APPLICANT TRACKING AND SCORING SYSTEM

By

ANKIT TIWARI (35)
SAGAR VAGHELA (36)
RAHIL NAGAR (63)

SUPERVISOR

Ms. MRUNALI DESAI

Department of Computer Engineering

K. J. Somaiya Institute of Engineering and Information Technology,

Ayurvihar, Sion, Mumbai-22

2018-2019

ABSTRACT

The biggest challenge of screening resumes by far is volume. The number of resumes received is one of the biggest time-consuming factors during recruiting process. An average job opening receives 250 resumes and up to 88% of them are considered unqualified. This means a recruiter can spend up to 23 hours screening resumes for a single hire. An ATSS (Applicant Tracking and Scoring System) is a must-have software for recruitment and talent acquisition departments, because it organizes all the resumes received for each role. With the growing competition on daily basis and the growing skills of the applicants, it gets difficult for the companies to sort the resumes according to their needs and this activity is very time consuming. Therefore, instead of going through each of the applicants, the company can get a sorted list of applicants that fulfil the company requisites. ATSS is an application that caters the company for sorting the resumes and helping the HR team to shortlist the applicants based on various criteria.

TABLE OF CONTENTS

LIST OF ABBREVIATION

OCR -	Optical Character Recognition
ATSS-	Applicant tracking and scoring system
NLP -	Natural Language Processing
CNN -	Convolutional Neural Network
WEAPSS -	Web-based Employment Application & Processing Support System
HR -	Human Resource

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 PROBLEM DEFINITION:

Recruitment is an important activity in any organization or a company. It is the process of finding and hiring the best-qualified candidate for a job opening, in a timely and cost-effective manner. The recruitment process includes analyzing the job requirements, attracting employees, screening and selecting applicants, hiring, and integrating the new employee to the organization. At the same time, it is a time consuming and tedious job for the recruiters. Applicant Tracking and Scoring System tracks job applicants as they go through each stage of the hiring process. Many ATSS systems offer additional features, such as note taking, bulk emailing and job posting. An ATSS system is an applicant tracking system, or a piece of web-based software used to track job candidates as they move through the hiring process.

1.2 PROJECT OBJECTIVE:

The proposed ATSS will act as an electronic gatekeeper for an employer. The ATSS parses a resume's content into categories and then scans it for specific keywords to determine if the job application should be passed along to the recruiter. Its job is to essentially weed out unqualified applicants, so the recruiter can devote his or her time to evaluating the candidates who are more likely to be a match for the position. In other words, the ATSS is apt to find the best fit candidates, rather than identify the least qualified applicants. The main idea behind is to remove the manmade process of hiring the employees by looking into the resume and their education qualifications and then pointing and noting the things that satisfy the company criteria. The main objective is to reduce the work of the HR recruitment team. System will sort the resume of the applicants by taking their educational qualification, work experience, languages known and many different keys as per the company criteria. Points will be allotted to each key criteria and applicants will be rated as per the points acquired. It will sort the applicants in the descending order of their points which will help the recruitment team to select the applicants.

1.3 FUNCTIONAL REQUIREMENT:

The resumes are required to be in a PDF or in a word format. Most of the freshers tend to include marks of all the semesters in their resume. Since this is not in coherence with the resumes already circulating in the industry, the system requires every resume to have aggregate scores of their graduation degree.

The resumes should be in English and for the system to work in an efficient way, the applicants are suggested to use familiar words pertaining to their domain. Applicants are suggested to have all their information on their resume, as an absence of even smallest of trait can lead them to have a lower score in sorting system.

CHAPTER 2
REVIEW OF LITERATURE

Literature survey consisted of study of several papers published in reputed journals related to the topic as well as some peripheral topics to give us better knowledge and clarity of the domain. A lot of informative material was found on the internet which helped us learn more about object tracking.

2.1 OVERVIEW OF THEORETICAL STUDY:

WEAPSS[6] is a powerful online recruitment and application processing support system which can store and maintaining different types of user accounts, resumes, applications, jobs, and keeping track of the steps in the hiring process. It allows applicants to search for jobs based on different criteria and to post application. In addition, WEAPSS allows applicant to create online resumes which can be posted for multiple jobs. In the two years, following the first approval for recruitment for a position, WEAPSS allows HR staff to re-initiate the pool search and re-advertise the position for several times in case of a vacancy. The focus of WEAPSS is to streamline advertisement, hiring processes, save administrative time, eliminate redundant processes, and accelerate communication with candidates. Throughout each phase of the recruitment process, WEAPSS facilitates much more streamlined, standardized approach than the existing, paper-based recruitment process. Tasks such as sorting, coding, filing, and routing application materials which were previously performed manually, can now be performed automatically.

Alternatively, the early experimental OCR systems were often rule-based, by the eighties these have been completely replaced by systems based on statistical → Pattern Recognition[7]. For clearly segmented printed materials such techniques offer virtually error-free OCR for the most important alphabetic systems including variants of the Latin, Greek, Cyrillic, and Hebrew alphabets. However, when the number of symbols is large, as in the Chinese or Korean writing systems, or the symbols are not separated from one another, as in Arabic or Devanagari print, OCR systems are still far from the error rates of human readers, and the gap between the two is also evident when the quality of the image is compromised e.g. by fax transmission. Until these

problems are resolved, OCR cannot play the pivotal role in the transmission of cultural heritage to the digital age that it is often assumed to have. In the recognition of handprint, algorithms with successive segmentation, classification, and identification (language modeling) stages are still in the lead. For cursive handwriting, → Hidden Markov Models that make the segmentation, classification, and identification decisions in parallel have proven superior, but performance still leaves much to be desired, both because the spatial and the temporal aspects of the written signal are not necessarily in lockstep (discontinuous constituents arising e.g. at the crossing of t-s and dotting of i-s) and because the inherent variability of handwriting is far greater than that of speech, to the extent that we often see illegible handwriting but barely hear unintelligible speech. Also another software [8] states that, at the present time to getting a good job is very intricate task for any job seekers. The same problem also a company can face to acquire intelligent and qualified employees. Therefore, to minimize the problem, there are many management systems were applied and out of them, computer-based management system is one of an appropriate elucidation for this problem. In the computer management system, software is made for jobseekers to find their suitable companies and as well as made for companies for finding their suitable employees. However, the available software in the market are not intelligent based, and to make privacy, security and robustness, the software should made with the application of expert system. In this proposed study, an attempt has been made for finding the solution for job seekers and the companies with the application of expert systems.

CHAPTER 3
REQUIREMENT SPECIFICATION

3.1 INTRODUCTION:

To be used efficiently, all computer software needs certain hardware components or the other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time.

3.2 HARDWARE REQUIREMENTS:

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware.

The following subsections discuss the various aspects of hardware requirements.

- Processor: Core i3 4th Gen or higher.
- GPU: Nvidia GTX 670 or higher
- Monitor: 15” VGA Color.
- Ram: 4 GB (minimum).

3.3 SOFTWARE REQUIREMENTS:

Software requirements deal with defining software resource requirements and prerequisites that need to be installed on a computer to provide optimal functioning of an application. These requirements or prerequisites are generally not included in the software installation package and need to be installed separately before the software is installed.

Requirement Specification:

- Operating System: Windows OS (XP,7 and further versions), Mac/ Linux OS.
- Programming Language: Python 3.6.4 or higher
- Graphics Support: CUDA 7.5 or higher and CuDNN 4.0 or higher

CHAPTER 4

ANALYSIS

4.1 PROJECT TIMELINES AND TASK DISTRIBUTION:

A timeline represents the flow of the working of the project with its list of events in chronological order, also known as a project artifact. It is a graphical design that shows the dates with the tasks completed on their sides. The highlighted areas indicate the period wherein they were completed.

Table 4.1: Project Timeline

Phase	Task	Months								
		July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	March
1	Problem Definition									
	Rigorous study & Analysis									
2	Project planning									
	Designing									
3	Implementation									
4	Publication of paper									
5	Testing									

The task distribution table is given below. It is a distributed view of the individual subtasks of the entire project and the group member(s) who have completed them. The 1-year plan has been

broken down into manageable parts and helps us set measurable goals that are realistic and can be accomplished easily in small amounts of time. Table 4.2 shows this required data.

Table 4.2: Task distribution

Task List	Assigned to	Status
Requirement Gathering	Sagar Vaghela	Completed
Analysis	Ankit Tiwari	Completed
Problem Definition	Rahil Nagar	Completed
System Design	Sagar Vaghela	Completed
Documentation	Rahil Nagar	Completed
Implementation	Ankit Tiwari Sagar Vaghela Rahil Nagar	Completed
Testing	Ankit Tiwari Sagar Vaghela Rahil Nagar	Completed
Publication of paper	Ankit Tiwari Sagar Vaghela Rahil Nagar Mrunali Desai	Completed

4.2 EXISTING SYSTEM:

As the industries have grown, there hiring needs has rapidly grown. To serve this hiring needs certain consultancy units have come into existence. They offered a solution in which the candidate must upload their information in a format and submit it to the agency. Then these

agencies would search the candidates based on certain keywords. These agencies were middle level organizations between the candidate and company. These systems were not flexible as the candidate must upload there resume in a layout, and these formats changed from system to system.

4.3 PROPOSED SYSTEM:

In this proposed system, which allow the candidates to upload their resumes in flexible format. These resumes are then analysed by the system, indexed and stored in a specific format. This makes the search process easy. The analyzing system works on the algorithm that uses Natural Language Processing, a sub domain of Artificial Intelligence. It reads the resumes and understands the natural language/format created by the candidate and transforms it into a specific format. This acquired knowledge is stored in the knowledge base and scored according to the ranking attributes.

Ranking Attributes are:

- Education
- Specialization
- Location
- Total Experience
- Relevant Experience
- Resume Content Quality

These ranking attributes are customizable according to the needs of the HR.

4.4 FEASIBILITY STUDY:

Feasibility steady is an important phase in the software development process. It enables the developer to have an assessment of the product being developed. It refers to the feasibility study of the product in terms of outcomes of the product, operational use and technical support required for implementing it. Feasibility study should be performed based on various

criteria and parameters. The various feasibility studies are:

- Economic Feasibility
- Operational Feasibility
- Technical Feasibility

Economic Feasibility:

It refers to the benefits or outcomes that are deriving from the product as compared to the total cost that is spent for developing the product. In case of this project, using set of computer commands/shell. So, it becomes feasible for most of the users to access the data and understand the logic behind it. And as it is a network-based project (especially terminal based), so there will be not any economic cost associated with it. Economic analysis is the most frequently used method for evaluating the effectiveness of a new system.

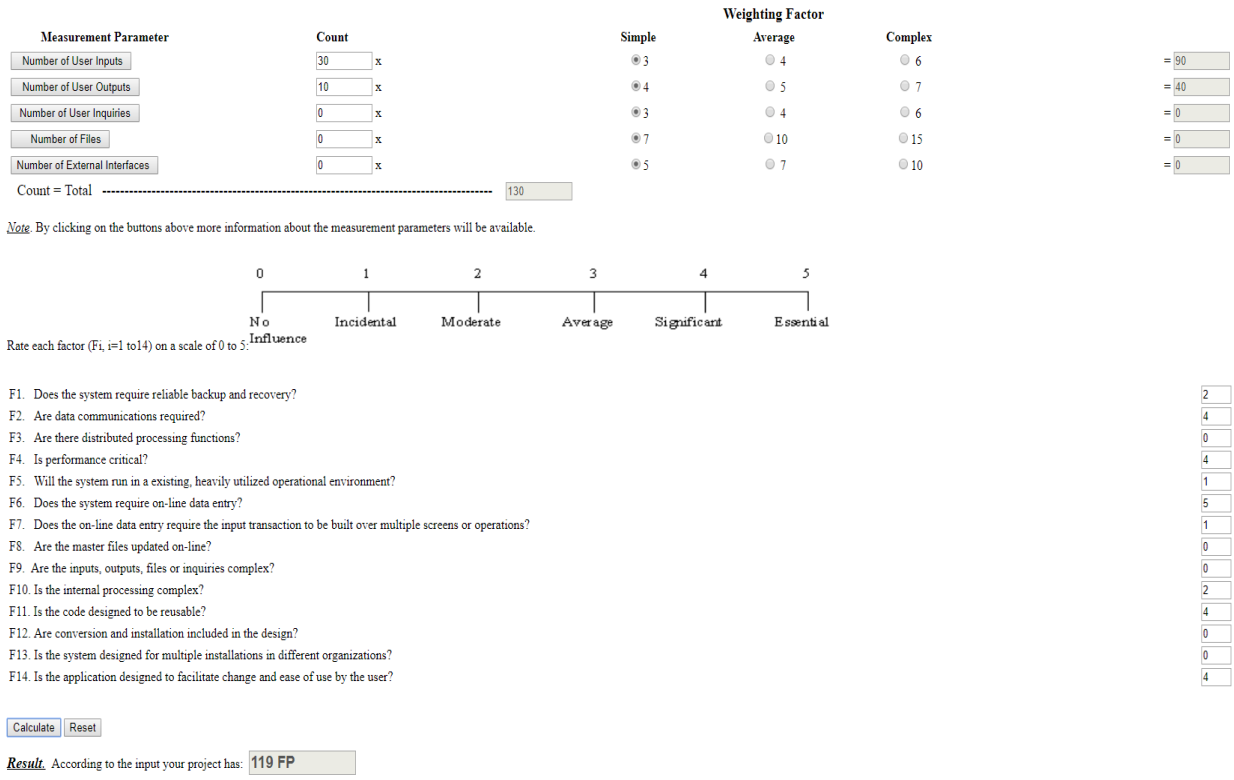


Fig. 4.1 Cost Estimation.

Operational Feasibility:

It refers to the feasibility of the product to be operational. In case of this project, it has been developed in such a way that it becomes very easy even for a person with little knowledge to operate it. This software is user friendly and does not require any technical person to operate. Thus, in this way the project is operationally feasible.

Technical Feasibility:

It refers to whether the software that is available in the market fully supports the present application. Technical Feasibility in the proposed system deals with the technology used in the system. It studies the pros and cons of using software for the development and its feasibility. In case of this project, using set of computer commands/shell. Thus, this project can be handled by any of the user who is technically sound. Hence, this project is technically feasible.

CHAPTER 5
DESIGN

The flow of the project needs to be decided in advance because of the obstacles which needs to be taken care of. The main aim is to make sure that all the disabilities which are categorized should be taken care of and needs to be handled correctly. Categorized learning disabilities can be classified into five categories. Each category of students faces different problems. These problems are tackled using Assistive technology. The learning modules that are implemented in this system focuses on these categories and help the students facing them successfully. Each module focuses on at least one category of LD. Once the learning is completed, report is generated. The generated report is analyzed and based on the report student's performance can be analyzed.

UML is a standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems. There are several goals for developing UML but the most important is to define some general-purpose modeling language, which all modelers can use, and it also needs to be made simple to understand and use. The biggest problem is how to share knowledge among end users, developers and managers. In order to upgrade quality of software development, this paper aims to present a novel operational process model that assist to share or reuse the knowledge through different stage of software engineering. The operational process model is the knowledge transformation and computing model and it will prompt the automation of software design; also enhance the standardization of knowledge representation.

5.1 ENTITY-RELATIONSHIP DIAGRAM:

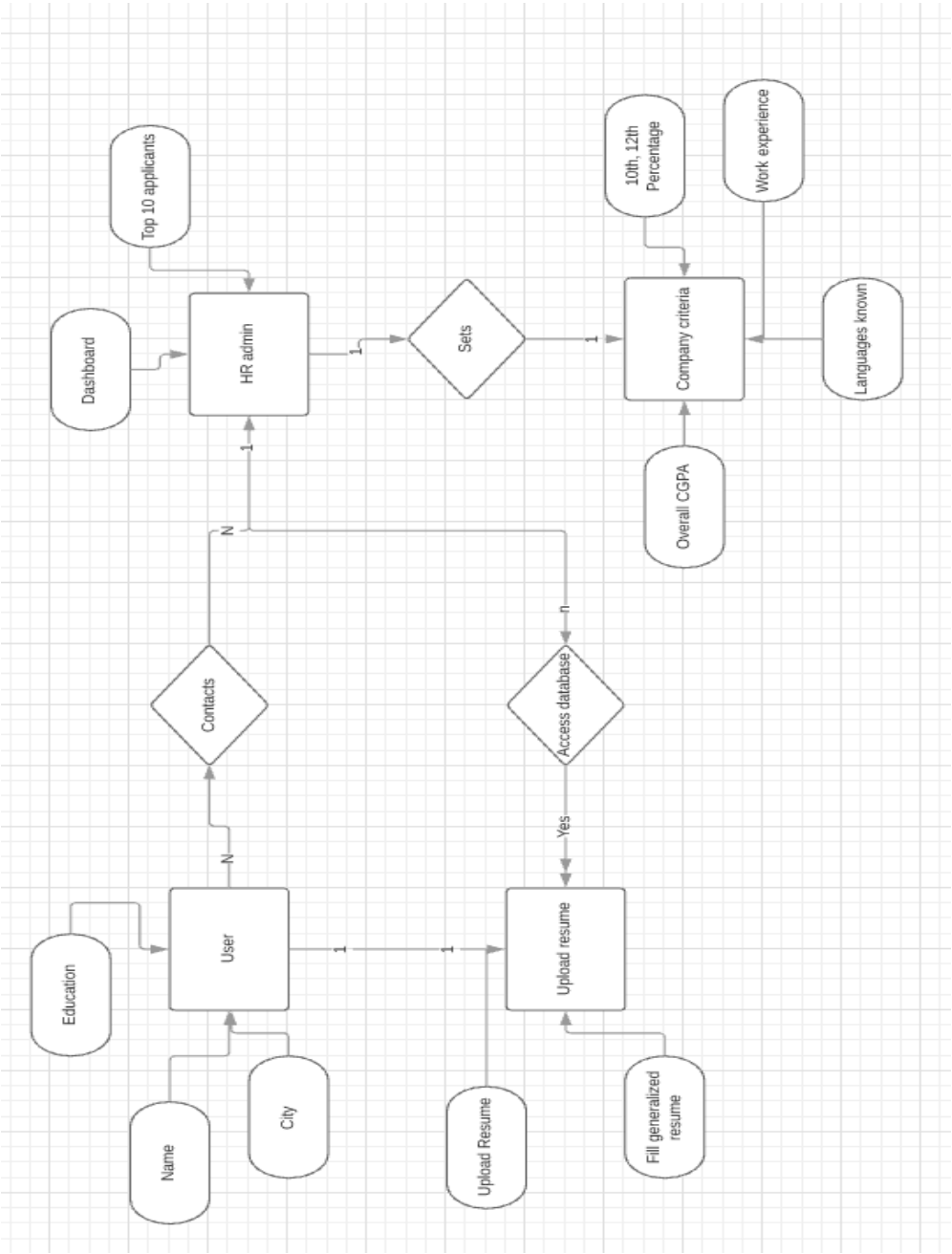


Fig. 5.1 Entity-Relationship diagram.

In the Figure 5.1, the entities are User, Resume, HR Admin and Company criteria.

5.2 OBJECT DIAGRAM:

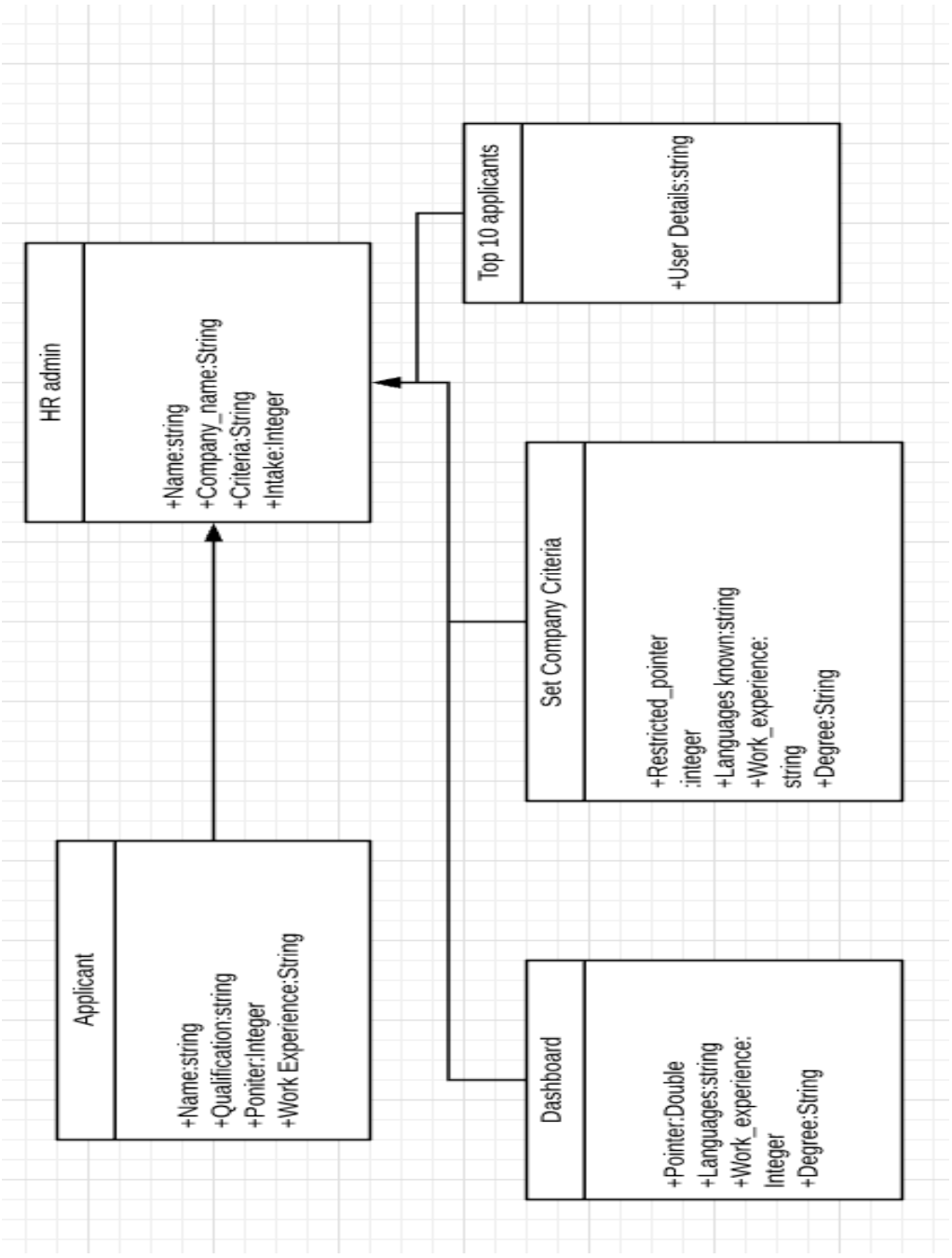


Fig. 5.2 Object diagram.

Object diagram shows the relationship between various objects of the system. In Figure 5.2, the objects are Applicant, HR Admin, Dashboard, Sorted applicants and Company Criteria

5.3 USE CASE DIAGRAM:

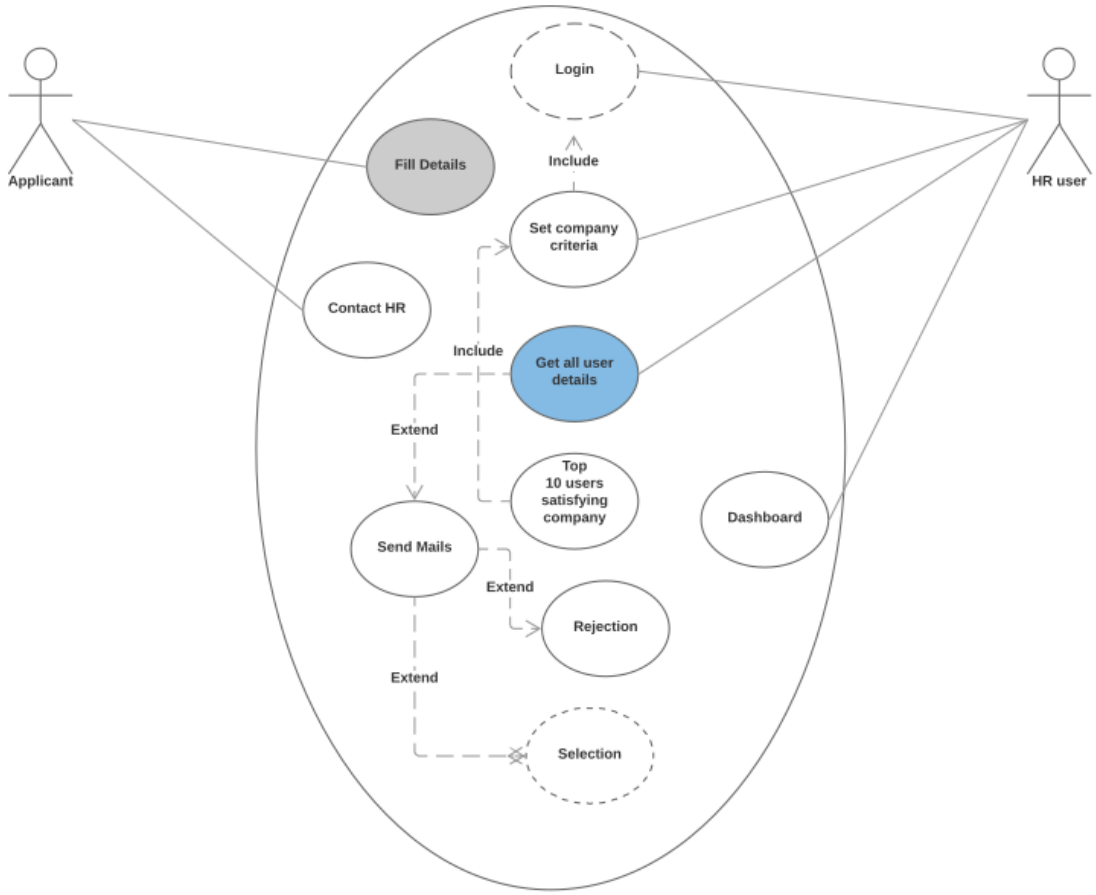


Fig. 5.3 Use-case diagram.

The Use case diagram of selected case study symbolizes the relationship between Applicant and HR User. Various Use-cases are demonstrated in the Figure 5.3

5.4 ACTIVITY DIAGRAM:

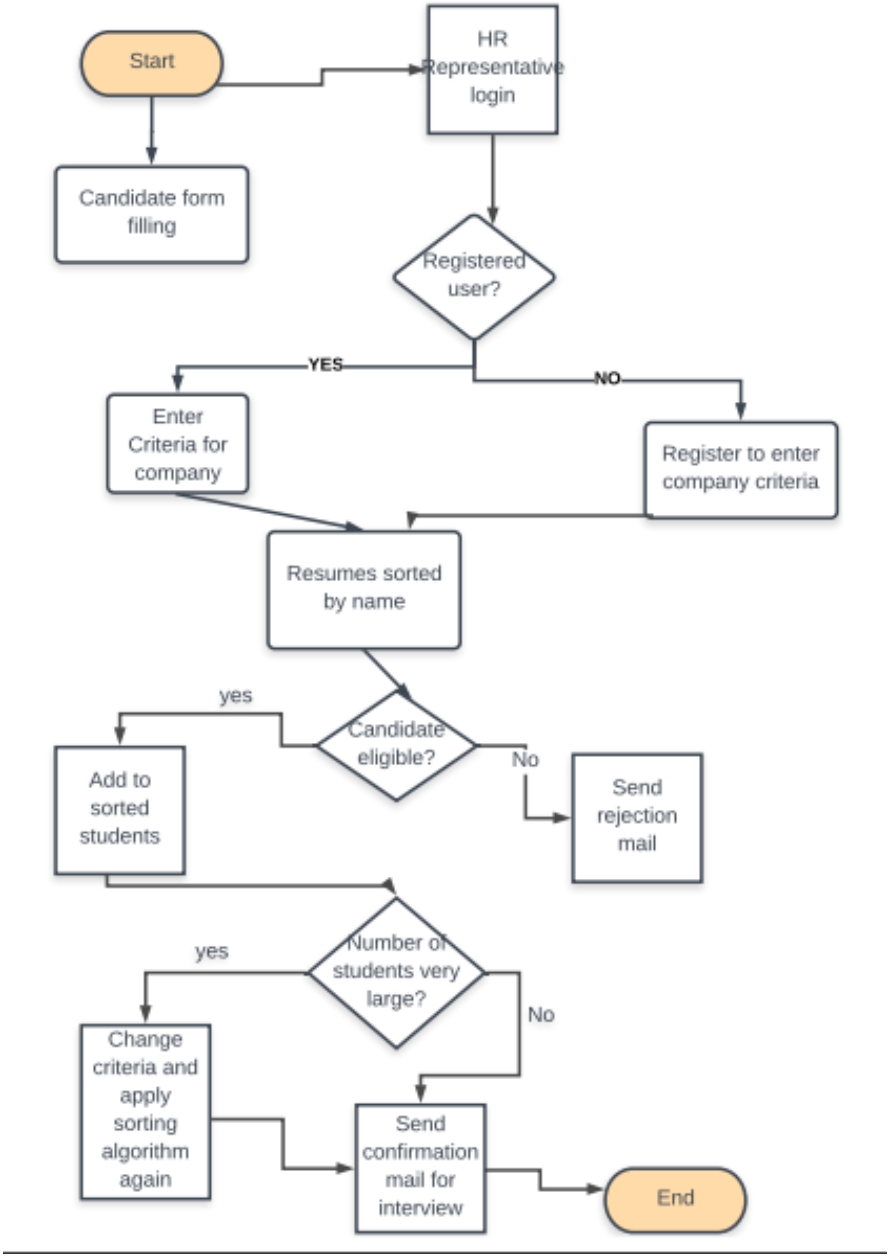


Fig. 5.4 Activity diagram.

In Figure 5.4, the Activity diagram for selected case study describes the activity flow of the system among the applicant and the HR users.

CHAPTER 6
IMPLEMENTATION AND TESTING

6.1 OVERVIEW OF MODULES:

The project is subdivided into three modules. These three modules are the three stages that the data of an applicant will have to go through before reaching its place in a list based on HR defined constraints.

The three modules are:

- Optical Character Recognition (OCR)
- Natural Language Processing (NLP)
- Applicant sorting and Dashboard interface

6.2 DESCRIPTION OF MODULES:

6.2.1 Text Extraction:

Many of the resumes that are stacked inside a company is already in a soft copy format, on which a software extraction technique can directly be applied. But to make this system more holistic, an Optical Character Recognition (also optical character reader, OCR) module is added to include the applicants who have sent a hard copy of their resume.

OCR is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image. For this project, there will be OCR to convert the digital formats like PDF, JPG etc. to a text format. Figure 6.1 shows the process of converting a hard copy document into a text blob[1] that can be fed into a Natural Language Processing (NLP) module.

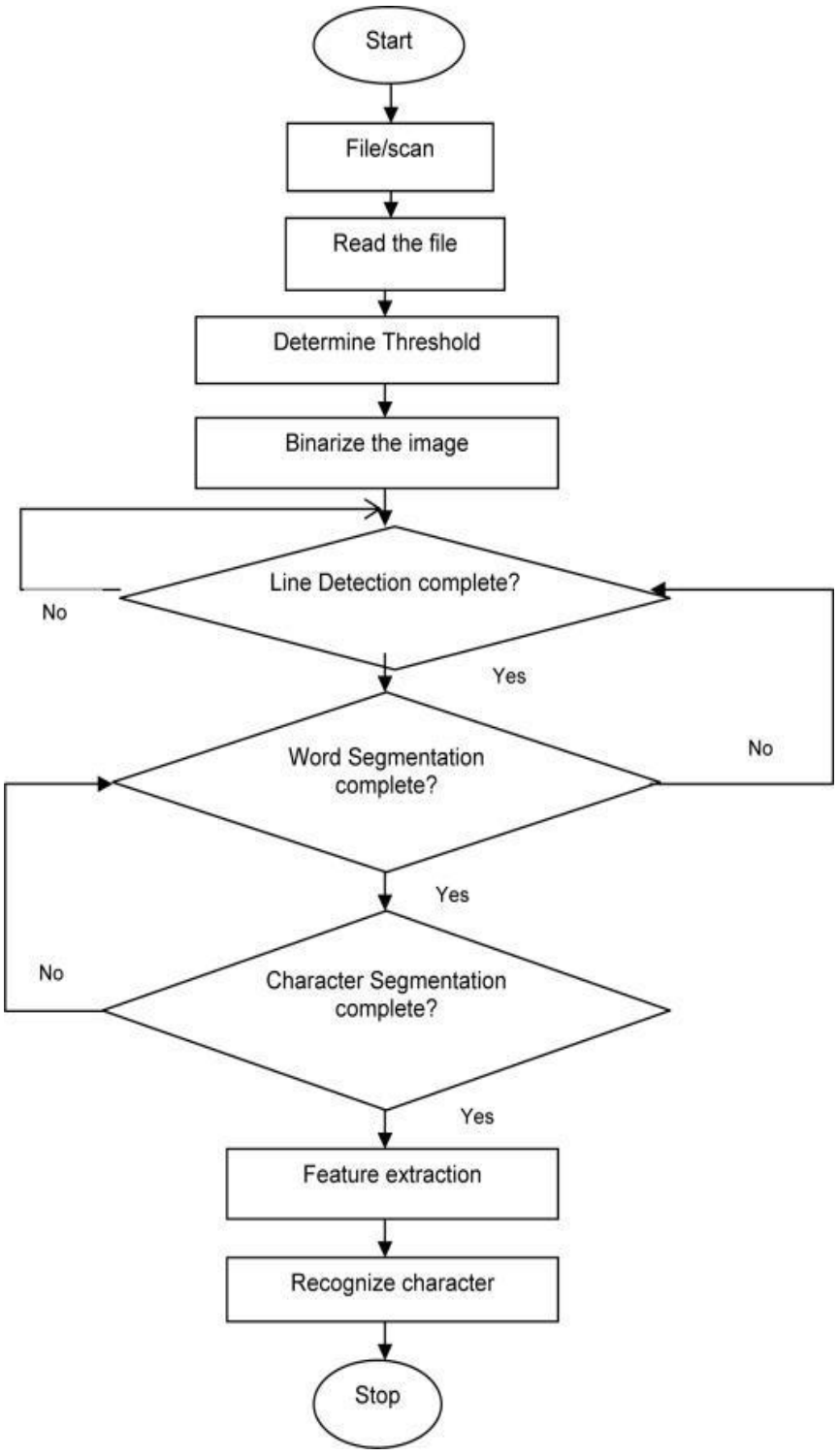


Fig. 6.1 OCR Flow chart

6.2.2 Natural Language Processing (NLP):

Natural Language Processing is an algorithm that takes the plain text as input and can convert into meaningful data.

Using NLP, system is going to parse the resume [5], NLP requires the following for parsing:

1) Lexical Analysis
2) Syntactic Analysis
3) Semantic Analysis
4) Named Entity Recognition (NER)

Lexical Analysis:

Lexical analysis is the first phase of NLP parsing, as shown in Figure 6.2, the plain text input is segmented into words and paragraphs and then the tokens are created.

Syntactic Analysis:

In Syntactic analysis the analysis of the grammar and the arrangement of words in a meaningful manner is checked, sentences like “College goes to girl” is rejected.

Semantic Analysis:

Semantic analysis checks the exact meaning of the text, sentences like “Sunny is Raining” will be rejected by the English semantic analysis.

Named Entity Recognition (NER):

One of the problems with using the same NLP module for all the companies is the jargons and words that mean something for that company’s domain and may mean something else in general. This hindrance is overcome in this system with the help of “Named Entity Recognition” or NER. A named entity is an object that exists in the real word. With NER, the system can fine tune NLP module to understand the real word objects from a domain[2][4][5]. For example, if a company wants to hire developers, they can use this system to differentiate between people who love “Python”- the programming language and the people who love “python”- the snake, based on the context in which the word is used.

THE MODEL

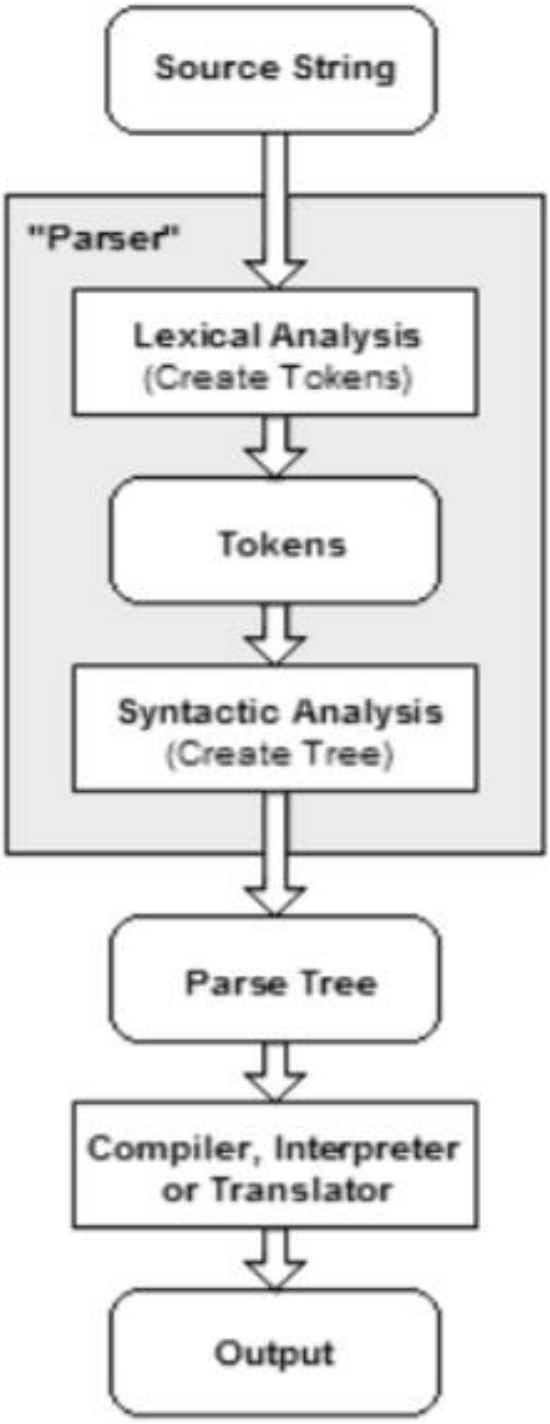


Fig. 6.2 NLP Model

6.2.2 Applicant sorting and Dashboard interface:

A web portal is provided to the HR, to define the constraints and the required skill sets of the company, on which the applicants are to be judged. After getting the output from the NLP, the data obtained will be used for a dashboard which will contain graphs and pie charts based on the data in the resumes. An HR can use this dashboard to prepare his/her query based on the requirement that the respective company has. In order to make this dashboard, all the data will be fed into ElasticSearch. Elasticsearch provides powerful tools to prepare such dashboards. ElasticSearch is included in order to make a more holistic system, which will be ready for basic level industry use. To make the system indifferent to many formats, the data to be fed will be fed by using Logstash.

Using queries inbuilt in ElasticSearch, the resumes will be scored and then they will be sorted according to the constraints that were provided by the HR. The individual traits of an applicant will be provided with a proportionate boost based on the priority of the trait. At the end, the cumulative score will be used to sort the applicants. A final sorted list of applicants will be displayed to the HR.

The process of feeding the data into the system, as said, begins with Logstash. The data is then scored and sorted using ElasticSearch. Afterwards, Kibana is used to make dashboards.

Logstash: Logstash can dynamically unify data from disparate sources and normalize the data into destinations of your choice. Cleanse and democratize all your data for diverse advanced downstream analytics and visualization use cases.

Elasticsearch: Elasticsearch provides scalable search, has near real-time search, and supports multitenancy. Elasticsearch is distributed, which means that indices can be divided into shards and each shard can have zero or more replicas. Each node hosts one or more shards, and acts as a coordinator to delegate operations to the correct shard(s).

Kibana: It provides visualization capabilities on top of the content indexed on an Elasticsearch cluster. Users can create bar, line and scatter plots, or pie charts and maps on top of large volumes of data

6.3SYSTEM ARCHITECTURE:

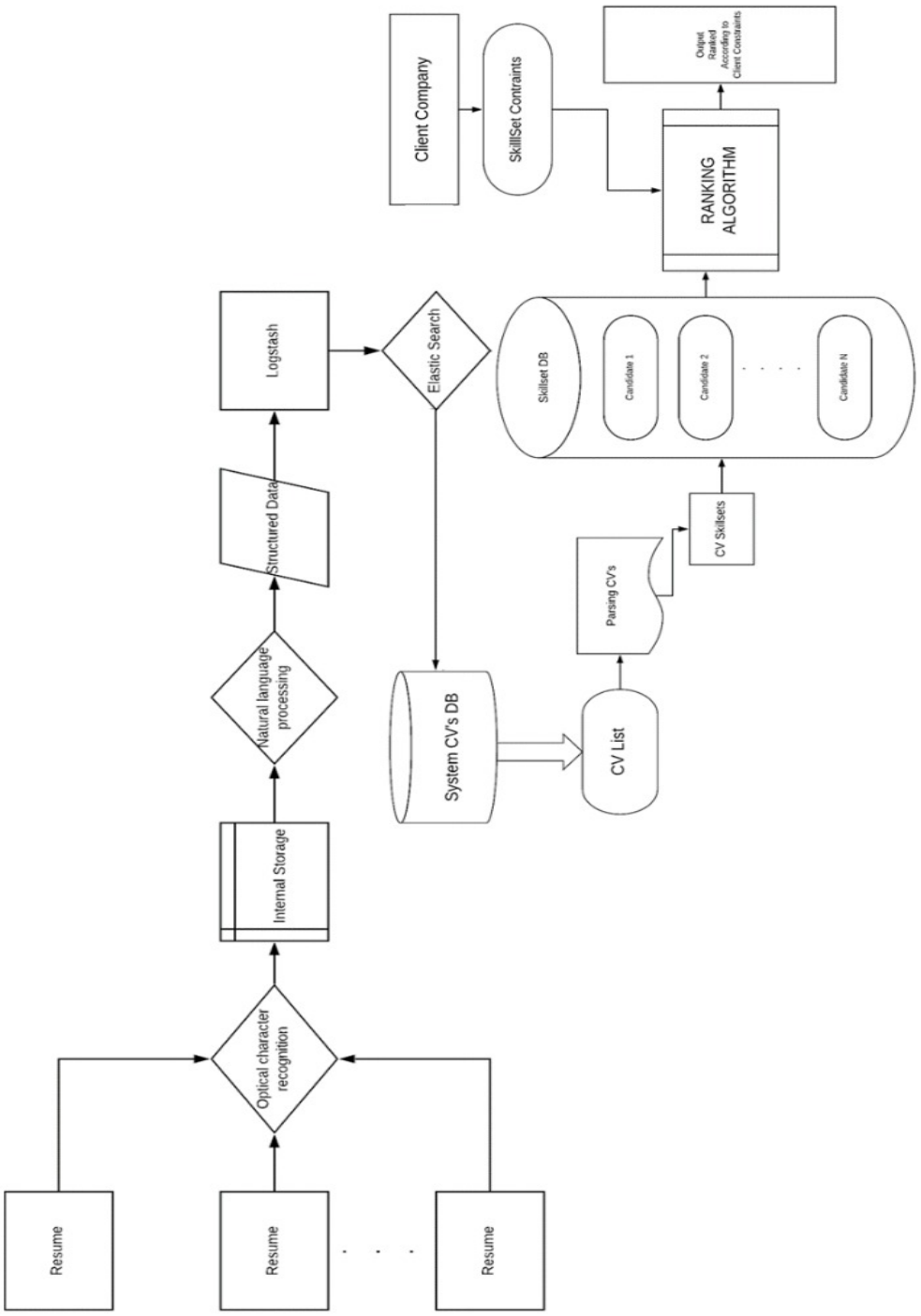


Fig. 6.3 System Architecture

6.4 SCREENSHOTS OF THE IMPLEMENTATION:

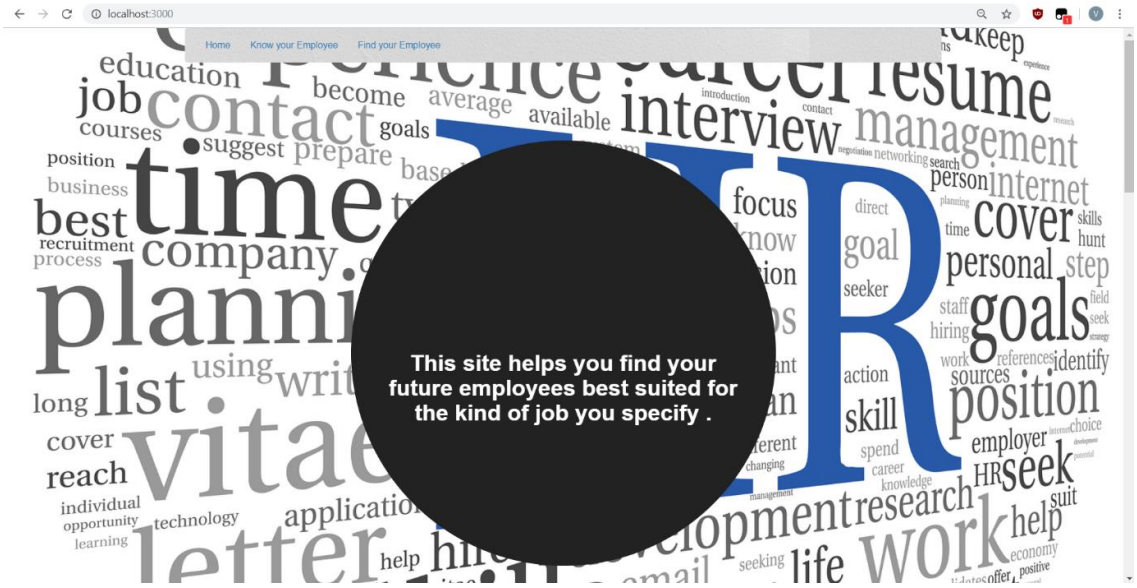


Fig. 6.4 Homepage

Figure 6.4 shows the homepage of the system for the recruiters.

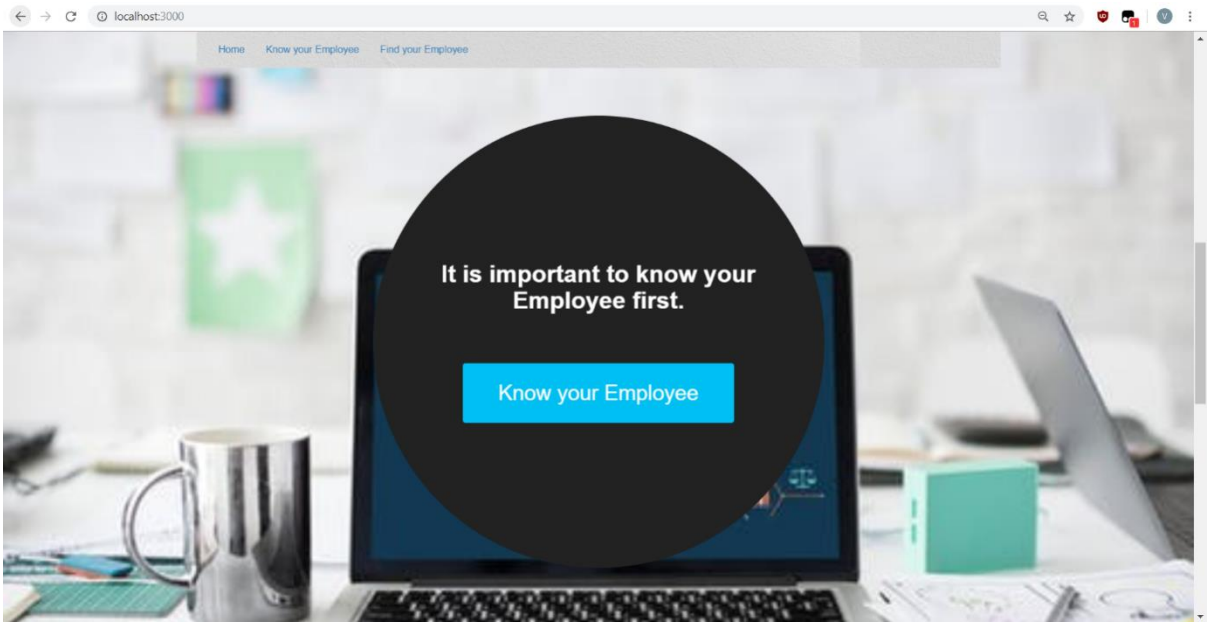


Fig. 6.5 Hyperlink to Form

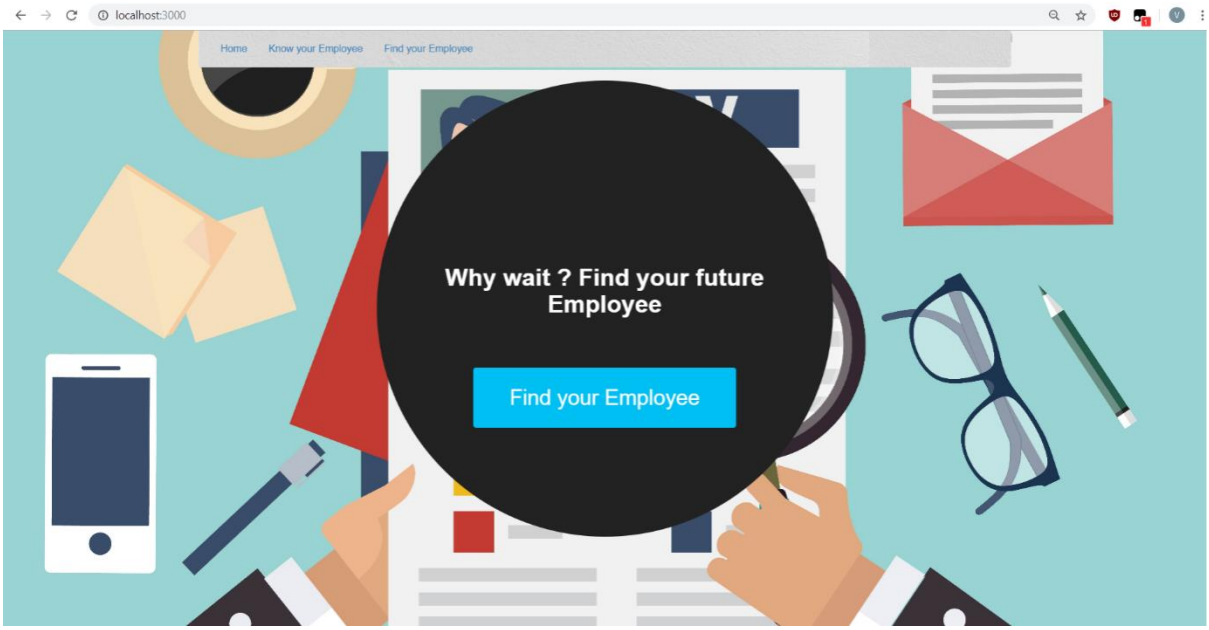


Fig. 6.6 Hyperlink to Dashboard

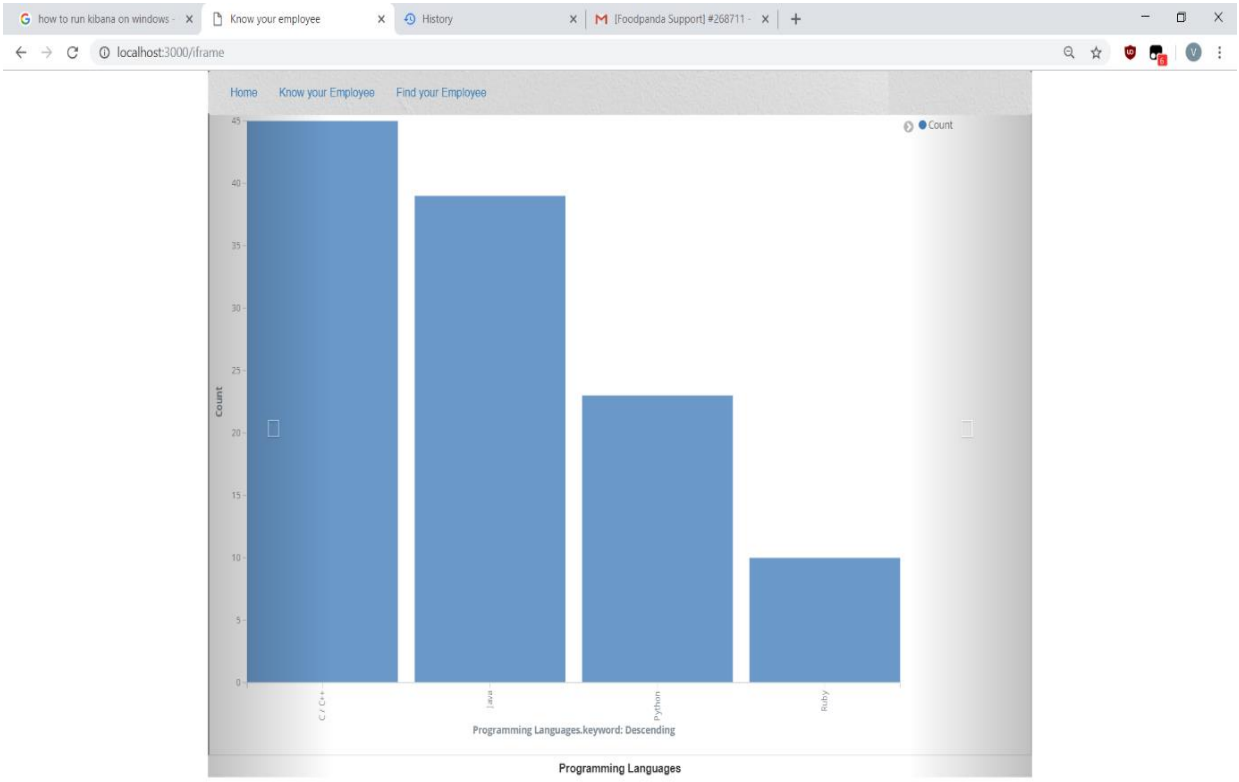


Fig. 6.7 Dashboard – Bar Graph

Figure 6.7 is a bar graph generated by the system displaying the number of applicants and their experience in different programming languages.

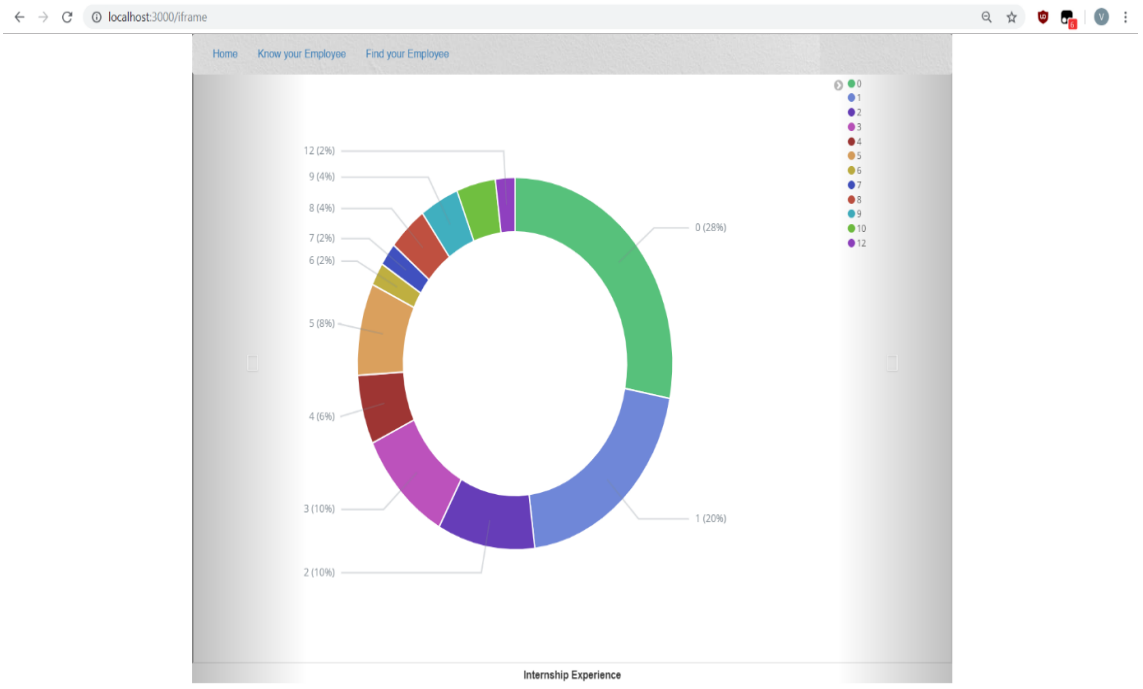


Fig. 6.8 Dashboard – Pie Chart

Figure 6.8 shows Pie chart generated by the system demonstrating years of experience of internship in industries separated by different colours according to the number of years.

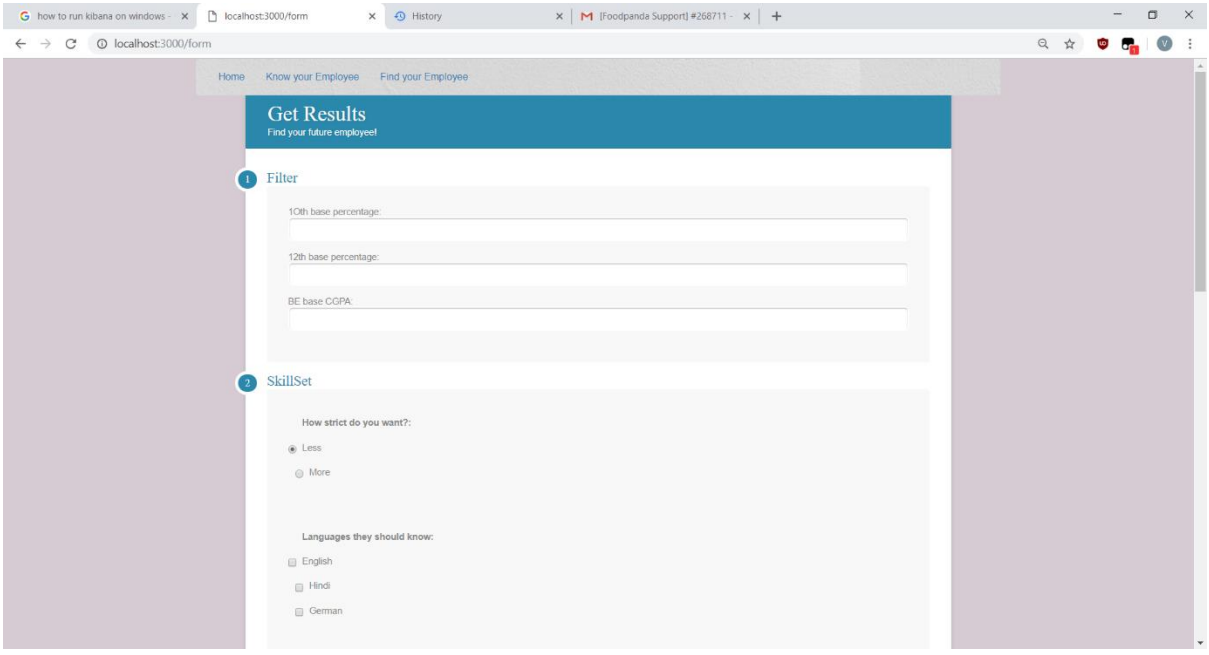


Fig. 6.9 Applicant Form – 1

Figure 6.9 is the Applicant Form for the recruiters

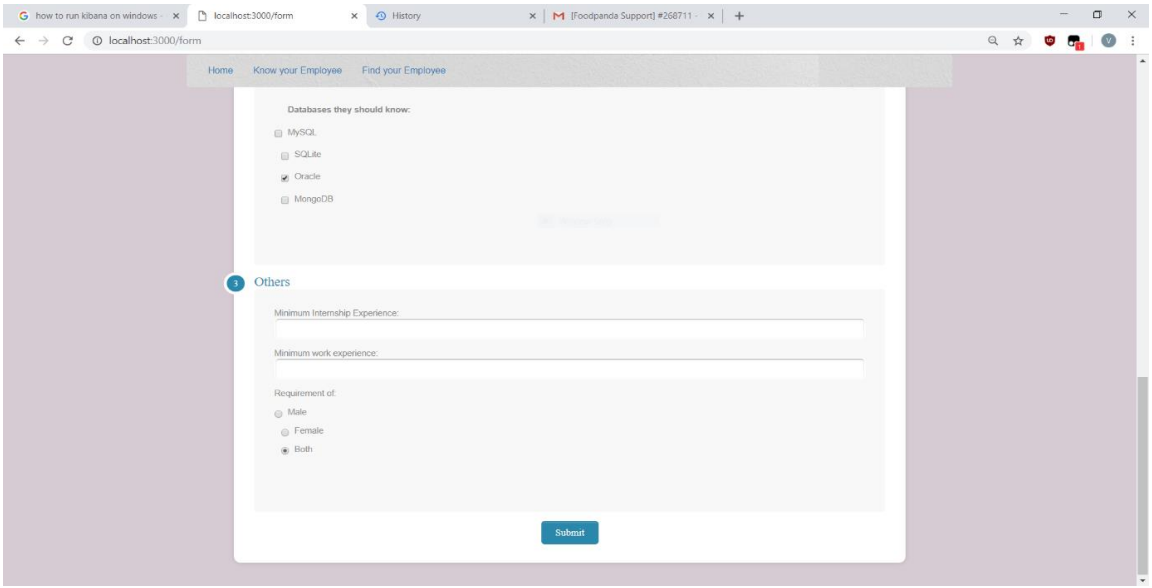


Fig. 6.10 Applicant Form - 2

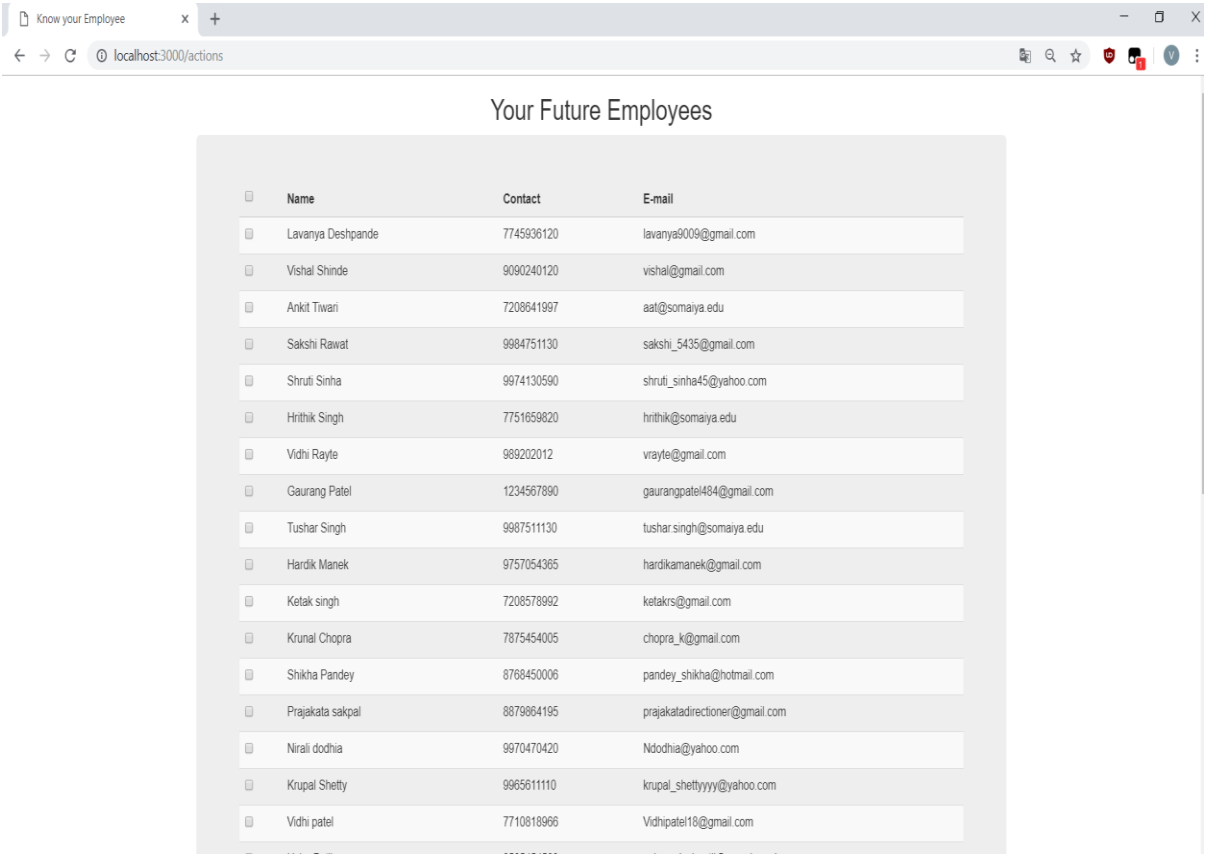


Fig. 6.11 Result – Sorted list

Figure 6.11 displays the Sorted list of the applicants according to the constrains selected by the recruiter.

6.5 TESTING:

The testing phase of the software development lifecycle (SDLC) is where you focus on investigation and discovery. During the testing phase, developers find out whether their code and programming work according to customer requirements. And while it's not possible to solve all the failures you might find during the testing phase; it is possible to use the results from this phase to reduce the number of errors within the software program.

Before testing can begin, the project team develops a test plan. The test plan includes the types of testing you'll be using, resources for testing, how the software will be tested, who should be the testers during each phase, and test scripts, which are instructions each tester uses to test the software. Test scripts ensure consistency while testing.

There are several types of testing during the test phase, including quality assurance testing (QA), system integration testing (SIT), and user acceptance testing (UAT).

Table 6.1: Test Case table

Test ID	Test	Expected result	Actual Result	Pass/Fail
1	Extract text from resumes	All the text is stored in the '.csv' format.	All the text is stored in the '.csv' format.	Pass
2	Take data from different format word, pdf, etc.	Analyse data from various formats	Data analysed From all the necessary formats.	Pass
3	Take professions from resumes	Professions to be stored in database	Different professions are stored in the database	Pass
4	Different sorting result when	Each key parameter of the	Resumes are sorted as per the	Pass

	different fields are boosted	resume decide should be boosted	boost given to the key parameters	
5	If a field is selected, it should have boosted score	Key criteria selected the by admin should have boosted score	Key criteria have boosted score selected by the admin.	Pass

CHAPTER 7
CONCLUSION

7.1 CONCLUSION:

The process of recruitment is one of the most stressful periods for both the parties, viz., the applicants and the recruiters. Corporate companies and recruitment agencies process numerous resumes daily. This is no task for humans. An automated intelligent system will take out all the vital information from the unstructured resumes and transform all of them to a common structured format which can then be ranked for a specific job position and candidate according to need.

This system aims to ease this process by making the deserving candidates, making them stand out against the crowd, which in turn, makes it easier for the recruiters. This system greatly automates the process of recruitment. The recruiters will have an idea about the quality of applicants beforehand. The applicants will be notified of the reasons because of which they were rejected, so they can improve their resume next time they submit.

Furthermore, the unfair and discriminatory practices than takes place during recruitment process can be dampened to some affect.

7.2 FUTURE SCOPE:

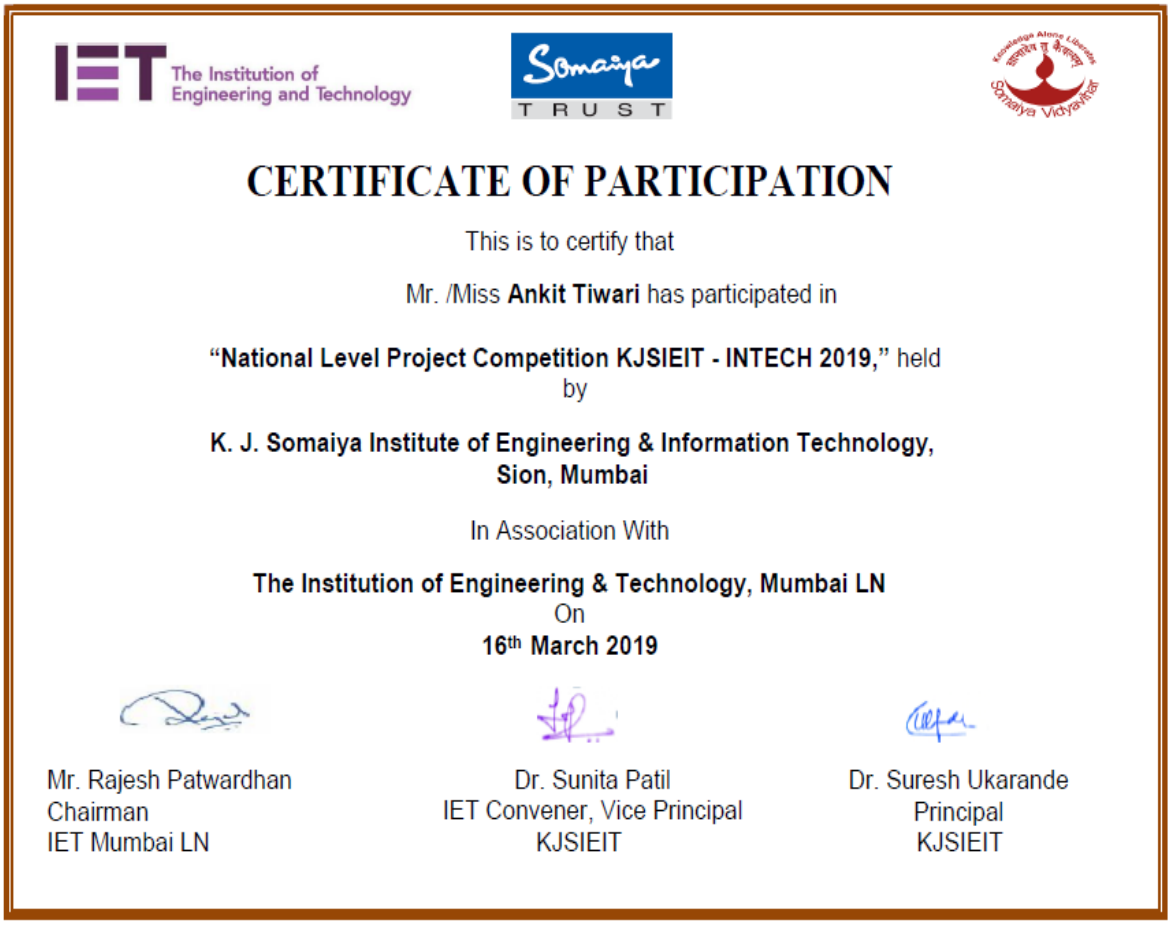
To scrape keywords from different social networking sites including Stack Overflow, LinkedIn, etc and find the similarity between them with which we could determine the genre of the resume (e.g.: Computer science, Management, Sales, human resource, etc). Future work includes ranking the resume and analysing information about the candidate from social networking sites like Facebook and Twitter so that we can decide more accurately and authentically whether to offer the candidate, a job.

CHAPTER 8

PUBLICATION AND CERTIFICATES

PUBLICATION AND CERTIFICATES

[1] Ankit Tiwari, Sagar Vaghela, Rahil Nagar, Mrunali Desai, "ATSS–Applicant tracking and scoring system", Institute of Engineering and technology (IET).



The Institution of Engineering and Technology

Somaiya TRUST

Somaiya Vidyavihar

CERTIFICATE OF PARTICIPATION

This is to certify that

Mr. /Miss **Ankit Tiwari** has participated in

"National Level Project Competition KJSIEIT - INTECH 2019," held by

K. J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai

In Association With

The Institution of Engineering & Technology, Mumbai LN

On

16th March 2019

Mr. Rajesh Patwardhan
Chairman
IET Mumbai LN

Dr. Sunita Patil
IET Convener, Vice Principal
KJSIEIT

Dr. Suresh Ukarande
Principal
KJSIEIT

CERTIFICATE OF PARTICIPATION

This is to certify that

Mr. /Miss **Sagar Vaghela** has participated in

"National Level Project Competition KJSIEIT - INTECH 2019," held by

K. J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai

In Association With

The Institution of Engineering & Technology, Mumbai LN

On

16th March 2019

Mr. Rajesh Patwardhan Chairman IET Mumbai LN	Dr. Sunita Patil IET Convener, Vice Principal KJSIEIT	Dr. Suresh Ukarande Principal KJSIEIT

CERTIFICATE OF PARTICIPATION

This is to certify that

Mr. /Miss **Rahil Nagar** has participated in

"National Level Project Competition KJSIEIT - INTECH 2019," held by

K. J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai

In Association With

The Institution of Engineering & Technology, Mumbai LN

On

16th March 2019

Mr. Rajesh Patwardhan Chairman IET Mumbai LN	Dr. Sunita Patil IET Convener, Vice Principal KJSIEIT	Dr. Suresh Ukarande Principal KJSIEIT

[2] Paper published by International Research Journal of Engineering and Technology (IRJET), www.irjet.net

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Applicant Tracking and Scoring System

Ankit Tiwari[1], Sagar Vaghela[2], Rahil Nagar[3], Mrunali Desai[4]

[1]Student, Department of Computer Engineering, K.J. Somaiya Institute of Engineering & I.T., Sion, Mumbai, Maharashtra, India
[2]Student, Department of Computer Engineering, K.J. Somaiya Institute of Engineering & I.T., Sion, Mumbai, Maharashtra, India
[3]Student, Department of Computer Engineering, K.J. Somaiya Institute of Engineering & I.T., Sion, Mumbai, Maharashtra, India
[4]Assistant Professor, Department of Computer Engineering, K.J. Somaiya Institute of Engineering & I.T., Sion, Mumbai, Maharashtra, India

---

Abstract - *Text in a resume file stored in any format is cleaned, extracted and fed to a Natural Language Processing (NLP) module. The NLP module is fine-tuned with the help of Named Entity Recognition (NER) to accommodate the varying needs of different companies and their technical jargons. The extracted, intelligent data is stored in a ".csv" format. The Human Resource Officer (HR) can define constraints on which the HR wants the applicants to be judged. These constraints are used to score each traits from the resume. At last the cumulative score is used to showcase the eligibility of the candidate to the HR.*

Key Words: NLP, NER, Resume Parser, text mining, JSON Resume.

1. PROBLEM STATEMENT

To design a system to extract the information, parse the extracted information from unstructured to structured JSON format, and rank those resumes according to the skill sets of the candidate and on the job description.

2. INTRODUCTION

The biggest challenge of screening resumes by far is its volume. The number of resumes received is one of the biggest time-consuming factors during the recruiting process. An average job opening receives 250 resumes and up to 88 per cent of them are considered unqualified. This means a recruiter can spend up to 23 hours screening resumes just for a single hire. An ATSS (Applicant Tracking and Scoring System) is a must-have software for recruitment and talent asset departments because it organizes all the resumes received for each role. With the growing competition on a daily basis and the growing skills of the applicants, it gets difficult for the companies to sort the resumes according to their needs and this activity is very time-consuming. Recruitment is an important activity in any organization or a company. Therefore, instead of going through each of the applicants, the company can get a sorted list of applicants that fulfil the company requisites. ATSS is an application that caters the company for sorting the resumes and helping the HR team to shortlist the applicants based on various criteria. Applicant Tracking and Scoring System tracks job applicants as they go through each stage of the hiring process. Many ATSS systems offer additional features, such as note taking, bulk emailing and job posting.

3. LITERATURE SURVEY

3.1 Shakya, Sujan, Web-based Employment Application Processing Support System, Master of Software Engineering, May 2008, (Dr Thomas Gendreau, Dr Kasi Periyasamy)

It is a powerful online recruitment and application processing support system which is capable of storing and maintaining different types of user accounts, resumes, applications, jobs, and keeping track of the steps in the hiring process. It allows applicants to search for jobs based on different criteria and to post application. In addition, WEAPSS allows the applicant to create online resumes which can be posted for multiple jobs. In the two years, following the first approval for recruitment for a position, WEAPSS allows HR staff to re-initiate the pool search and re-advertise the position for a number of times in case of a vacancy. The focus of WEAPSS is to streamline advertisement, hiring processes, save administrative time, eliminate redundant processes, and accelerate communication with candidates. Throughout each phase of the recruitment process, WEAPSS facilitates a much more streamlined, standardized approach than the existing, paper-based recruitment process. Tasks such as sorting, coding, filing, and routing application materials which were previously performed manually, can now be performed automatically.

3.2 Bazzi, Issam, Richard Schwartz and John Makhoul (1999) An omnifont open-vocabulary OCR system for English and Arabic. Pattern Analysis and Machine Intelligence 21 495-504

The early experimental OCR systems were often rule-based by the '80s these have been completely replaced by systems which are based on statistical pattern recognition. For clearly segmented printed materials such techniques offer us

virtually error-free OCR for the most important alphabetic systems including variants of the Latin, Greek, Cyrillic, Hebrew alphabets and many more. However, when the number of symbols is large, as in the Chinese or say Korean writing systems or the symbols are just not separated from one another, as in Devanagari print, OCR systems are still far from the error rates of human readers, and the gap between the two is also very much evident when the quality of the image is compromised for example by fax transmission. Until these problems are resolved, OCR cannot play a pivotal role in the transmission of cultural heritage to the digital age that it is often assumed to have. In the recognition of handprints, algorithms with successive segmentation, classification, and identification stages are still said to be in the lead. For cursive handwriting, Hidden Markov Models that make the segmentation, classification, and identification decisions in parallel have proven superior and topmost, but performance still leaves much to be desired, both because the spatial and temporal aspects of the written signal are not necessarily in lockstep (discontinuous constituents arising, for example, at the crossing of t-s and dotting of i-s) and because the inherent variability of handwriting is far greater than that of the speech, to the extent that we often see illegible handwriting but we barely hear unintelligible speech.

3.3 Santosh Kumar Nanda, Department of Computer Science and Engineering, Eastern Academy of Science and Technology, Development of Intelligence Process Tracking System for Job Seekers

At the present time to getting a good job is a very intricate task for any job seekers. The same problem also a company can face acquiring intelligent and qualified employees. Therefore, to reduce the problem of manually searching for the right candidate, there are many management systems were applied and out of them, the computer-based management system is one of an appropriate solution for this problem. In the computer management system, software is made for job-seekers to find their suitable companies and as well as made for companies finding their suitable employees. However, the current solutions in the market are not Artificial Intelligence (AI) based, and to make privacy, security and robustness, the solution should be made with the application of an AI system. In this proposed study, an attempt has been made for finding the solution for job seekers and companies with the application of expert systems.

4. PROPOSED SYSTEM

4.1 Text Extraction

Many of the resumes that are stacked inside a company is already in a soft copy format, on which a software extraction technique can directly be applied. But to make this system more holistic, an Optical Character Recognition (also optical character reader, OCR) module is added to include the applicants who have sent a hard copy of their resume.
OCR is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image[6]. For our project, we will be OCR to convert the digital formats like PDF, JPG etc. to a text format. Fig - 1 shows the process of converting a hard copy document into a text blob[1] that can be fed into a Natural Language Processing (NLP) module.

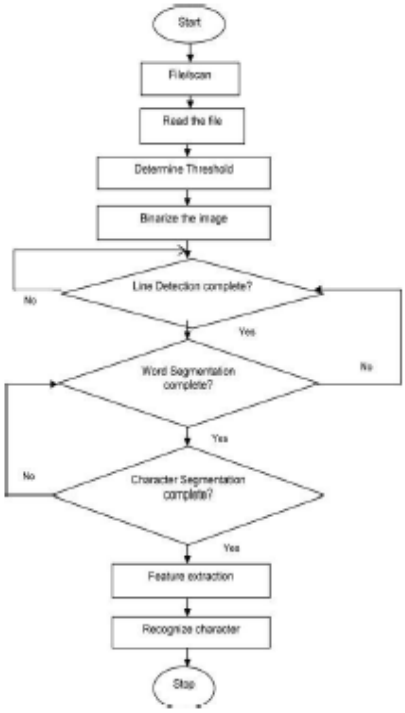


Fig -1: OCR Flow Chart

4.2 Natural Language Processing (NLP)

Natural Language Processing is an algorithm that takes the plain text as input and can convert into meaningful data. Using NLP, we are going to parse the resume, NLP requires the following for parsing:

1) Lexical Analysis
2) Syntactic Analysis
3) Semantic Analysis
4) Named Entity Recognition (NER)

Lexical Analysis:

Lexical analysis is the first phase of NLP parsing, as shown in Fig -2, the plain text input is segmented into words and paragraphs and then the tokens are created.

Syntactic Analysis:
In Syntactic analysis the analysis of the grammar and the arrangement of words in a meaningful manner is checked, sentences like "College goes to girl" is rejected.

Semantic Analysis:
Semantic analysis checks the exact meaning of the text, sentences like "Sunny is raining" will be rejected by the English semantic analysis.

Named Entity Recognition (NER):
One of the problems with using the same NLP module for all the companies is the jargons and words that mean something for that company's domain and may mean something else in general. This hindrance is overcome in our system with the help of "Named Entity Recognition" or NER. A named entity is an object that exists in the real word. With NER, we can fine tune our NLP module to understand the real word objects from a domain[2][4][5]. For example, if a company wants to hire developers, they can use our system to differentiate between people who love "Python"- the programming language and the people who love "python"- the snake, based on the context in which the word is used.

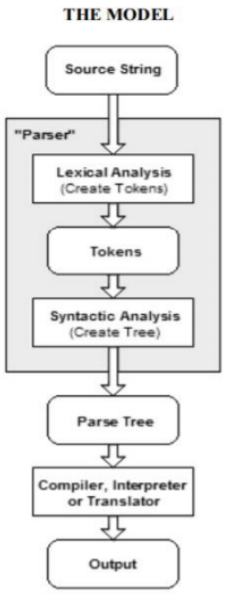


Fig -2: NLP Model

4.3 Applicant sorting and Dashboard interface

A web portal is provided to the HR, to define the constraints and the required skill sets of the company, on which the applicants are to be judged. After getting the output from the NLP, the data obtained will be used for a dashboard which will contain graphs and pie charts based on the data in the resumes. An HR can use this dashboard to prepare his/her query based on the requirement that the respective company has. In order to make this dashboard, all the data will be fed into ElasticSearch. ElasticSearch provides powerful tools to prepare such dashboards. ElasticSearch is included in order to make a more holistic system, which will be ready for basic level industry use. To make the system indifferent to many formats, the data to be fed will be fed by using LogStash.
Using queries inbuilt in ElasticSearch, the resumes will be scored and then they will be sorted according to the constraints that were provided by the HR. The individual traits of an applicant will be provided with a proportionate boost based on the priority of the trait. At the end, the cumulative score will be used to sort the applicants. A final sorted list of applicants will be displayed to the HR (Fig -3).

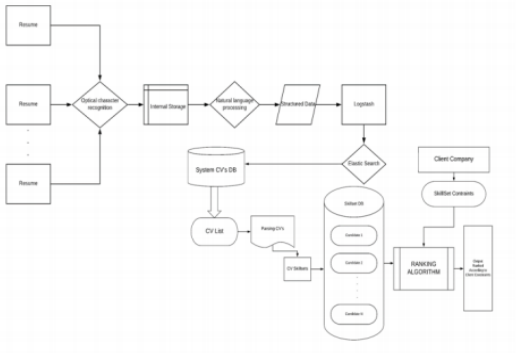


Fig -3: Proposed System

5. EXPECTED OUTCOME

Both employers and candidates will be benefited by our system, our system will parse all the applicants resume and store the summary by extracting key field in the database (Fig -4). The system will use this database to display information about the applicants using pie charts and bar graph (Fig -5) Then it will rank them (Fig -7) according to the constraints defined by the HR (Fig -6), thus reducing the unfair hiring practices and making the hiring system authentic.

Fig -4: Resume Summary

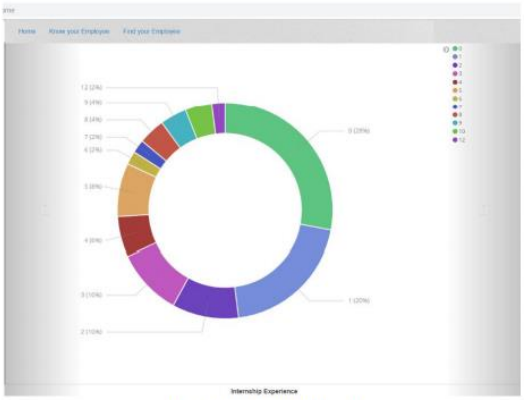


Fig -5: Pie Dashboard

Fig -6: HR constraint defining form

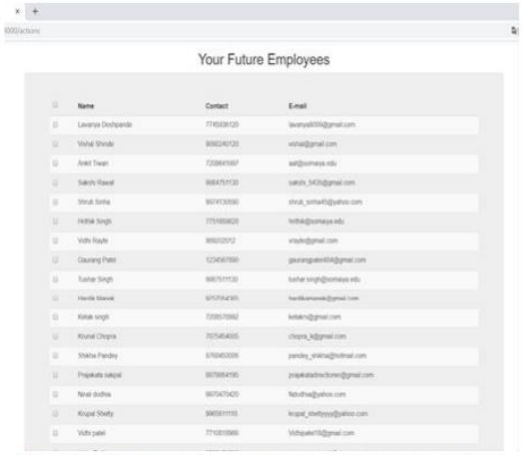


Fig -7: List of deserving applicants according to defined constraints

6. CONCLUSIONS

The process of recruitment is one of the most stressful periods for both the parties, viz., the applicants and the recruiters. Corporate companies and recruitment agencies process numerous resumes daily. This is no task for humans. An automated intelligent system will take out all the vital information from the unstructured resumes and transform all of them to a common structured format which can then be ranked for a specific job position and candidate according to need. This system aims to ease this process by making the deserving candidates, making them stand out against the crowd, which in turn, makes it easier for the recruiters. This system greatly automates the process of recruitment. The recruiters will have an idea about the quality of applicants beforehand. The applicants will be notified of the reasons because of which they were rejected, so they can improve their resume next time they submit. Furthermore, the unfair and discriminatory practices that take place during the recruitment process can be dampened to some effect.

REFERENCES

[1] F. Ciravegna, "Adaptive information extraction from the text by rule induction and generalisation," in Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI2001), 2001.

[2] A. Chandel, P. Nagesh, and S. Sarawagi, "Efficient batch top-k search for dictionary-based entity recognition," in Proceedings of the 22nd IEEE International Conference on Data

[3] M. J. Cafarella and O. Etzioni, "A search engine for natural language applications," in WWW, pp. 442–452, 2005.

[4] Swapnil Sonar, Resume Parsing with Named Entity Clustering Algorithm, IEEE Research, May 2012, http://www.slideshare.net/swapnilmsonar/resume-parsing-with-named-entity-clustering-algorithm.

[5] Sovren Resume/CV Parser, http://www.sovren.com

[6] http://vlabs.iitb.ac.in/vlabs-dev/labs_local/machine_learning/labs/exp11/theory.php

CHAPTER 9
REFERENCES

[1]. F. Ciravegna, "Adaptive information extraction from text by rule induction and generalisation," in Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI2001), 2001.

[2]. A. Chandel, P. Nagesh, and S. Sarawagi, "Efficient batch top-k search for dictionary-based entity recognition," in Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), 2006.

[3]. M. J. Cafarella, D. Downey, S. Soderland, and O. Etzioni, "KnowItNow: Fast, scalable information extraction from the web," in Conference on Human Language Technologies (HLT/EMNLP), 2005.

[4]. Swapnil Sonar, Resume Parsing with Named Entity Clustering Algorithm, IEEE Research, may 2012, http://www.slideshare.net/swapnilmsonar/resume-parsing-with-named-entity-clustering-algorithm

[5]. Sovren Resume/CV Parser, http://www.sovren.com

[6] Shakya, Sujan, "Web-based Employment Application & Processing Support System", Master of Software Engineering, May 2008, (Dr. Thomas Gendreau, Dr. Kasi Periyasamy).

[7] Bazzi, Issam, Richard Schwartz and John Makhoul (1999) An omnifont open-vocabulary OCR system for English and Arabic. Pattern Analysis and Machine Intelligence 21 495-504.

[8] Santosh Kumar Nanda, Department of Computer Science and Engineering, Eastern Academy of Science and Technology, Development of Intelligence Process Tracking System for Job Seekers.