated development tools

SQLYOG ENTERPRISE

SQLyog is a GUI tool for the RDBMSMySQL. It is developed by Webyog, Inc. based in Bangalore, India and Santa Clara California. SQLyog is being used by more than 30,000 customers worldwide and has been downloaded more than 2,000,000 times.

SQLyog was first released to the public in 2001 as after eight months of development. SQLyog was available free of charge, but with closed source code, until when it was made fully commercial software. Nowadays SQLyog is distributed both as free software as well as several paid, proprietary versions.

Prominent features of SQLyog are:

- Editor with syntax highlighting and various automatic formatting options.
- Intelligent Code Completion.
- Visual Schema Designer.
- Visual Query Builder.
- Query Formatter.
- Foreign key lookup.
- Visual Data Compare

JAVA

Java is a general-purposecomputer-programming language that is concurrent, class-based, object-oriented and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA) meaning that compiled Java code can run on all platforms that support Java without the need for recompilation. Java applications are typically compiled to "byte code" that can run on any Java virtual machine (JVM) regardless of the underlying computer architecture.

Java was originally developed by a Canadian James Gosling at Sun Microsystems (which has since been acquired by Oracle) and released in 1995 as a core component of Sun Microsystems' Java platform.

There were five primary goals in the creation of the Java language:

- It must be "simple, object-oriented, and familiar".
- It must be "robust and secure".
- It must be "architecture-neutral and portable".
- It must execute with "high performance".
- It must be "interpreted, threaded, and dynamic".

JSP

Michael A. Jackson originally developed JSP in the 1970s. He documented the system in his 1975 book Principles of Program Design. Jackson structured programming (JSP) is a method for structured programming based on correspondences between data stream structure and program structure. JSP structures programs and data in terms of sequences, iterations and selections, and as a consequence it is applied when designing a program's detailed control structure. The method applies to processing of any data structure or data stream that is describable as a hierarchical structure of sequential, optional and iterated elements.

JAVA SERVLET

A Java servlet is a Javasoftware component that extends the capabilities of a server. Although servlets can respond to any types of requests, they most commonly implement web containers for hosting web applications on web servers and thus qualify as a server-side servlet web API.

A Java servlet processes or stores a Java class in Java EE that conforms to the Java Servlet API, a standard for implementing Java classes that respond to requests. Servlets could in principle communicate over any client–server protocol, but they are most often used with the HTTP. Thus "servlet" is often used as shorthand for "HTTP servlet".Thus, a software developer may use a servlet to add dynamic content to a web server using the Java platform. The generated content is commonly HTML, but may be other data such as XML and more commonly, JSON. Servlets can maintain state in session variables across many server transactions by using HTTP cookies, or URL mapping.

A Servlet is an object that receives a request and generates a response based on that request. The basic Servlet package defines Java objects to represent servlet requests and responses, as well as objects to reflect the servlet's configuration parameters and execution environment. The package javax.servlet.http defines HTTP-specific

subclasses of the generic servlet elements, including session management objects that track multiple requests and responses between the web server and a client. Servlets may be packaged in a WAR file as a web application.

MYSQL

MySQL is an open sourcerelational database management system (RDBMs). MySQL is free and open-source software under the terms of the GNU General Public License, and is also available under a variety of proprietary licenses.

MySQL is offered under two different editions: the open source MySQL Community Server and the proprietary Enterprise Server.

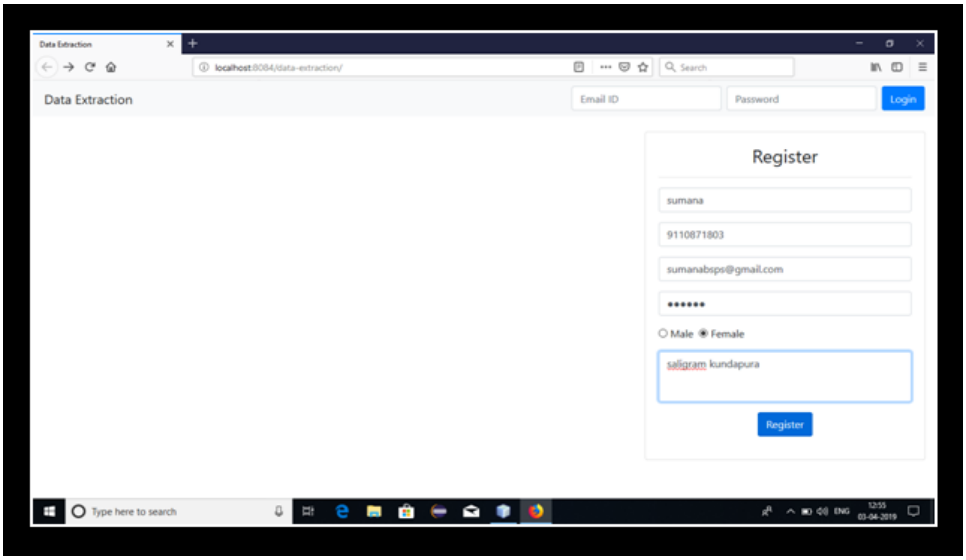
Major features of MySQL:

- Cross-platform support.
- Triggers
- Cursors
- Updatable views
- Information Schema
- Query Caching.

VI. RESULTS

REGISTRATION MODULE

In fig 2.1, the registration module is used by every user to do the registration. User need to provide name, email-id, Address and password during registration.



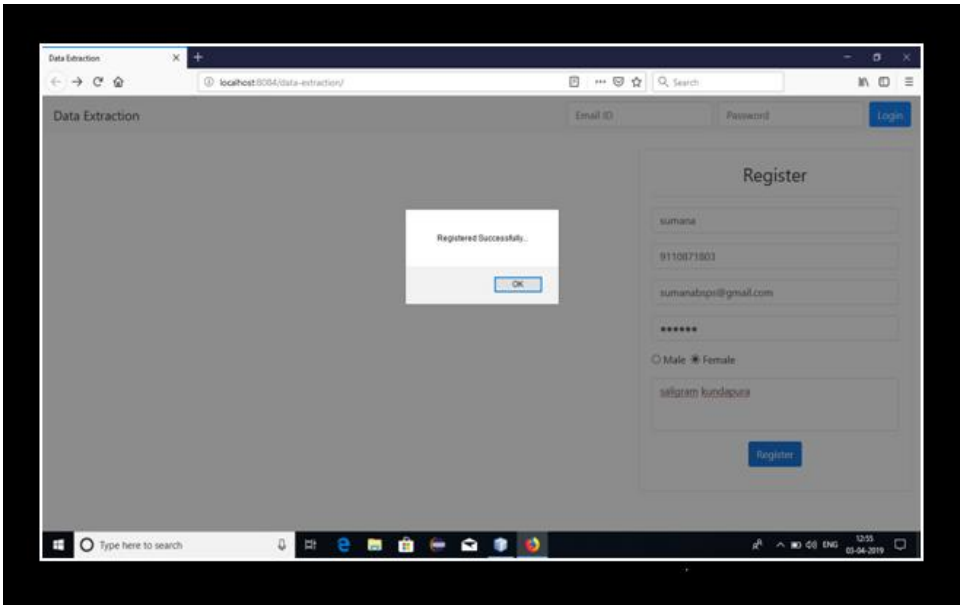


Fig 2.1 Registration Module

LOGIN MODULE

Fig 2.2 represents login module, here the user can login into the system. They need to provide username and password to login. The system consists of two kinds of user. First is the admin who controls other users in the system. Second is the user who is using the system.

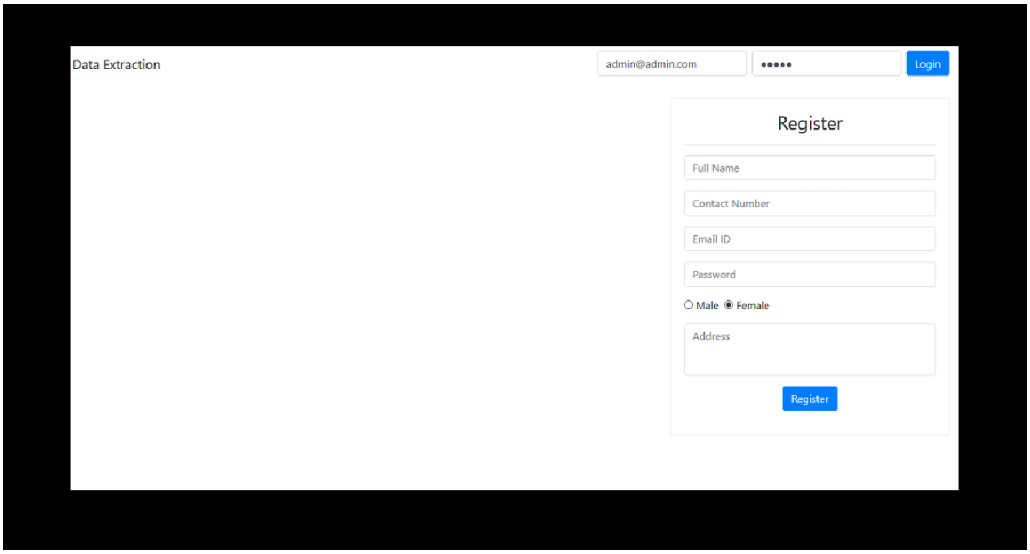


Fig 2.2 Login Module

ADD FILE

Once the registration is successful, the registered user can login into the system. Next the user can add their resumes by clicking on add files, as shown in fig 2.3.

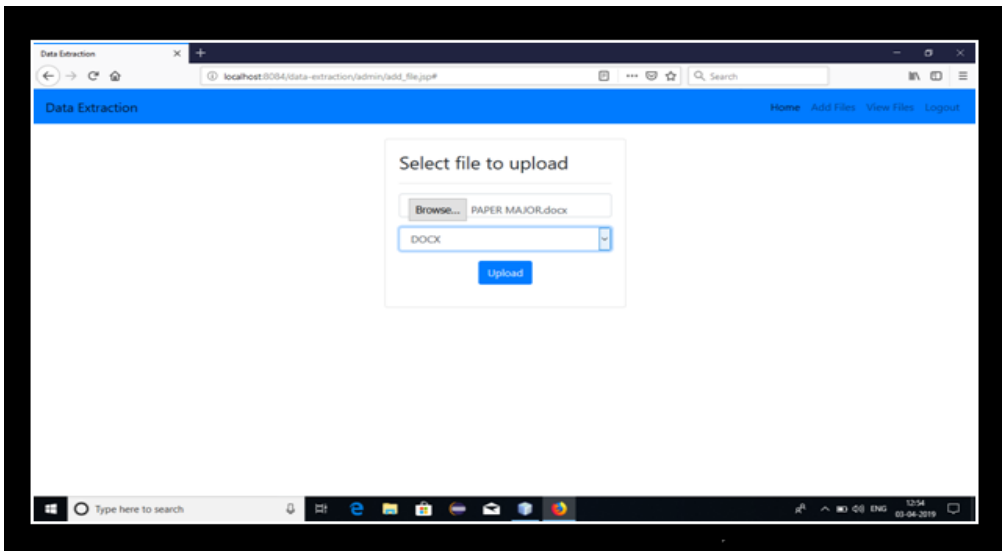
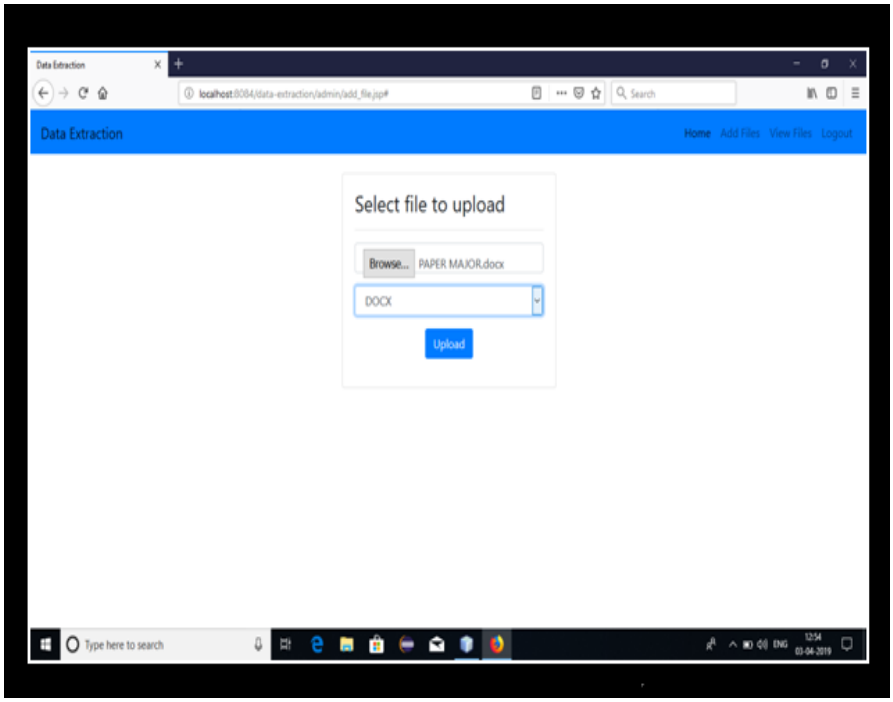


Fig 2.3 Add File

UPLOAD FILE

Here the registered user can upload their resume in the form of docx or pdf, as shown in fig 2.4.



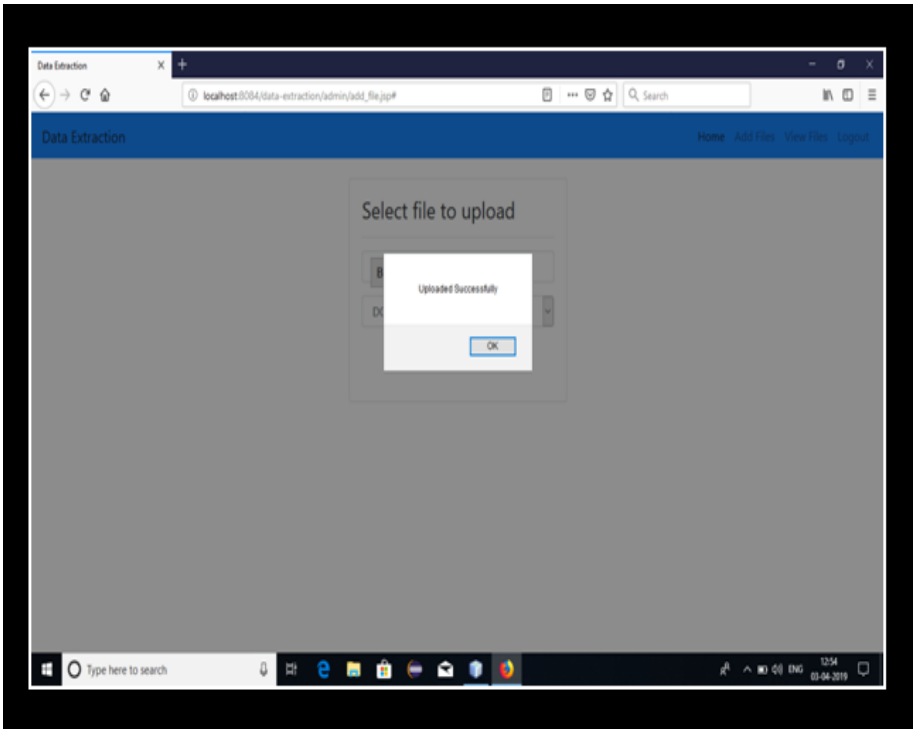


Fig 2.4 Upload File

VI.CONCLUSION

The conclusions of this paper have described the implementation of usage of Text Mining. This has reduced the execution time of a particular task and hence has been helpful in accomplishment of the objective. Here by implementing this paper, which are able to avoid the manual work & increase the efficiency of the recruitment process.

REFERENCES

[1] Ramzan Talib, Muhammad Kashif Hanify, *"Text Mining: Techniques, Applications and Issues"*, (IJACSA) International Journal of Advanced Computer Science and Applications, *Vol. 7 No. 11, 2016.*
[2] Samiddha Mukherjee, Ravi Shaw, Nilanjan Halder, Satyasaran Changdar, *"A Survey Of Data Mining Applications And Techniques"*, International Journal of Computer Applications, *Volume 6(5), 2015.*
[3] Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil, *"Text Mining Methods And Techniques"*, International Journal of Computer Applications, *Volume.85, No.17, January 2014*.
[4] K.L.Sumathy, M.Chidambaram, *"Text Mining: Concepts, Applications, Tools and Issues – An Overview"*, International Journal of Computer Applications (0975 – 8887) *Volume 80 – No.4, October 2013.*
[5] Shaidah Jusoh, Hejab M. Alfawareh, *"Techniques, Applications And Change in Issue In Text Mining"*, International Journal of Computer Applications *Volume.9, Issue.6 N0.2, November 2012*.