SHRIDEVI
EDUCATION
Sri Shridevi Charitable Trust (R.)
SHRIDEVI INSTITUTE OF ENGINEERING & TECHNOLOGY
(Recognised by Govt. of Karnataka, Affiliated to VTU, Belagavi and Approved by AICTE, New Delhi)
Sira Road, Tumakuru - 572 106. Karnataka.
FS 543667
An ISO 9001:2015 Certified Institution
Phone: 0816-2212629 | Fax: 0816-2212628 | Email: info@shridevíengineering.org | Web: http://www.shrideviengineering.org

**Main Project Work**

The main objective of final year main projects in the course curriculum is to encourage students to apply their theoretical knowledge to practical use. Working on final year projects allows students to gain practical knowledge and skills to solve real-world engineering and business problems.

Students can choose their final year projects based on their interest and specialization areas to acquire complete knowledge and build skills in that domain. It is a great opportunity to build hands on practical knowledge which is vital for their career.

Shridevi Institute of Engineering & Technology is following VTU examination guidelines to assess the final year students Main projects.

PRINCIPAL
SHRIDEVI INSTITUTE OF
ENGINEERING AND TECHNOLOGY
TUMKUR - 572106.

| | |
|---|---|
| | **(h)** Seminar: Deliverable at the Institution under the supervision of a Faculty.<br>**(i)** Internship: Preferably at an industry/R and D organization/IT company/Government organization or elsewhere of significant repute for a specified period as mentioned in Scheme of Teaching and Examinations.<br>**(j)** Mandatory Courses (MC): These Courses are mandatory, without the benefit of a grade or credit, for students admitted to B.E./B.Tech. Programme. A pass in each mandatory Course is required to qualify for the award of degree. |
| **18OB3.2** | The minimum number of students registered to any Elective Course offered by the Departments shall be not less than ten.<br>However, the above condition shall not be applicable to Programmes having class strength of less than 10. In such cases, only one elective course shall be offered. |
| **18OB3.3** | A student shall exercise his option in respect of Elective Course/s and registered for the same at the beginning of the concerned semester. The student may be permitted to opt for a change of Elective Course/s within 15 days from the date of commencement of the semester as per the calendar of the University. |
| **18OB3.4** | **Course Registration:**<br>In order to maintain proper academic record of each student at the Institution, every student shall register for the Courses of a semester (Credits) under the supervision of a Faculty Advisor (also called Mentor, Counselor, etc.,) in each semester. |
| **18OB4.0** | **Internship/Professional Practice** |
| **18OB4.1** | Internship / Professional Practice<br>The Internship shall be completed during the period specified in the Scheme of Teaching and Examinations.<br>**1)** The internship shall preferably be at an industry/R and D organization/IT company/ Government organization of significant repute for a specified period as mentioned in Scheme of Teaching and Examinations.<br>**2)** The Department/college shall nominate staff member/s to facilitate, Guide and supervise students under internship.<br>**3)** The students shall report progress of the internship to the Guide in regular intervals and seek his/her advice. The Guide shall maintain the progress record of the candidates undergoing internship.<br>**4)** After the completion of Internship, students shall submit a report with completion certificate and attendance certificate to the Head of the Department with the approval of both internal and external Guides.<br>**5)** There shall be 40 marks for CIE and 60 marks for SEE. The minimum requirement of CIE marks shall be 50% of the maximum marks.<br>**6)** The internal Guide shall be the internal examiner for the SEE.<br>**7)** The external Guide for Internship shall be the external examiner for SEE. Examination for internship shall be conducted at the college and the date shall be fixed in consultation with the external Guide. The Examiners shall jointly award the SEE marks. [To be read along with 18OB8.9 (f)]<br>**8)** In case the external Guide expresses his inability to conduct the Examination, the Principal /Chief Superintendent of the Institute shall appoint a senior faculty of the Department to conduct the Examination along with the internal Guide.<br>**9)** Non-availability of Internal guide due to inevitable situations for the conduct of SEE, the Principal /Chief Superintendent of respective institute shall appoint a senior faculty of the Department to conduct the Examination.<br>**10)** The students are permitted to carry out the internship anywhere in India or abroad. The University will not provide any kind of financial assistance to any student for carrying out the Internship. |
| **18OB 5.0** | **Technical Seminar and Project** |
| **18OB 5.1** | **Technical Seminar:** Technical Seminar is one of the head of passing.<br>**(i)** Each candidate shall deliver Technical seminar as per the Scheme of Teaching and Examinations on the topic chosen from the relevant field.<br>**(ii)** The Head of the Department shall make arrangements for the conduct of seminars |

REGISTRAR
Visvesvaraya Technological University,
BELAGAVI - 590 018.

(c) For Practical/ Mini-project/Internship/Project work– Phase 2 the maximum CIE marks shall be 40. To appear for the SEE, the minimum CIE marks to be secured shall be 50 % of the maximum marks i.e., 20 marks.

(d) For all other theory Courses of the Programme, the maximum CIE marks shall be 40. To appear for the SEE, the minimum CIE marks to be secured shall be 40 % of the maximum marks i.e., 16 marks.

(f) For Additional Mathematics I and II (to be completed by diploma lateral entry students) the maximum CIE marks shall be 40. To appear for the SEE, the minimum CIE marks to be secured shall be 40 % of the maximum marks i.e., 16 marks.

(g) For Engineering Graphics and Elements of Civil Engineering and Mechanics (of First Year Engineering and to be completed by B.Sc graduates under lateral entry) the maximum CIE marks shall be 40. To appear for the SEE, the minimum CIE marks to be secured shall be respectively 50 % and 40 % of the maximum marks i.e., 20 and 16 marks.

**18OB8.2** **Continuous Internal Evaluation Procedure:** [To be read along with 18 OB 8.1and 8.3]

**(a) Theory Courses:**

**(i)** CIE Marks in each theory Course [including 'Technical English I and II', 'Constitution of India, Professional Ethics and, Cyber Law', 'Environmental Studies', 'Additional Mathematics I and II'], shall be the sum of marks prescribed for tests and assignments. Marks prescribed for tests shall be 30 and that for assignments 10.

**(ii)** The CIE marks awarded for tests in the theory Courses shall be based on three tests generally conducted at the end of fifth, tenth and fifteenth week of each semester. Each test shall be conducted for a maximum of 50 marks and the final test marks shall be the average of three tests, proportionately reduced to a maximum of 30 marks.

**(iii)** The remaining 10 marks shall be awarded based on the evaluation of assignments/unit tests/written quizzes that support to cover both lower and higher order thinking skills as per Revised Bloom's Taxonomy.

**(iv)** Final CIE marks awarded shall be the sum of 18OB8.2 (a) (ii) and (iii) for a maximum of 40 marks.

(v) The candidates shall write the tests, assignments/unit-tests /written quizzes in Blue Books which shall be preserved by the Principal/ Head of the Department for at least six months after the announcement of University results and shall be made available for verification at the direction of the Registrar (Evaluation).

REGISTRAR
Visvesvaraya Technological University
BELAGAVI - 590 018.

**(b) Engineering Graphics/ Drawing/Field work Courses:**

The CIE marks awarded for I year Engineering Graphics Course shall be based on

(i) Classwork for 24 marks (sketching and Computer Aided Engineering Drawing).

(ii) Two Tests conducted in the same pattern as that of SEE for 16 marks (The marks secured can be taken as best of the two tests).

(iii) Final CIE marks awarded for Engineering Graphics shall be the sum of 18OB8.2 (b) (i) and (ii) for a maximum of 40 marks.

(iv) The CIE marks awarded for higher semester Drawings/ Design Drawings offered by various branches shall be based on the evaluation of the sheets and one test in the ratio 60:40.

(v) The CIE marks awarded for field work (like Surveying Practice) shall be based on the evaluation of the associated field work and one test in the ratio 60:40.

**(c) Practical Courses:**

The CIE marks awarded in case of Practical shall be based on the weekly evaluation of laboratory journals/ reports after the conduction of every experiment and one practical test in the ratio 60:40.

**(d) Internship:**

The CIE marks awarded for internship shall be based on the evaluation of Internship Report, Presentation skill and Question and Answer session in the ratio 50:25:25.

**(e) Technical Seminar:**

The CIE marks awarded for Technical Seminar shall be based on the evaluation of Seminar Report, Presentation skill and Question and Answer session in the ratio 50:25:25.

**(f) Mini – Project:**

| | |
|---|---|
| | The CIE marks awarded for Mini - Project, shall be based on the evaluation of Mini - Project Report, Project Presentation skill and Question and Answer session in the ratio 50:25:25.The marks awarded for Mini - Project report shall be the same for all the batch mates.<br>**(g) Main Project Work:**<br>**(i)** Project Work Phase – 1<br>The CIE marks awarded for project work phase -1 shall be based on the evaluation of project work phase -1 Report, Project Presentation skill and Question and Answer session in the ratio 50:25:25.The marks awarded for the Project report shall be the same for all the batch mates.<br>**(ii)** Project Work Phase - 2<br>The CIE marks awarded for project work phase -2 shall be based on the evaluation of project work phase -2 Report, Project Presentation skill and Question and Answer session in the ratio 50:25:25.The marks awarded for Project report shall be the same for all the batch mates. |
| | **(h) Vyavaharika Kannada (Balake Kannada)/Aadalitha Kannada** (Samskruthika **Kannada)**<br>**(i)** CIE Marks in Vyavaharika Kannada (Balake Kannada)/Aadalitha Kannada (Samskruthika Kannada) shall be the sum of marks prescribed for tests and assignments. Marks prescribed for tests shall be 75 and that for the assignments shall be 25.<br>**(ii)**The CIE marks awarded for the tests shall be based on three tests generally conducted at the end of fifth, tenth and fifteenth week of each semester. Each test shall be conducted for a maximum of 25 marks and the final CIE marks shall be the sum of the marks of all the three tests.<br>**(iii)** The remaining 25 marks shall be awarded based on the evaluation of assignments/oral discussions/ quizzes that supports communication skills.<br>**(iv)** Final marks awarded shall be the sum of 18OB8.2 (h) (ii) and (iii) for a maximum of 100 marks.<br>**(v)** Students shall write the tests in Blue Books ,and complete the exercises/activates/ questions given in the University Kannada textbook. These shall be preserved by the Principal/ Head of the Department for at least six months after the announcement of University results and shall be made available for verification at the direction of the Registrar (Evaluation). |
| **18OB8.3** | **(a)** The CIE marks in the case of Internship/Technical Seminar/Mini-Project and Project Work Phase 1 and 2 shall be awarded by a committee consisting of the Head of the concerned Department and two senior faculty members of the Department, one of whom shall be the Guide.<br>**(b)** A committee constituted by the Head of the Department of Humanities and Social Science shall inspect and authenticate the award the CIE marks for the Course Vyavaharika Kannada (Balake Kannada)/Aadalitha Kannada (Samskruthika Kannada). The committee shall consist of two senior faculty members of the Department and the senior most acting as the Chairperson. |
| **18OB8.4** | **(i)** Students satisfying the attendance requirement but failing to secure the minimum percentage of CIE marks, in any Course/s, shall not be eligible for the SEE conducted by the University and they shall be considered as fail in that Course /those Courses. However, they can appear for University examinations conducted in other Courses of the same semester and backlog Course/s if any.<br>**(ii)** Students who have satisfied the attendance requirement but not the CIE requirements shall be permitted to register afresh and appear for SEE after satisfying the CIE requirements in the same Course/s (with or without satisfying the attendance requirement) when offered during subsequent semester/s.<br>**(iii)** Each appearance to SEE to complete a course shall be treated as an attempt. |
| **18OB8.5** | CIE marks of those students, who come under 18OB8.4, shall also be sent to the Registrar (Evaluation) along with other course CIE Marks. |

REGISTRAR
Visvesvaraya Technological University,
BELAGAVI - 590 018.

SHRIDEVI
EDUCATION

# SHRIDEVI INSTITUTE OF ENGINEERING & TECHNOLOGY

## Department of Electronics & Communication & Engineering

ACADEMIC YEAR 2021-22 (Even Semester)

FS 543667
An ISO 9001:2015 Certified Institution

## SEM: VIII SEM

## PROJECT WORK GUIDE ALLOTMENT LIST

| Sl No | Groups | USN | Name of the Student | Project Title | Project Guide & Signature |
|---|---|---|---|---|---|
| 1 | Group 1 | 1SV18EC002 | ANUSHA G S | Data analysis of students job entry | Dr.Umesha G B |
| | | 1SV18EC005 | BARATHI M | | |
| | | 1SV18EC006 | BRUNDA K | | |
| | | 1SV18EC009 | GURANNA GOUDA | | |
| 2 | Group 2 | 1SV18EC016 | PRASHANTH M | Automatic sql query generation using natural language processing unit | Prof.Pradeepkumar S S |
| | | 1SV18EC011 | LATHASHREE K R | | |
| | | 1SV18EC007 | CHANDANA D | | |
| | | 1SV18EC018 | RACHANA S R | | |
| 3 | Group 3 | 1SV18EC013 | MOUNIKA Y | Data analysis of job portal | Prof.Raghavendra D |
| | | 1SV18EC019 | SADAF NAZ | | |
| | | 1SV19EC401 | JYOTHI R | | |
| | | 1SV18EC012 | MOUNESH GOUDA | | |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 4 | Group 4 | 1SV18EC008 | DEVIKA L | Autonomus vechile | Dr.Pradeep KGM |
|  |  | 1SV18EC021 | SHIRISHA R T |  |  |
|  |  | 1SV18EC024 | YASHASWINI K Y |  |  |
|  |  | 1SV19EC403 | NAVYASHREE S M |  |  |
| 5 | Group 5 | 1SV18EC004 | BASAVARAJ | Smartstick for visually impaired people | Prof.Raghavendra D |
|  |  | 1SV18EC022 | SIDRAM |  |  |
|  |  | 1SV18EC020 | SAMEER BICHAGATTI |  |  |
|  |  | 1SV19EC400 | ARUNA R N |  |  |
| 6 | Group 6 | 1SV19EC405 | SWAMY M | Speech to text converter using embedded C | Prof.Pradeepkumar S S |
|  |  | 1SV19EC404 | PRAVEEN G D |  |  |
|  |  | 1SV18EC014 | NAGESH D R |  |  |
| 7 | Group 7 | 1SV18EC010 | KETANRAJ S | Coin based vechile battery charging | Dr.Pradeep KGM |
|  |  | 1SV18EC023 | SRINIVAS C |  |  |
|  |  | 1SV19EC402 | MAHADEVAIAH M B |  |  |
|  |  | 1SV17EC012 | RAVISH KUMAR |  |  |
| 8 | Group 8 | 1SV17EC011 | RAKESH L | Solar powered cooling system for car parked in sun light | Prof.Aijaz ahamed sharief |
|  |  | 1SV17EC016 | TEJASWINI D |  |  |

**Dr.Pradeep K G M**
**(Project Coordinator)**

**Prof. Aijaz Ahamed Sharief**
**( HOD )**

HOD
Dept of E&C
SIET, Tumkur-6

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
"JNANA SANGAMA", BELGAVI-590018 KARNATAKA

Project Report (18ECP83)
ON

# *"Data Analysis of Students Job Entry"*

**Submitted in partial fulfillment of the requirement for the award of degree**

**BACHELOR OF ENGINEERING**

**IN**

**ELECTRONICS & COMMUNICATION ENGINEERING**

**Submitted by:**

**ANUSHA G S (1SV18EC002)**
**BHARATHI M (1SV18EC005)**
**BRUNDA K (1SV18EC006)**
**GURANNAGOUDA (1SV18EC009)**

**Under the Guidance of:**
**Dr Umesha G B** BE, M. Tech, Ph.D.

Assoc. Prof., Dept. of ECE, SIET, Tumakuru-06

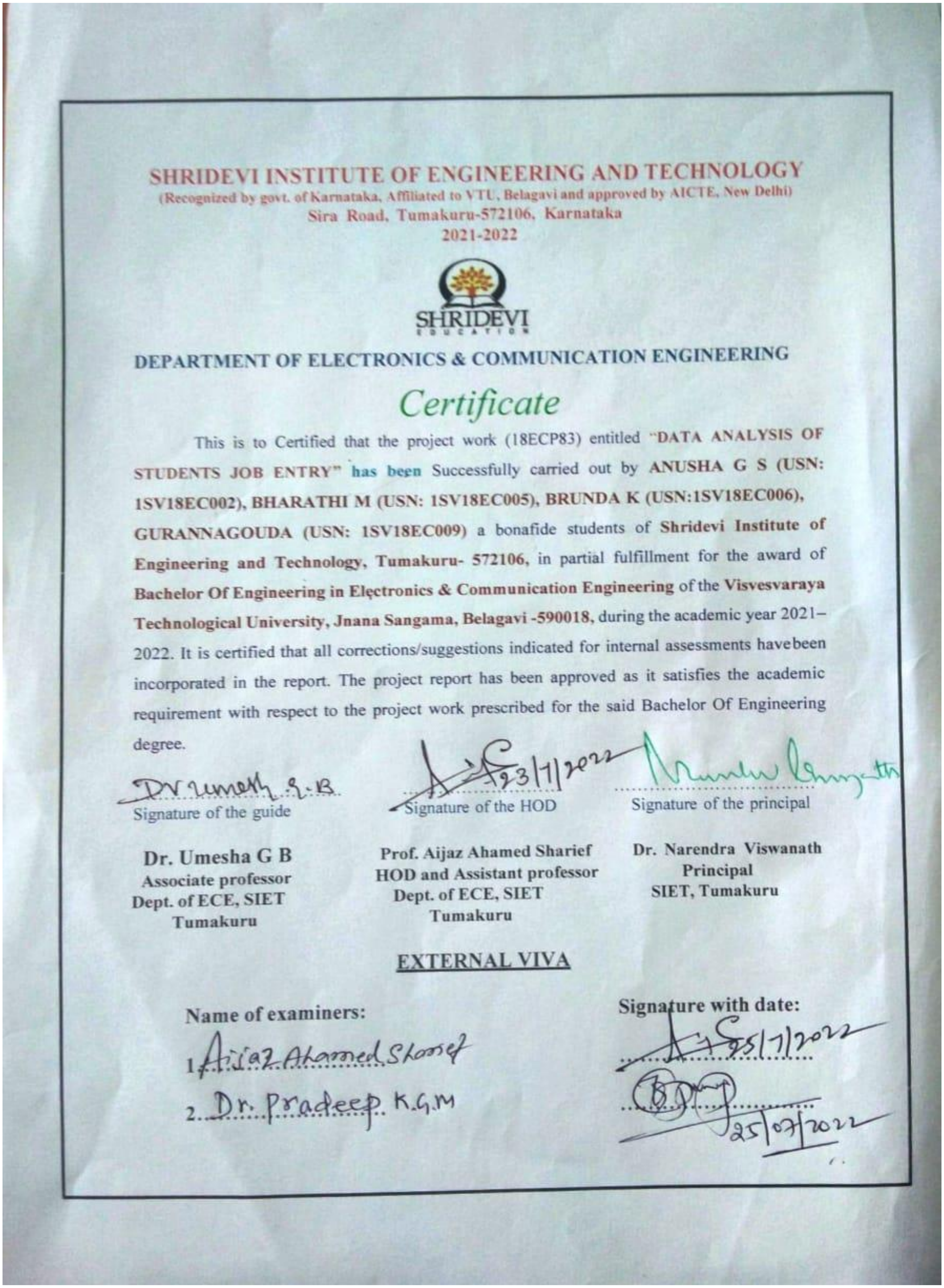
DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

**SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY**

(Recognized by govt. of Karnataka, Affiliated to VTU, Belagavi and approved by AICTE, New Delhi)

Sira Road, Tumakuru-572106

**2021– 2022**



SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Recognized by govt. of Karnataka, Affiliated to VTU, Belagavi and approved by AICTE, New Delhi)

Sira Road, Tumakuru-572106, Karnataka

2021-2022

SHRIDEVI
EDUCATION

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

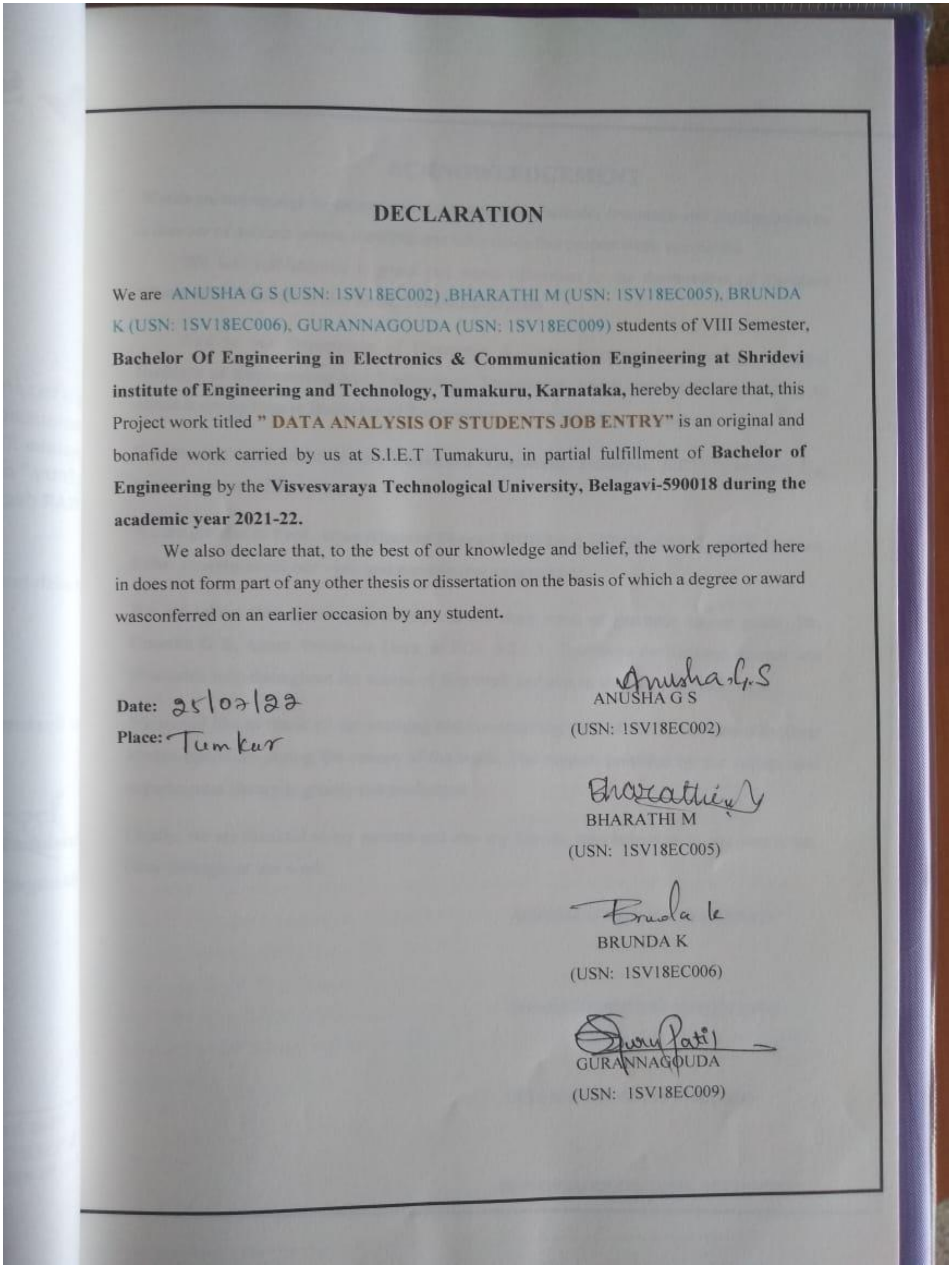
## *Certificate*

This is to Certified that the project work (18ECP83) entitled "**DATA ANALYSIS OF STUDENTS JOB ENTRY**" **has been** Successfully carried out by **ANUSHA G S (USN: 1SV18EC002), BHARATHI M (USN: 1SV18EC005), BRUNDA K (USN:1SV18EC006), GURANNAGOUDA (USN: 1SV18EC009)** a bonafide students of **Shridevi Institute of Engineering and Technology, Tumakuru- 572106,** in partial fulfillment for the award of **Bachelor Of Engineering in Electronics & Communication Engineering** of the **Visvesvaraya Technological University, Jnana Sangama, Belagavi -590018,** during the academic year 2021–2022. It is certified that all corrections/suggestions indicated for internal assessments havebeen incorporated in the report. The project report has been approved as it satisfies the academic requirement with respect to the project work prescribed for the said Bachelor Of Engineering degree.

| | | |
|---|---|---|
| Dr. umesh G.B | 23/7/2022 | |
| Signature of the guide | Signature of the HOD | Signature of the principal |
| **Dr. Umesha G B**<br>**Associate professor**<br>**Dept. of ECE, SIET**<br>**Tumakuru** | **Prof. Aijaz Ahamed Sharief**<br>**HOD and Assistant professor**<br>**Dept. of ECE, SIET**<br>**Tumakuru** | **Dr. Narendra Viswanath**<br>**Principal**<br>**SIET, Tumakuru** |

**EXTERNAL VIVA**

| | |
|---|---|
| **Name of examiners:** | **Signature with date:** |
| 1. Aijaz Ahamed Sharief | 25/7/2022 |
| 2. Dr. Pradeep K.G.M | 25/07/2022 |



## DECLARATION

We are ANUSHA G S (USN: 1SV18EC002) ,BHARATHI M (USN: 1SV18EC005), BRUNDA K (USN: 1SV18EC006), GURANNAGOUDA (USN: 1SV18EC009) students of VIII Semester, **Bachelor Of Engineering in Electronics & Communication Engineering at Shridevi institute of Engineering and Technology, Tumakuru, Karnataka,** hereby declare that, this Project work titled **" DATA ANALYSIS OF STUDENTS JOB ENTRY"** is an original and bonafide work carried by us at S.I.E.T Tumakuru, in partial fulfillment of **Bachelor of Engineering** by the **Visvesvaraya Technological University, Belagavi-590018 during the academic year 2021-22.**

We also declare that, to the best of our knowledge and belief, the work reported here in does not form part of any other thesis or dissertation on the basis of which a degree or award wasconferred on an earlier occasion by any student.

Date: 25/07/22

Place: Tumkur

Anusha.G.S

ANUSHA G S

(USN: 1SV18EC002)

Bharathi

BHARATHI M

(USN: 1SV18EC005)

Brunda k

BRUNDA K

(USN: 1SV18EC006)

GuruPatil

GURANNAGOUDA

(USN: 1SV18EC009)

# ACKNOWLEDGEMENT

Words are not enough to express our deep sense of gratitude, reverence and indebtedness to a number of persons whose contributions have made this project work successful.

We feel self-honored to place our warm salutation to the **Authorities of Shridevi Charitable Trust (R.),** Sira Road, Tumakuru-572106.

And to our Department of Electronic & Communication Engineering, **Shridevi Institute of Engineering and Technology, Tumakuru,** which give us the opportunity to obtain a strong base in **Bachelor of Engineering** and profound knowledge**.**

We are highly grateful to **Dr. Narendra Viswanath,** Principal, SIET, Tumakuru for encouragement of this work.

We are grateful to **Prof. Aijaz Ahamed Sharief**, HOD Dept. of ECE, S.I.E.T, and Tumakuru forhis concern about our work and constructive suggestions.

We are taking this opportunity to express our deep sense of gratitude to our guide, **Dr. Umesha G B,** Assoc. Professor, Dept. of ECE, S.I.E.T, Tumakuru for his keen interest and invaluable help throughout the course of this work and also in shaping our profession.

We would like to thank all the teaching and non-teaching staff of SIET, Tumakuru for their kindco-operation during the course of the work. The support provided by the college and departmental library is greatly acknowledged.

Finally, we are thankful to my parents and also my friends, who helped us in one way or the other throughout our work.

ANUSHA G S (USN: 1SV18EC002)

BHARATHI M (USN: 1SV18EC005)

BRUNDA K (USN: 1SV18EC006)

GURANNAGOUDA (USN: 1SV18EC009)

# CONTENTS

# ABSTRACT

Our Job Portal consists of 3 modules. Admin, Recruiter and Jobseeker. The admin has authority over the complete portal. He can see the recruiter requirements & search the relevant candidates for that profile. Recruiter has to buy some packages after registration and can post jobs, view jobseeker profile, download their resumes as per the restrictions of the package for the given validity period.

Jobseeker can register for free in our portal and can search and apply for jobs matching their profile. This research aims to develop a job web portal for the students in the Faculty of Computer Science and Information Technology The main aims of this portal are to connect to the industries and acts as an online recruitment to support the students to find the right IT job after graduation.

Furthermore, this system enhances the understanding concept and importance of the job portal for students in the universities. A survey was conducted to identify the students' problems with the existing portal of the faculty and to gather their requirements which can be incorporated in to the portal to be developed. Job seeking is a very difficult task in India and determining a career path is a much more difficult challenge. There is a lack of planning in the life of job aspirants. Job aspirants don't get enough updated information regarding the job opportunities in different sectors. There is a need of a proper medium through which job aspirants could get information on the number of job vacancies, number of intakes, salary etc. in each job industry for the future years. Using such predictions, job aspirants can plan on a specific career path and the specific qualification required for the job can be attained in the following years. Nowadays, websites act as a powerful medium to reach to the job aspirants.

The major contribution of this dissertation will be a web application using angular platform, through which job aspirants get updated knowledge regarding future job opportunities in the form of statistics, charts and graphs. Therefore, this dissertation endeavors to minimize the challenges faced by job aspirants through machine learning techniques. The study was conducted by adopting a mixed method approach: both statistical and predictive in nature. KNN and neural network regression are some of the machine learning algorithms considered in this study.

SCII

Syscon India

**System Consultant Information India (P) Ltd.**

*Solutions Crafted with Intelligence and Innovation*

# CERTIFICATE

This is to certify that **Ms. ANUSHA G S** with USN **1SV18EC002**, student of BE - Electronics and Communication Engineering, Shridevi Institute of Engineering & Technology Tumakuru, Karnataka 572106 has successfully completed her project titled " Data Analysis of Students Job Entry " on our organization for the partial fulfillment of her **Bachelor of Engineering (ECE) IV YEAR** degree.

We wish her all the very best in her future endeavours.

**For System Consultant Information India (P) Ltd.**

SYSTEM CONSULTANT INFORMATION INDIA PVT. LTD. ★ TUMKUR

**Biswanath Misra**

**(Manager)**

Syscon India

**SCII System Consultant Information India (P) Ltd.**

*Solutions Crafted with Intelligence and Innovation*

# CERTIFICATE

This is to certify that **Ms. BHARATHI M** with USN **1SV18EC005**, student of BE - Electronics and Communication Engineering, Shridevi Institute of Engineering & Technology Tumakuru, Karnataka 572106 has successfully completed her project titled " Data Analysis of Students Job Entry " on our organization for the partial fulfillment of her **Bachelor of Engineering (ECE)** IV YEAR degree.

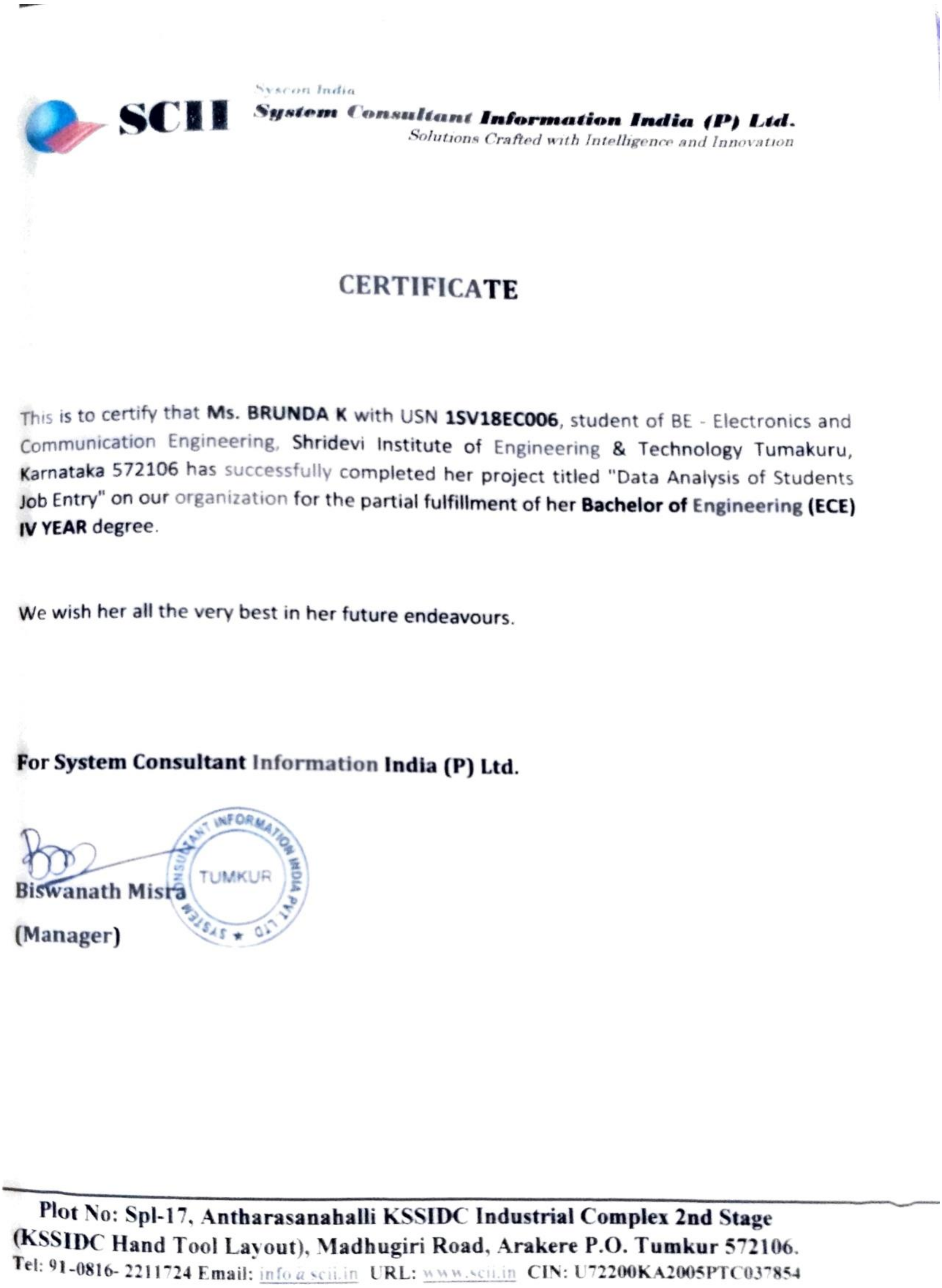
We wish her all the very best in her future endeavours.

**For System Consultant Information India (P) Ltd.**

SYSTEM CONSULTANT INFORMATION INDIA PVT LTD ★ TUMKUR

**Biswanath Misra**

**(Manager)**



Syscon India

SCII System Consultant Information India (P) Ltd.

Solutions Crafted with Intelligence and Innovation

# CERTIFICATE

This is to certify that **Ms. BRUNDA K** with USN **1SV18EC006**, student of BE - Electronics and Communication Engineering, Shridevi Institute of Engineering & Technology Tumakuru, Karnataka 572106 has successfully completed her project titled "Data Analysis of Students Job Entry" on our organization for the partial fulfillment of her **Bachelor of Engineering (ECE) IV YEAR** degree.

We wish her all the very best in her future endeavours.

**For System Consultant Information India (P) Ltd.**

SYSTEM CONSULTANT INFORMATION INDIA PVT LTD ★ TUMKUR

**Biswanath Misra**

**(Manager)**

Plot No: Spl-17, Antharasanahalli KSSIDC Industrial Complex 2nd Stage
(KSSIDC Hand Tool Layout), Madhugiri Road, Arakere P.O. Tumkur 572106.
Tel: 91-0816- 2211724 Email: info@scii.in URL: www.scii.in CIN: U72200KA2005PTC037854

Syscon India

SCII **System Consultant Information India (P) Ltd.**

*Solutions Crafted with Intelligence and Innovation*

# CERTIFICATE

This is to certify that **Mr. GURANNAGOUDA** with USN **1SV18EC009**, student of BE - Electronics and Communication Engineering, Shridevi Institute of Engineering & Technology Tumakuru, Karnataka 572106 has successfully completed his project titled " Data Analysis of Students Job Entry” on our organization for the partial fulfillment of his/her **Bachelor of Engineering (ECE) IV YEAR** degree.

We wish him all the very best in his future endeavours.

**For System Consultant Information India (P) Ltd.**

**Biswanath Misra**
**(Manager)**

SYSTEM CONSULTANT INFORMATION INDIA PVT. LTD. ★ TUMKUR.

# CHAPTER 1

# INTRODUCTION

Unemployment is one of the serious social issues faced by both developing and developed and countries. For example, in Europe the rate of unemployment has been increasing rapidly since the 1970’s.Dorn and Naz Mentioned that one of the reasons for this problem is the unfair distribution or lack of information on job opportunities so people are unableto now the new job vacancies. It means that there are some jobs available, but jobseekers do not have access to that information.

An efficient search of the internet might help to jobseekers in their job hunt. There are some web portals that provide an efficient way to search the web for online informationon job vacancies for jobseekers. Today, the internet has changed many aspects of our life ,such as the way we look for jobs If one person wants to find a new job, he/she can submit a resume using word processing software like Microsoft Office Word, open a web browser to send the resume and receive an e-mail. Online recruitment has become the standard method for employers and jobseekers to meet their respective objectives.

The Student Job Entry is a platform between job seeker. The Student or Job seeker can easily find and apply for job by login into system. Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

A recent study suggests enormous amount of data is being created daily. This datacontains a lot of useful information that needs to be filtered out for further useful purposes. This process of filtering data is called data analysis.

The basic idea behind this HR analytics tool is to provide ease to people for finding the best jobs based on their skills that the person mentioned in their resume in the first versionor the skills inputted manually in the second version. It’s very difficult for everyone tofind out jobs, internship etc. on the basics of skills they have. Everyone to spend a lot of time to search jobs on various job posting sites like glass door, indeed, intern Shala, etc. checking each and every posting in the list with having applying some filters to it but it consume very much time because when we have to check each and every time that the job posting we are going to apply,

Is we are worthy to apply to this job? Sometimes, we are not having the skills that are mentioned in the job description. So, as I mentioned the basic idea. We are suggesting the job on the basics of the skills of the person. It has many benefits that we have not to check each and every posting of jobs so it consumes less time. Just we have to provide the input, i.e., the skills and output, i.e., the suggested jobs will be shown on the UI or on the local host.

# CHAPTER 2

## 2.1 LITERATURE SURVEY

### A. Job Procurement: Old and New Ways:

Job seeking usually involves different ways to look for jobs such as through personal contacts, direct telephone calls to employers, job agency office, scanning online job listings, etc. Before the Internet, became widely uses as a method of seeking jobs, jobseekers spent a lot of time using various methods to look for job openings. Today, jobseekers use online methods which are very convenient and save a lot of time.

Lists the following methods to be the traditional (old) ways for recruitment:

- Employment recruitment agencies
- Job fairs
- Advertising in the mass media such as newspapers
- Advertisement in television and radio
- Management Consultants
- Existing employee contacts
- School's colleges or universities student's services department
- Workers or professional referrals

These old job seeking methods are too slow, stressful, challenging and also lack quality . In addition, the applicants have to consider the cost and the amount of time to get the information they need, and other preparations they have to make.

Finding all available job vacancies is a main step at in the job-seeking process. The Internet is now a powerful tool that jobseekers can use. Today, there are many sites that advertise job positions to be filled by people with certain skills in various fields. The Internet plays an important role in the area of human resource planning and development. Most planning and development organizations are now using computer technology and theInternet for staff recruitment. It should be noted that although the Internet has facilitated the process of job-seeking, it has not replaced the traditional methods, completely.

### B. Importance of Job Portals:

In the age of technology, the Internet has become the main source of information for jobseekers. Large corporations, Institutions, and universities include information on career Prospects on their websites. According to a survey, 70% of the workforce uses websites or portals on the Internet to Search for jobs in France. These websites or portals provide search engine to access information on job opportunities.

## 2.2 EXISTING SYSTEM AND PROPOSED SYSTEM

### 2.2.1 EXISTING SYSTEM

The present job portal System works on the manual basis, so the information is stored on the sheet of paper. Storing the information in the hard copies increases the workload for maintaining the data while it also increase the time consumption required. The organization faces difficulties in finding the candidate with the required skill for the job while a jobseeker has to perform lots effort to find job. while in the organization, it also a cumbersome process to check the details of each candidate by the HR.

To try the resume software, just one have to upload the resume and copy-and- paste a job description job seeker interested in applying for. Job scan will then analyze your resume for formatting errors, key qualifications, hard skills, best practices, word count, tone, and more. Job scan also features a potent lineup of other job search tools, such as the free resume builder, Power Edit real-time resume editor, and LinkedIn Optimization.

### 2.2.2 PROPOSED SYSTEM

The proposed job portal System is internet based so it can be a useful tool to maintain the data related to any person. This Job Portal System Project will keep the data of a job seeker into the System, and when an organization needs the candidate with required skill, it will show the candidate profile to the HR of that organization, Once the candidate will be shortlisted by the organization it will notify about it to the organization and the job seeker to get the right job with correct skill.

Data Analysis of Students Job Entry, as the name suggests is the tool to suggest the top 10 jobs from the dataset based on the skills of the person.

- There are two versions of this tool:
- In the First version, I have used pyre parser for extracting the skills from the resume and the second version is manually providing the skills as the input. Both have there UI based on their input type.
- The dataset is created via scrapping glass door website using selenium and beautiful

soup, around 1923 jobs are extracted.

And for matching the job description and the skills, I have used the NLP algorithm which is based on n-grams, tf-idf, and KNN is used for finding the best jobs for the person. At last deployed both the version using Flask and rendering is done using simple html templates.

- Version 1: (Input is the resume)
- Version 2: (Skills is entered manually)

# CHAPTER 6

# SOFTWARE AND HARDWARE REQUIREMENTS

## 3.1 SOFTWARE REQUIREMENTS

- ❖ Web Browser(Google chrome, Mozilla Firefox)
- ❖ Anaconda Software
- ❖ Technical Tool kits Used:
  - Python
  - Jupyter notebook
  - PyCharm
- ❖ Libraries Used:
  - Nearest Neighbors
  - TFIDF
  - S K Learn
  - Flask
  - Resume Parser
- ❖ Packages Used:
  - NumPy
  - Pandas
  - Regular Expression(re)

## 3.2 HARDWARE REQUIREMENTS

- ❖ Operating System: Windows 10/11
- ❖ RAM: 8GB of RAM
- ❖ Hard Disk: 1Tera Byte
- ❖ Memory: 500GB Hard drive

# CHAPTER 4

# SOFTWARE SPECIFICATIONS AND TECHNICAL TOOLKITS

## 4.1 SOFTWARE SPECIFICATIONS

### 4.1.1 Web Browser:

Web Browsers are software installed on your PC. To access the Web you need a web browsers, such as Netscape Navigator, Microsoft Internet Explorer or Mozilla Firefox. Currently you must be using any sort of Web browser while you are navigating through my site tutorialspoint.com. On the Web, when you navigate through pages of information this is commonly known as browsing or surfing.

### 4.1.2 Google Chrome:

This web browser is developed by Google and its beta version was first released on September 2, 2008 for Microsoft Windows. Today, chrome is known to be one of the most popular web browsers with its global share of more than 50%.

Google Chrome is a cross-platform web browser developed by google. It was first released in 2008 for Microsoft Windows, built with free software components from Apple web kit and Mozilla Firefox. It was later ported to Linux, macOS, iOS, and Android, where it is the default browser. The browser is also the main component of Chrome OS, where it serves as the platform for web Applications.

Most of Chrome's source code comes from Google's free – open source code project chromium, but Chrome is licensed as proprietary freeware. Web kit was the original rendering engine but Google eventually forked it to create the Blink engine; all Chrome variants except iOS now use Blink.

Chromium projects and browser

The Chromium projects are open-source, community-driven projects to develop technologies for Chrome and Chrome OS. The Chromium browser is similar to Chrome, but is developed exclusively with Chrome's open-source components.

Ungoggled Chromium project and browser

Ungoggled Chromium is a development fork of the Chromium browser which strips out selected browser components.

The project's stated goals are to:

- Disable or remove offending services and features that communicate with Google or weaken privacy.
- Strip binaries from the source tree, and use those provided by the system or build them from source.
- Add, modify, or disable features that inhibit control and transparency.

The Ungoggled Chromium browser source code can be downloaded from its repository on GitHub

### 4.1.3 Mozilla Firefox:

Mozilla Firefox is nothing but a Web Bowser, with which one can access the internet. The web browser lets one access information in form of text, audio, images, and videos from all around the world. Mozilla Firefox was developed by Mozilla Foundation in 2002 under the Phoenix community. Nowadays, it is called Firefox only as it is derived from Mozilla Web Browser it is also known as Mozilla Firefox.

Mozilla Firefox can also be used to browse via android and iOS. Firefox officially was released on Nov 2004 and gave tough competition to Microsoft’s Internet Explorer with its add-ons, security, and speed. Firefox got it’s the highest popularity at the end of 2009 when it had reached of total market usage. But after the introduction of Google Chrome, the popularity of Firefox declined. As of now, it has around 5%, market users.

Mozilla Firefox is known for its speed. Though the Firefox browser needs a lot of memory for operating efficiently. It may limit the multiple tasking of computers. However, It provides better network security. It has advanced security options that protect your system from spyware and malwares. It has strong popup broker and authentication protocols which makes it safe from potential attackers using any unauthorized codes. Further to enhance security users can use enhanced security options like No Script and Flash block. It enables user to execute advanced code so that certain new features which can make the browser more Intuitive.

Firefox has an interface which is very user friendly and the user can use a number of add-ons on top of that user can customize the browsing also. It has more than 6000 extensions, user can customize the browser with more than 500 themes. Mozilla offers Tabbed Browsing which can let the user open unlimited tabs in a single window. It also has got embedded memory which makes it capable of remembering pages, in case if your systems is turned off by mistake, you can recall all the open pages.

Firefox is a safe browser when it comes to protecting personal data. It provides freedom of browsing and protection as no other browser does. Firefox is a non-profit organization, which means it does not intend to get profit from collecting personal browsing information of users. Firefox has an Open-Source-Project which makes it for anyone to see the code and have a look at it how it works. And it does not share any information about the user with third-party partners.

Firefox is a web browser to get information using the internet from different servers available all around the world. Firefox was introduced around 2002 but was fully released around 2004 and since then it has got a lot of popularity.

Firefox consumes more memory but highly efficient and safer over the other browsers. Firefox does not use personal settings and personal information of users to gain profits of any sort. Moreover, with the introduction of Firefox Quantum, now it consumes less memory and is quite faster also.

### 4.1.4 Anaconda Software:

Anaconda is an amazing collection of scientific Python packages, tools, resources, and IDEs. This package includes many important tools that a Data Scientist can use to harness the incredible force of Python. Anaconda individual edition is free and open source. This makes working with Anaconda accessible and easy. Just go to the website and download the distribution.

With over 20 million users, covering 235 regions, and with over 2.4 billion package downloads; Anaconda has grown an exceptionally large community. Anaconda makes it easy to connect to several different scientific, Machine Learning, and Data Science packages.

**The key features:**

- Neural Networks
- Machine Learning
- Predictive Analytics
- Data Visualization
- Bias Mitigation

If you are interested in Data Science, then you should know about this Python Distribution. Anaconda is great for deep models and neural networks. You can build models, deploy them, and integrate with leading technologies in the subject. Anaconda is optimized to run efficiently for machine learning tasks and will save you time when developing great algorithms. Over 250 packages are included in the distribution. You can install other third-party packages through the Anaconda terminal with conda install. With over 7500 data science and machine learning packages available in their cloud-based repository, almost any package you need will be easily accessible. Anaconda offers individual, team, and enterprise editions. Included also is support for the R programming language.

The Anaconda distribution comes with packages that can be used on Windows, Linux, and MacOS. The individual edition includes popular package nameslike numpy, pandas, scipy, sklearn, tensorflow, pytorch, matplotlib, and more. TheAnaconda Prompt and PowerShell makes working within the filesystem easy and manageable. Also, the GUI interface on Anaconda Navigator makes working with the everything exceptionally smooth. Anaconda is an excellent choice if you are looking for a thriving community of Data Scientists and ever-growing support in the industry. Conducting Data Science projects is an increasingly simpler task with the help of great tools like this.

Open-Source software that allows Data Scientist to conduct workflows and effectively realize scientific and computational solutions. With an emphasis on presentation and readability, Jupyter Notebooks are a smart choice for collaborative projects as well and insightful publications. Jupyter Notebooks are open source and developed on GitHub publicly by the Jupyter community.

### First of All, what is Anaconda & Why Should I bother about it?

You probably already have Python installed and will be wondering why you need this at all. Firstly, since Anaconda comes with a bunch of data science packages, you'll be all set to start working with data. Secondly, using conda to manage your packages and environments will reduce future issues dealing with the various libraries you'll be using. In most of the real-world Data Science projects, conda based package and environments are widely used and I personally preferred conda based package installation and maintenance of project then installing and maintaining directly PIP based packages.

### So, Why Anaconda?

Anaconda is a distribution of packages built for data science. It comes with conda, a package, and environment manager. We usually used conda to create environments for isolating our projects that use different versions of Python and/or different version of packages. We also use it to install, uninstall, and update packages in our project environments. When you download Anaconda first time it comes with conda, Python, and over 150 scientific packages and their dependencies. Anaconda is a fairly large download (~500 MB) because it comes with the most common data science packages in Python, for people who are conservative about disk space, there is also Manikonda, a smaller distribution that includes only conda and Python. You can still install any of the available packages with conda, that comes by default with the standard version. Conda is a program we will be using exclusively from the command line, so if you aren't comfortable using it, check out these learn by doing videos on Lynda.com command prompt tutorial for Windows and Linux Command Line Basics for Mac OSX/Linux

### Installing Anaconda

Anaconda is available for Windows, Mac OS X, and Linux. You can find the installers and installation instructions at https://www.continuum.io/downloads If you already have Python installed on your computer, this won't break anything. Instead, the default Python used by your scripts and programs will be the one that comes with Anaconda. Choose the Python 3.5 version, you can install Python 2 versions later. Also, choose the 64-bit installer if you have a 64-bit operating system, otherwise go with the 32-bit installer. Go ahead and choose the appropriate version, then install it. Continue on afterward!

After installation, you’re automatically in the default conda environment with all packages installed which you can see below. You can check out your own install by entering conda list into your terminal.

## 4.2 TECHNICAL TOOL KITS USED :

### 4.2.1 PYTHON:

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- Python is Interpreted: Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP

. • Python is Interactive: You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

- Python is Object-Oriented: Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- Python is a Beginner's Language: Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

### History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands. Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, Smalltalk, Unix shell, and other scripting languages. Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL). Python is now maintained by a core development team at the institute, although Guidovan Rossum still holds a vital role in directing its progress.

Python Features

Python's features include:

- Easy-to-learn: Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.

- Easy-to-read: Python code is more clearly defined and visible to the eyes.
- Easy-to-maintain: Python's source code is fairly easy-to-maintain.
- A broad standard library: Python's bulk of the library is very portable and cross platform compatible on UNIX, Windows, and Macintosh.
- Interactive Mode: Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- Portable: Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- Extendable: You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- Databases: Python provides interfaces to all major commercial databases.

- GUI Programming: Python supports GUI applications that can be created and ported to many system calls, libraries, and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- Scalable: Python provides a better structure and support for large programs than shell scripting.
- Apart from the above-mentioned features, Python has a big list of good features, few are listed below:
- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

## 4.2.2 JUPYTER NOTE BOOK:

The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. The Jupyter notebook combines two components: A web application: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output. Notebook documents: a representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects. See also: See the installation guide on how to install the notebook and its dependencies.

The Jupyter Notebook is an incredibly powerful tool for interactively developing and presenting data science projects. This article will walk you through how to use Jupyter Notebooks for data science projects and how to set it up on your local machine.

First, though: **what is a “notebook”?**

A notebook integrates code and its output into a single document that combines visualizations, narrative text, mathematical equations, and other rich media. In otherwords: it’s a single document where you can run code, display the output, and also add explanations, formulas, charts, and make your work more transparent, understandable, repeatable, and shareable.

Using Notebooks is now a major part of the data science workflow at companies across the globe. If your goal is to work with data, using a Notebook will speed up your workflow and make it easier to communicate and share your results.

Best of all, as part of the open source Project Jupyter, Jupyter Notebooks are completely free. You can download the software on its own, or as part of the Anaconda data science toolkit.

Although it is possible to use many different programming languages in Jupyter Notebooks, this article will focus on Python, as it is the most common use case. (Among R users, R Studio tends to be a more popular choice).

### Installation

The easiest way for a beginner to get started with Jupyter Notebooks is by installing Anaconda.

Anaconda is the most widely used Python distribution for data science and comes pre-loaded with all the most popular libraries and tools.

Some of the biggest Python libraries included in Anaconda include NumPy, pandas, and Matplotlib, though the full 1000+ list is exhaustive.

Anaconda thus lets us hit the ground running with a fully stocked data science workshop without the hassle of managing countless installations or worrying about dependencies and OS-specific (read: Windows-specific) installation issues.

To get Anaconda, simply:

1. Download the latest version of Anaconda for Python 3.8.
2. Install Anaconda by following the instructions on the download page and/or in the executable.

If you are a more advanced user with Python already installed and prefer to manage your packages manually, you can just use pip:

### 4.2.3 PYCHARM:

**PyCharm** is an integrated development environment (IDE) used in computer programming, specifically for the Python programming language. It is developed by the Czech company JetBrains (formerly known as IntelliJ). It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems , and supports web development with Django as well as data science with Anaconda.

PyCharm is cross-platform, with Windows, macOS and Linux versions. The Community Edition is released under the Apache License, and there is also an educational version, as well as a Professional Edition with extra features (released under a subscription- funded proprietary license)

- Project and code navigation: specialized project views, file structure views and quick jumping between files, classes, methods and usages
- Coding assistance and analysis, with code completion, syntax and error highlighting, linter integration, and quick fixes
- Python refactoring: includes rename, extract method, introduce variable, introduce constant, pull up, push down and others
- Support for web frameworks: Django, web2py and Flask [professional edition only]
- Integrated Python debugger

- Integrated unit testing, with line-by-line code coverage
- Google App Engine Python development [professional edition only]
- Version control integration: unified user interface for Mercurial, Git, Subversion, Perforce and CVS with change lists and merge
- Support for scientific tools like Matplotlib, NumPy and SciPy [professional edition only]

## History:

History helps you constantly track all changes made to a project independently of version control.

Unlike version control systems, which only keep track of the differences made between commits, Local History offers much more. It automatically records your project's state as you edit code, run tests, deploy applications, and so on, and maintains revisions for all meaningful changes made both from the IDE and externally.

Acting as your personal version control system, Local History lets you restore deleted files, bring back separate changes, or roll back to any state of a file even if no version control is enabled for your project yet, or if an unwanted change was made after your last commit. It may also serve as a recovery source if your computer restarts unexpectedly before you can take any action.

## Restore changes

Let's imagine you made a series of changes to a file since your last commit before your realized you've deleted a meaningful chunk of code. The Undo action can't help you here because that change is too far away and you'll be forced to discard other changes if you use it.

With PyCharm you can restore that change in a couple of clicks.

1. Right-click anywhere in the editor and choose Local History Show History from the context menu. In the dialog that opens, the left-hand pane shows a list of all saved revisions of the current file with timestamps. The right-hand pane shows a diff viewer which displays the differences between each revision and the current state of the file.

2. Do one of the following:
    a. To revert the whole file or directory to the state of this revision, right-click it and choose Revert from the context menu or click on the toolbar.
    b. To restore a specific code fragment, select the revision that contains that fragment. In the diff view on the right locate the piece of code you want to restore click the chevron button to copy it from the left pane.

## 4.3 LIBRARIES USED:

### 4.3.1 Nearest Neighbours:

The intuition underlying Nearest Neighbour Classification is quite straightforward, examples are classified based on the class of their nearest neighbours. It is often useful to take more than one neighbour into account so the technique is morecommonly referred to as k-Nearest Neighbour (k-NN) Classification where k nearest neighbours are used in determining the class. Since the training examples are needed at run-time, i.e. they need to be in memory at run-time, it is sometimes also called Memory- Based Classification. Because induction is delayed to run time, it is considered a Lazy Learning technique. Because classification is based directly on the training examples it is also called Example-Based Classification or Case-Based Classification. The basic idea is asshown in Figure 1 which depicts a 3-Nearest Neighbour Classifier on a two-class problem in a two-dimensional feature space. In this example the decision for q1 is straightforward – all three of its nearest neighbours are of class O so it is classified as an O.

The situation for q2 is a bit more complicated at it has two neighbours of class X and one of class O. This can be resolved by simple majority voting or by distance weighted voting (see below). So k−NN classification has two stages; the first is the determination of the nearest neighbours and the second is the determination of the class using those neighbours.

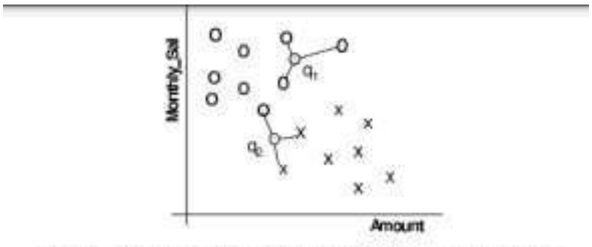


Fig. 1. A simple example of 3-Nearest Neighbour Classification

Let us assume that we have a training dataset D made up of (xi) I ∈[1,|D|] training samples. The examples are described by a set of features F and any numeric features have been normalised to the range [0,1]. Each training example is labelled with a class label yj ∈ Y . Our objective is to classify an unknown example q. For each xi ∈ D we can calculate the distance between q and xi as follows:

$$d(\mathbf{q},\mathbf{x}_i) = \sum_{f \in F} w_f \delta(\mathbf{q}_f, \mathbf{x}_{if})$$

There are a large range of possibilities for this distance metric; a basic version for continuous and discrete attributes would be:

$$\delta(\mathbf{q}_f, \mathbf{x}_{if}) = \begin{cases} 0 & f \text{ discrete and } \mathbf{q}_f = \mathbf{x}_{if} \\ 1 & f \text{ discrete and } \mathbf{q}_f \neq \mathbf{x}_{if} \\ |\mathbf{q}_f - \mathbf{x}_{if}| & f \text{ continuous} \end{cases}$$

The k nearest neighbours are selected based on this distance metric. Then there are a variety of ways in which the k nearest neighbours can be used to determine the class of q. The most straightforward approach is to assign the majority class among the nearest neighbours to the query. It will often make sense to assign more weight to the nearer neighbours in deciding the class of the query. A fairly general technique to achieve this is distance weighted voting where the neighbours get to vote on the class of the query case with votes weighted by the inverse of their distance to the query.

$$Vote(y_j) = \sum_{c=1}^{k} e^{-\frac{d(\mathbf{q},\mathbf{x}_c)}{h}} 1(y_j, y_c)$$

Thus the vote assigned to class $y_j$ by neighbour xc is 1 divided by the distance to that neighbour, i.e. 1($y_j$ , $y_c$) returns 1 if the class labels match and 0 otherwise. In equation 3 n would normally be 1 but values greater than 1 can be used to further reduce the influence of more distant neighbours.

Another approach to voting is based on Shepard's work [25] and uses an exponential function rather than inverse distance, i.e.

$$Vote(y_j) = \sum_{c=1}^{k} e^{-\frac{d(\mathbf{q},\mathbf{x}_c)}{h}} 1(y_j, y_c)$$

In this paper we consider three important issues that arise with the use of k-NN classifiers. In the next section we look at the core issue of similarity and distance measures and explore some exotic (dis)similarity measures to illustrate the generality of the k-NN idea. In section 3 we look at computational complexity issues and review some speed-uptechniques for k-NN. In section 4 we look at dimension reduction – both feature selection and sample selection. Dimension reduction is of particular importance with k-NN as it hasa big impact on computational performance and accuracy. The paper concludes with a summary of the advantages and disadvantages of k-NN.

### 4.3.2 TFIDF

TF-IDF stands for term frequency-inverse document frequency and it is a measure, used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc.) in a document amongst a collection of documents (also known as a corpus).

### Overview of TF-IDF

TF-IDF can be broken down into two parts TF (term frequency) and *IDF* (inverse document frequency).

What is TF (term frequency)?

Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document. There are multiple measures, or ways, of defining frequency:

Number of times the word appears in a document (raw count).

Term frequency adjusted for the length of the document (raw count of occurrencesdivided by number of words in the document).

Logarithmically scaled frequency (e.g. log(1 + raw count)).

Boolean frequency (e.g. 1 if the term occurs, or 0 if the term does not occur, in the document).

What is IDF (inverse document frequency)?

Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus. IDF is calculated as follows where ***t*** is the term (word) we are looking to measure the commonness of and ***N*** is the number of documents (d) in the corpus (D). The denominator is simply the number of documents in which the term, ***t***, appears in.

Note: It can be possible for a term to not appear in the corpus at all, which can result in a divide-by-zero error. One way to handle this is to take the existing count and add 1.Thus making the denominator (1 + count). An example of how the popular library scikit-learn handles this can be seen below.

$$idf(t, D) = \log \left( \frac{N}{count\ (d \in D : t \in d)} \right)$$

Scikit-Learn

- $\mathrm{IDF}(t) = \log \frac{1+n}{1+df(t)} + 1$

Standard notation

- $\mathrm{IDF}(t) = \log \frac{n}{df(t)}$

The reason we need IDF is to help correct for words like “of”, “as”, “the”, etc. since they appear frequently in an English corpus. Thus by taking inverse document frequency, we can minimize the weighting of frequent terms while making infrequent terms have a higher impact.

Finally IDFs can also be pulled from either a background corpus, which corrects for sampling bias, or the dataset being used in the experiment at hand.

**Putting it together: TF-IDF**

To summarize the key intuition motivating TF-IDF is the importance of a term is inversely related to its frequency across documents.TF gives us information on how often a term appears in a document and IDF gives us information about the relative rarity of a term in the collection of documents. By multiplying these values together we can get our final TF-IDF value.

$$tf\,idf\,(t, d, D) = tf\,(t, d)\ .\ idf\,(t, D)$$

## 4.3.3 SK-LEARN:

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon **NumPy, SciPy** and **Matplotlib**.

## Features:

Rather than focusing on loading, manipulating and summarizing data, Scikit-learn library is focused on modeling the data. Some of the most popular groups of models provided by

**Supervised Learning algorithms** − Almost all the popular supervised learning algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree etc., are the part of scikit-learn.

**Unsupervised Learning algorithms** − On the other hand, it also has all the popular unsupervised learning algorithms from clustering, factor analysis, PCA (Principal Component Analysis) to unsupervised neural networks.

**Clustering** − This model is used for grouping unlabeled data.

**Cross Validation** − It is used to check the accuracy of supervised models on unseen data.

**Dimensionality Reduction** − It is used for reducing the number of attributes in datawhich can be further used for summarization, visualization and feature selection.

**Ensemble methods** − As name suggest, it is used for combining the predictions ofmultiple supervised models.

**Feature extraction** − It is used to extract the features from data to define the attributes in image and text data.

**Feature selection** − It is used to identify useful attributes to create supervised models.

**Open Source** − It is open source library and also commercially usable under BSD License.

## Dataset Loading

A collection of data is called dataset. It is having the following two components −

**Features** − The variables of data are called its features. They are also known as predictors, inputs or attributes.

**Feature matrix** − It is the collection of features, in case there are more than one.

**Feature Names** − It is the list of all the names of the features.

**Response** − It is the output variable that basically depends upon the feature variables. They are also known as target, label or output.

**Response Vector** − It is used to represent response column. Generally, we have just one response column.

**Target Names** − It represent the possible values taken by a response vector.

Scikit-learn have few example datasets like **iris** and **digits** for classification and the **Boston house prices** for regression.

### Installation:

Scikit-learn requires:

NumPy

SciPy as its dependencies.

Before installing scikit-learn, ensure that you have NumPy and SciPy installed. Onceyou have a working installation of NumPy and SciPy, the easiest way to install scikit- learn is using pip:

```
pip install -U scikit-learn
```

Let us get started with the modelling process now.

**Step 1: Load a dataset**

A dataset is nothing but a collection of data. A dataset generally has two main components:

**Features**: (also known as predictors, inputs, or attributes) they are simply the variables of our data. They can be more than one and hence represented by a **feature matrix** (‘X’ is a common notation to represent feature matrix). A list of all the feature names is termed **feature names**.

**Response**: (also known as the target, label, or output) This is the output variable depending on the feature variables. We generally have a single response column and it is represented by a **response vector** (‘y’ is a common notation to represent response vector). All the possible values taken by a response vector are termed **target names**.

**Loading exemplar dataset:** scikit-learn comes loaded with a few example datasets like the iris and digits datasets for classification and the boston house prices dataset for regression.

**Step 2: Splitting the dataset**

One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model.

But this method has several flaws in it, like:

The goal is to estimate the likely performance of a model on **out-of-sample** data.

Maximizing training accuracy rewards overly complex models that won't necessarily generalize our model.

Unnecessarily complex models may over-fit the training data.

A better option is to split our data into two parts: the first one for training our machine learning model, and the second one for testing our model.

**To summarize:**

Split the dataset into two pieces: a training set and a testing set.

Train the model on the training set.

Test the model on the testing set, and evaluate how well our model did.

**Advantages of train/test split:**

The model can be trained and tested on different data than the one used for training.

Response values are known for the test dataset, hence predictions can be evaluated

Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

## 4.3.4 FLASK:

**Flask** is a micro web framework written in Python. It is classified asa microframework because it does not require particular tools or libraries. It hasno database abstraction layer, form validation, or any other components where pre- existing third-party libraries provide common functions. Flask is a web framework, it's a Python module that lets you develop web applications easily. It's has a small and easy-to- extend core: it's a microframework that doesn't include an ORM (Object Relational Manager) or such features.

## History:

Flask was created by Armin Rancher of Pocoo, an international group of Python enthusiasts formed in 2004. According to Rancher, the idea was originally an April Fool's joke that was popular enough to make into a serious application. The name is a play on the earlier Bottle framework.

When Rancher and Georg Brand created a bulletin board system written in Python in 2004, the Pocoo projects Werkzeug and Jinja were developed.

In April 2016, the Pocoo team was disbanded and development of Flask and related libraries passed to the newly formed Pallets project.

Flask has become popular among Python enthusiasts. As of October 2020, it has second most stars on GitHub among Python web-development frameworks, only slightly behind Django, and was voted the most popular web framework in the Python Developers Survey 2018.

## Features

Development server and debugger

Integrated support for unit testing

RESTful request dispatching

Uses Jinja templating

Support for secure cookies (client-side sessions)

100% WSGI 1.0 compliant

Unicode-based

Complete documentation

Google App Engine compatibility

Extensions available to extend functionality

## 4.3.5 RESUME PARSING

A Résumé parsing technology converts an unstructured form of resume data into a structured format. A Resume parser analyses resume data and extract it into machine-readable output such as XML, JSON. A CV/resume parser helps automatically store, organize, and analyze resume data to find the best candidate.

A Resume Parser helps organizations eliminate the error-prone and time-consuming process and improves recruiters' efficiency.

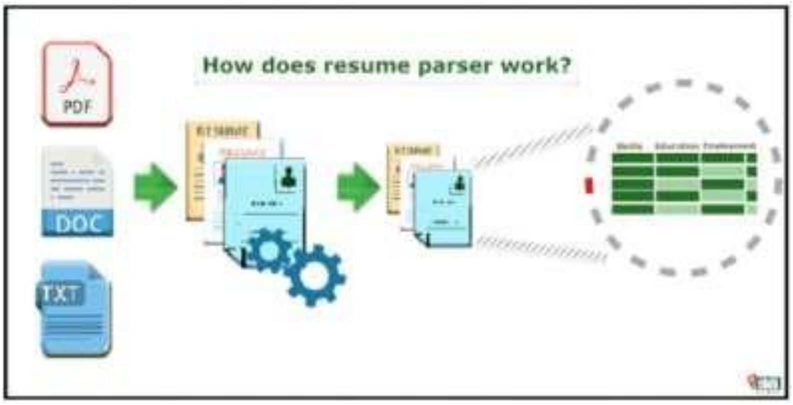
I am the HR Manager of an Enterprise. On average, I receive thousands of resumes per year. It is a challenging task to handle resumes manually. I heard about resume parsing technology. I was wondering **what a resume parser is?** And **how does it automatically parse resumes?** I took a demo with one of the leading resume parser providers and used parsing software to parse information from resumes in bulk.

**What is a Resume Parser?**

A resume parser is a deep learning/AI framework that identifies complete information from resumes, analyses, store, organize, and enriches it through its taxonomies. Resume parsing software makes the hiring process quicker and more productive.

Fast and accurate **resume parsing** technology improves efficiency and offers an enhanced candidate experience.

What Does a CV/Resume Parser Do?



1. A **resume parser** is a **compiler or interpreter** that converts the unstructured data into a structured form.
2. It is a component that **automatically segregates the information** into **various fields** and parameters like contact information, educational qualification, work experience, skills, achievements, professional certifications to quickly help you identify the most relevant resumes based on your criteria.
3. A parser takes input in the form of a sequence of program instructions and tends to build a data structure, a **"parse tree,"** or an abstract syntax tree.

## 4.4 PACKAGES USED:

### 4.4.1 NumPy

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

NumPy is a Python package. It stands for ‘Numerical Python’. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

**Numeric**, the ancestor of NumPy, was developed by Jim Hugunin. Another package

Num array was also developed, having some additional functionalities. In 2005, Travis Oliphant created NumPy package by incorporating the features of Num array into Numeric package. There are many contributors to this open-source project.

**History**

The Python programming language was not originally designed for numerical computing, but attracted the attention of the scientific and engineering community early on. In 1995 the special interest group (SIG) matrix-sig was founded with the aim of definingan array computing package; among its members was Python designer and maintainer Guido van Rossum, who extended Python's syntax (in particular the indexing syntax) to make array computing easier.

An implementation of a matrix package was completed by Jim Fulton, then generalized by Jim Hugunin and called (also variously known as the "Numerical Python extensions" or "NumPy"). Hugunin, a graduate student at the Massachusetts Institute of Technology (MIT), joined the Corporation for National Research Initiatives (CNRI) in 1997 to work on J Python, leaving Paul Dubois of Lawrence Livermore National Laboratory (LLNL) to take over as maintainer. Other early contributors include David Ascher, Konrad Hinson and Travis Oliphant.

A new package called *Num array* was written as a more flexible replacement for Numeric. Like Numeric, it too is now deprecated. Num array had faster operations for large arrays, but was slower than Numeric on small ones, so for a time both packageswere used in parallel for different use cases. The last version of Numeric (v24.2) was released on 11 November 2005, while the last version of Num array (v1.5.2) was released on 24 August 2006.

There was a desire to get Numeric into the Python standard library, but Guido van Rossum decided that the code was not maintainable in its state then.

In early 2005, NumPy developer Travis Oliphant wanted to unify the community around a single array package and ported Num array's features to Numeric, releasing the result as NumPy 1.0 in 2006. This new project was part of SciPy. To avoid installing the large SciPy package just to get an array object, this new package was separated and called NumPy. Support for Python 3 was added in 2011 with NumPy version 1.5.0.

In 2011, PyPy started development on an implementation of the NumPy API for PyPy.] It is not yet fully compatible with NumPy.

## Features

NumPy targets the C Python reference implementation of Python, which is a non-optimizing bytecode interpreter. Mathematical algorithms written for this version ofPython often run much slower than compiled equivalents due to the absence of compiler optimization. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays; using these requires rewriting some code, mostly inner loops, using NumPy.

Using NumPy in Python gives functionality comparable to MATLAB since they are both interpreted, and they both allow the user to write fast programs as long as most operations work on arrays or matrices instead of scalars. In comparison, MATLAB boasts a large number of additional toolboxes, notably Simulink, whereas NumPy is intrinsically integrated with Python, a more modern and complete programming language. Moreover, complementary Python packages are available; SciPy is a library that adds more MATLAB-like functionality and Matplotlib is a plotting package that provides MATLAB-like plotting functionality. Internally, both MATLAB and NumPy relyon BLAS and LAPACK for efficient linear algebra computations.

Python bindings of the widely used computer vision library OpenCV utilize NumPyarrays to store and operate on data. Since images with multiple channels are simply represented as three-dimensional arrays, indexing, slicing or masking with other arraysare very efficient ways to access specific pixels of an image. The NumPy array asuniversal data structure in OpenCV for images, extracted feature points, filter kernels and many more vastly simplifies the programming workflow and debugging.

## Limitations

Inserting or appending entries to an array is not as trivially possible as it is with Python's lists. The nipped(...) routine to extend arrays actually creates new arrays of the desired shape and padding values, copies the given array into the new one and returns it. NumPy's np.concatenate([a1,a2]) operation does not actually link the two arrays but returns a new one, filled with the entries from both given arrays in sequence. Reshaping the dimensionality of an array with np.reshape(...) is only possible as long as the number of elements in the array does not change. These circumstances originate from the fact that NumPy's arrays must be views on contiguous memory buffers. A replacement package called Blaze attempts to overcome this limitation.

Algorithms that are not expressible as a vectorized operation will typically run slowly because they must be implemented in "pure Python", while vectorization mayincrease memory complexity of some operations from constant to linear, because temporary arrays must be created that are as large as the inputs. Runtime compilation of numerical code has been implemented by several groups to avoid these problems; open source solutions that interoperate with NumPy include SciPy. weave , numeri and Numb. Cython and Pythran are static-compiling alternatives to these.

Many modern large-scale scientific computing applications have requirements that exceed the capabilities of the NumPy arrays. For example, NumPy arrays are usually loaded into a computer's memory, which might have insufficient capacity for the analysis of large datasets. Further, NumPy operations are executed on a single CPU. However, many linear algebra operations can be accelerated by executing them on clusters of CPUs or of specialized hardware, such as GPUs and TPUs, which many deep learning applications rely on. As a result, several alternative array implementations have arisen in the scientific python ecosystem over the recent years, such as Disk for distributed arraysand TensorFlow or JAX for computations on GPUs. Because of its popularity, these often implement a subset of NumPy’s API or mimic it, so that users can change their array implementation with minimal changes to their code required. A recently introduced library named CuPy, accelerated by Nvidia's CUDA framework, has also shown potential for faster computing, being a 'drop-in replacement' of NumPy.

## 4.4.2 PANDAS:

**Pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "**panel data**",an econometrics term for data sets that include observations over multiple time periods forthe same individuals. Its name is a play on the phrase "Python data analysis" itself. Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010.

**Library features**

- Data Frame object for data manipulation with integrated indexing.
- Tools for reading and writing data between in-memory data structures and different file formats.

- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of data sets.
- Label-based slicing, fancy indexing, and sub setting of large data sets.
- Data structure column insertion and deletion.
- Group by engine allowing split-apply-combine operations on data sets.
- Data set merging and joining.
- Hierarchical axis indexing to work with high-dimensional data in a lower-dimensional data structure.
- Time series-functionality: Date range generation and frequency conversions, moving window statistics, moving window linear regressions, date shifting and lagging.
- Provides data filtration.

The library is highly optimized for performance, with critical code paths written in Cython or C.

**Dataframes**

Pandas is mainly used for data analysis and associated manipulation of tabular data in Dataframes. Pandas allows importing data from various file formats such as comma-separated values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features. Thedevelopment of pandas introduced into Python many comparable features of working withDataframes that were established in the R programming language. The pandas library is built upon another library NumPy, which is oriented to efficiently working with arrays instead of the features of working on Dataframes.

**History**

Developer Wes McKinney started working on pandas in 2008 while at AQR Capital Management

out of the need for a high performance, flexible tool to perform quantitative analysis on financial data. Before leaving AQR he was able to convince management to allow him to open source the library.

Another AQR employee, Chang She, joined the effort in 2012 as the second major contributor to the library.

In 2015, pandas signed on as a fiscally sponsored project of NumFOCUS, a 501(c)(3) non-profit charity in the United States.

### 4.4.3 REGULAR EXPRESSION

A **regular expression** (shortened as **regex** or **reexpel**; sometimes referred to as **rational expression**) is a sequence of characters that specifies a search pattern in text. Usuallysuch patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation. Regular expression techniques are developed in theoretical computer science and formal language theory.

The concept of regular expressions began in the 1950s, when the American mathematician Stephen Cole Kleene formalized the concept of a regular language. They came into common use with Unix text-processing utilities. Different syntaxes for writing regular expressions have existed since the 1980s, one being the POSIX standard andanother, widely used, being the Perl syntax.

Regular expressions are used in search engines, in search and replace dialogs of word processors and text editors, in text processing utilities such as sed and AWK, andin lexical analysis. Most general-purpose programming languages support regex capabilities either natively or via libraries, including for example Python, C, C++, Java, and JavaScript.

### History

Stephen Cole Kleene, who introduced the concept Regular expressions originated in 1951, when mathematician Stephen Cole Kleene described regular languages using his mathematical notation called regular events. These arose in theoretical computer science, in the subfields of automata theory (models of computation) and the description and classification of formal languages. Other early implementations of pattern matching include the SNOBOL language, which did not use regular expressions, but instead its own pattern matching constructs.

Regular expressions entered popular use from 1968 in two uses: pattern matching in a text editor and lexical analysis in a compiler. Among the first appearances of regular expressions in program form was when Ken Thompson built Kleene's notation into the editor QED as a means to match patterns in text files. For speed, Thompson implemented regular expression matching by just-in-time compilation (JIT) to IBM 7094 code on the Compatible Time-Sharing System, an important early example of JIT compilation. He later added this capability to the Unix editor ed, which eventually led to the popularsearch tool grep's use of regular expressions ("grep" is a word derived from the command for regular expression searching in the ed editor: g/*re*/p meaning "Global search for Regular Expression and Print matching lines"). Around the same time when Thompson developed QED, a group of researchers including Douglas T. Ross implemented a tool based on regular expressions that is used for lexical analysis in compiler design.

Many variations of these original forms of regular expressions were usedin Unix programs at Bell Labs in the 1970s, including vi, lex, sed , AWK, and expr, andin other programs such as Emacs. Regexes were subsequently adopted by a wide range of programs, with these early forms standardized in the POSIX.2 standard in 1992.

In the 1980s, the more complicated regexes arose in Perl, which originally derived from a regex library written by Henry Spencer (1986), who later wrote an implementation of Advanced.

Regular Expressions for Tcl. The Tcl library is a hybrid NFA/DFA implementation with improved performance characteristics. Software projects that have adopted Spencer's Tcl regular expression implementation include PostgreSQL. Perl later expanded on Spencer's original library to add many new features. Part of the effort in the design of Raku (formerly named Perl 6) is to improve Perl's regex integration, and to increase their scope and capabilities to allow the definition of parsing expression grammars.[21] Theresult is a mini-language called Raku rules, which are used to define Raku grammar as well as provide a tool to programmers in the language. These rules maintain existing features of Perl 5.x regexes, but also allow BNF-style definition of a recursive descent parser via sub-rules.

The use of regexes in structured information standards for document and database modelling started in the 1960s and expanded in the 1980s when industry standards like ISO SGML (precursored by ANSI "GCA 101-1983") consolidated. The kernel ofthe structure specification language standards consists of regexes. Its use is evident inthe DTD element group syntax. Prior to the use of regular expressions, many search languages allowed simple wildcards, for example "*" to match any sequence of characters, and "?" to match a single character. Relics of this can be found today in the glob syntax for filenames, and in the SQL LIKE operator.

Starting in 1997, Philip Hazel developed PCRE (Perl Compatible Regular Expressions), which attempts to closely mimic Perl's regex functionality and is used by many modern tools including PHP and Apache HTTP Server.

Today, regexes are widely supported in programming languages, text processing programs (particularly leers), advanced text editors, and some other programs. Regex support is part of the standard library of many programming languages, including Java and Python, and is built into the syntax of others, including Perl and ECMAScript. Implementations of regex functionality is often called a **regex engine**, and a number of libraries are available for reuse. In the late 2010s, several companies started to offer hardware, FPGA, GPU implementations of PCRE compatible **regex engines** that are faster compared to CPU implementations**.**

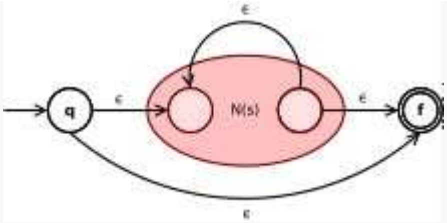
**Patterns**

The phrase *regular expressions*, or *regexes*, is often used to mean the specific, standard textual syntax for representing patterns for matching text, as distinct from the mathematical notation described below. Each character in a regular expression (that is, each character in the string describing its pattern) is either a metacharacter, having a special meaning, or a regular character that has a literal meaning. For example, in the

regex b., 'b' is a literal character that matches just 'b', while '.' is a metacharacter that matches every character except a newline. Therefore, this regex matches, for example, 'b%', or 'bx', or 'b5'. Together, metacharacters and literal characters can be used to identifytext of a given pattern or process a number of instances of it. Pattern matches may vary from a precise equality to a very general similarity, as controlled by the metacharacters.
For example, is a very general pattern, [a-z] (match all lower-case letters from 'a' to 'z') is less general and b is a precise pattern (matches just 'b'). The metacharacter syntax is designed specifically to represent prescribed targets in a concise and flexible way to direct the automation of text processing of a variety of input data, in a form easy to type using a standard ASCII keyboard.

A very simple case of a regular expression in this syntax is to locate a word spelled two different ways in a text editor, the regular expression serial [Sze] matches both "serialise" and "serialize". Wildcard characters also achieve this, but are more limited in what they can pattern, as they have fewer metacharacters and a simple language-base.

The usual context of wildcard characters is in globing similar names in a list of files, whereas regexes are usually employed in applications that pattern-match text strings in general. For example, the regex ^[ \t]+|[ \t]+$ matches excess whitespace at the beginning or end of a line. An advanced regular expression that matches any numeral is [+-]?(\d+(\.\d*)?|\.\d+)([ewe][+-]?\d+)?.

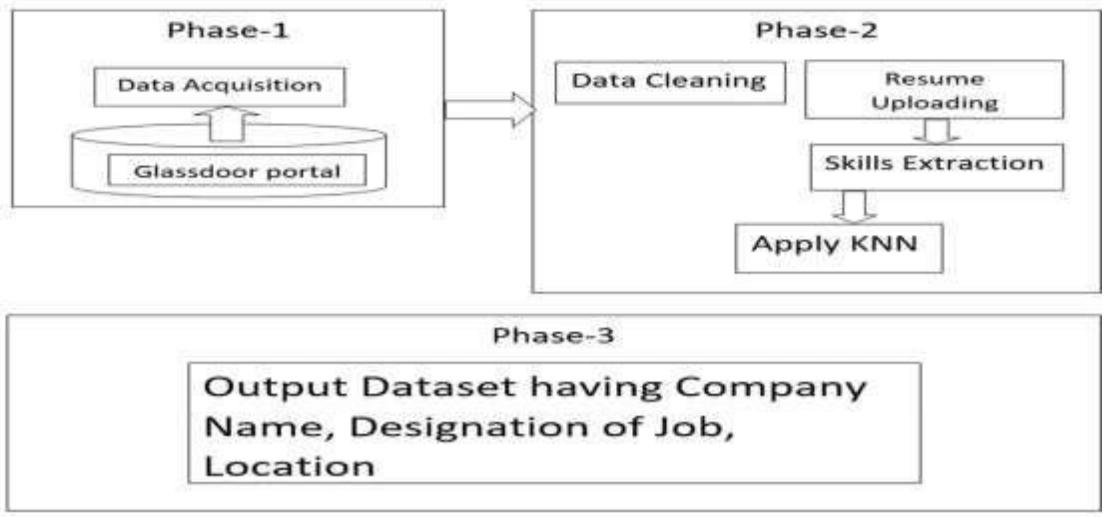


Translating the Kleene star
(*s** means "zero or more of *s*")

A **regex processor** translates a regular expression in the above syntax into an internal representation that can be executed and matched against a string representing the text being searched in. One possible approach is the Thompson's construction algorithm to construct a nondeterministic finite automaton (NFA), which is then made deterministic and the resulting deterministic finite automaton (DFA) is run on the target text string to recognize substrings that match the regular expression. The picture shows the NFA scheme *N*(*s**) obtained from the regular expression *s**, where *s* denotes a simpler regular expression in turn, which has already been recursively translated to the NFA *N*(*s*).

# CHAPTER 5

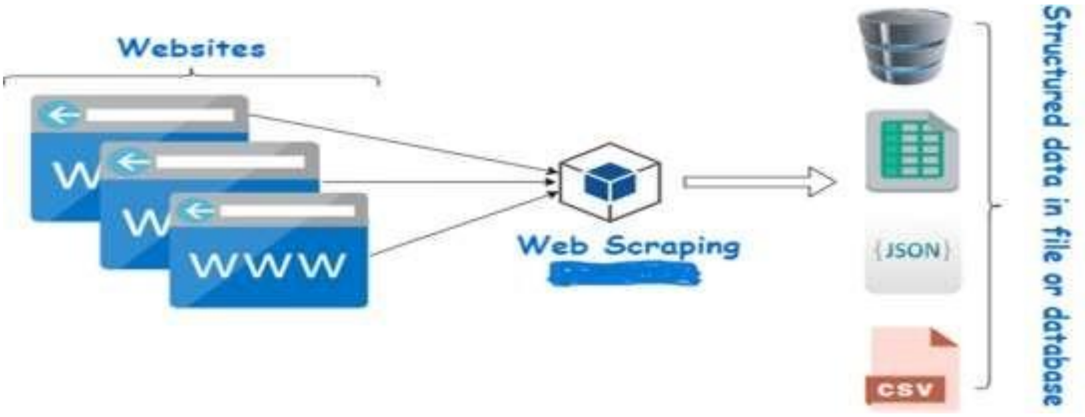
**FLOWCHART:**



## 5.1 WEB SCRAPING

Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. There are many different ways to perform web scraping to obtain data from websites. These include using online services, particular API's or even creatingyour code for web scraping from scratch. Many large websites, like Google, Twitter, Facebook, Stack Overflow, etc. have API's that allow you to access their data in a structured format. This is the best option, but there are other sites that don't allow users to access large amounts of data in a structured form or they are simply not that technologically advanced. In that situation, it's best to use Web Scraping to scrape the website for data.

Web scraping requires two parts, namely the **crawler** and the **scraper**. The crawler is an artificial intelligence algorithm that browses the web to search for the particular data required by following the links across the internet. The scraper, on the other hand, is a specific tool created to extract data from the website. The design of the scraper can vary greatly according to the complexity and scope of the project so that it can quickly and accurately extract the data.



**How  Web Scrapers Work?**

Web Scrapers can extract all the data on particular sites or the specific data that a user wants. Ideally, it’s best if you specify the data, you want so that the web scraper only extracts that data quickly. For example, you might want to scrape an  Amazon  page for the types of juicers available, but you might only want the data about the models of different juicers and not the customer reviews.

So, when a web scraper needs to scrape a site, first the URLs are provided. Then it loads all the HTML code for those sites and a more advanced scraper might even extract allthe CSS and JavaScript elements as well. Then the scraper obtains the required datafrom this HTML code and outputs this data in the format specified by the user. Mostly, this is in the form of an Excel spreadsheet or a CSV file, but the data can also be saved in other formats, such as a JSON file.

**Different Types of Webs Scrapers**

Web Scrapers can be divided on the basis of many different  criteria, including Self-built or Pre-built Web Scrapers, Browser extension or Software Web Scrapers, and Cloud or Local Web Scrapers.

You can have **Self-built Web Scrapers** but that requires advanced knowledge of programming. And if you want more features in your Web Scrapper, then you need even more knowledge. On the other hand, pre-built **Web Scrapers** are previously created scrapers that you can download and run easily. These also have more advanced options that you can customize.

**Browser extensions Web Scrapers** are extensions that can be added to your browser. These are easy to run as they are integrated with your browser, but at the same time, they are also limited because of this. Any advanced features that are outside the scope of your browser are impossible to run on Browser extension Web Scrapers. But **Software Web Scrapers** don't have these limitations as they can be downloaded and installed on your computer. These are more complex than Browser web scrapers, but they also have advanced features that are not limited by the scope of your browser.

**Cloud Web Scrapers** run on the cloud, which is an off-site server mostly provided by the company that you buy the scraper from. These allow your computer to focus on other tasks as the computer resources are not required to scrape data from websites. **Local Web Scrapers**, on the other hand, run on your computer using local resources. So, if the Web scrapers require more CPU or RAM, then your computer will become slow and not be able to perform other tasks.

## 5.2 DATA CLEANING:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

### STEPS INVOLVED IN DATA CLEANING:

Step 1: Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyse. For example, if you want to analyse data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient and minimize distraction from your primary target—as well as creating a more manageable and more performant dataset.

Step 2: Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabelled categories or classes. For example, you may find "N/A" and "Not Applicable" both appear, but they should be analysed as the same category.

Step 3: Filter unwanted outliers

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analysing. If you have a legitimate reason to remove an outlier,like improper data-entry, doing so will help the performance of the data you are working with.

However, sometimes it is the appearance of an outlier that will prove a theory you are working on. Remember: just because an outlier exists, doesn't mean it is incorrect.

This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

Step 4: Handle missing data

You can't ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.

As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.

As a third option, you might alter the way the data is used to effectively navigate null values.

**Removing stop words with NLTK :**

The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

What are Stop words?
Stop Words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.
We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stop words stored in 16 different languages. You can find them in the nltk_data directory.

To check the list of stopwords you can type the following commands in the python shell.
Removing stop words with NLTK
The following exemplar program for removing stop words from a piece of text:

```
from nltk.corpus import stopwords
from nltk.tokenize import word tokenize

example sent = """This is a sample sentence, showing off the stop words filtration."""

stopwords = set(stopwords. Words('English'))

word tokens = word tokenize(example sent)

filtered sentence = [w for w in word tokens if not wallower() in stopwords]

filtered_sentence = []

for w in word_tokens:
        if w not in stop_words:
                filtered_sentence. Append(w)

print(word_tokens)
print(filtered_sentence)
```

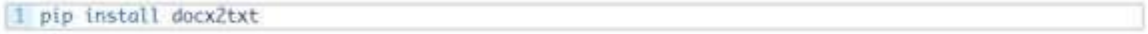
## 5.3 SKILL EXTRACTION :

We are using Python 3 for its wide range of libraries that is already available and for its general acceptance in the data sciences area.

We are also be using nltk for NLP (natural language processing) tasks such as stop word filtering and tokenization, docx2txt a for extracting text from PDF formats.

Extracting text from docx files

In order to extract text from docx files, the procedure is pretty similar to what we've done for PDF files. Let's install the required dependency (docx2txt) using pip and then write some code to do the actual work.

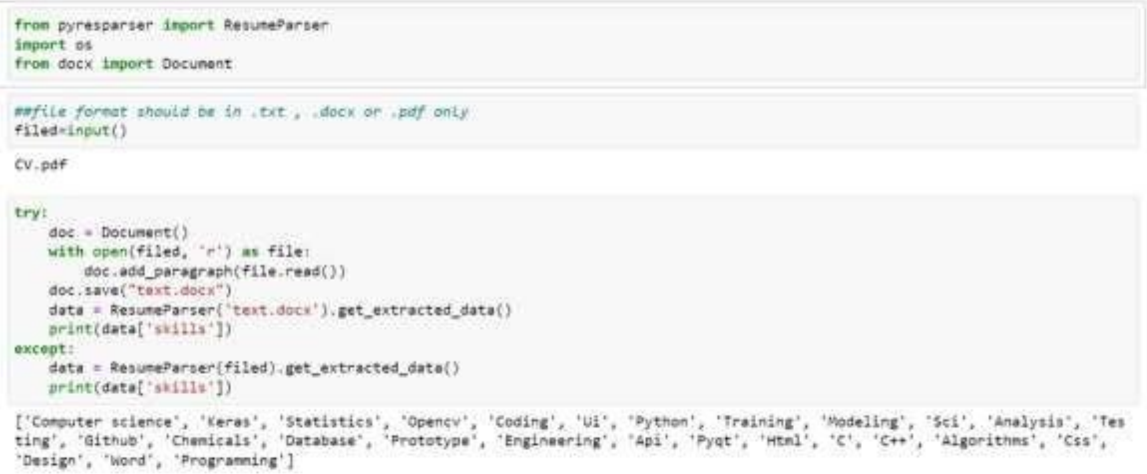


```
pip install docx2txt
```

Extracting skills from the resumes

This is the section where things get trickier. Exporting skills from a text is a verychallenging task and in order to increase accuracy, we need a database or an API to verify if a text is a skill or not.

And the code is as follows:



```
from pyresparser import ResumeParser
import os
from docx import Document
```

```
##file format should be in .txt , .docx or .pdf only
filed=input()
```

```
CV.pdf
```

```
try:
    doc = Document()
    with open(filed, 'r') as file:
        doc.add_paragraph(file.read())
    doc.save("text.docx")
    data = ResumeParser('text.docx').get_extracted_data()
    print(data['skills'])
except:
    data = ResumeParser(filed).get_extracted_data()
    print(data['skills'])
```

```
['Computer science', 'Keras', 'Statistics', 'Opencv', 'Coding', 'Ui', 'Python', 'Training', 'Modeling', 'Sci', 'Analysis', 'Tes
ting', 'Github', 'Chemicals', 'Database', 'Prototype', 'Engineering', 'Api', 'Pyqt', 'Html', 'C', 'C++', 'Algorithms', 'Css',
'Design', 'Word', 'Programming']
```

# CHAPTER 40

## STEPS AND PROCESS:

### 1. Research and business understanding

The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something actionable.

### 2. Web Scraping

Web scraping is the process of using bots to extract content and data from a website. Unlike screen scraping, which only copies pixels displayed onscreen, web scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.

### 3. Data pre-processing

Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process.

### 4. Nearest Neighbours and cosine similarity

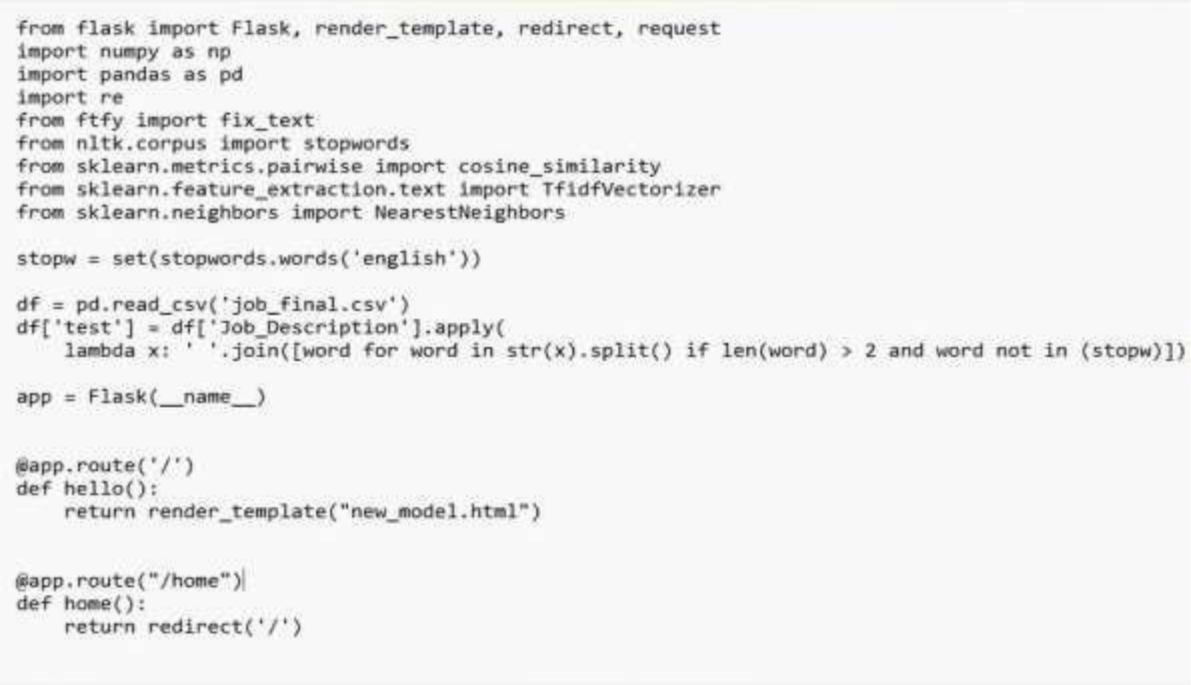
Nearest Neighbours Analysis measures the spread or distribution of something over a geographical space. It provides a numerical value that describes the extent to which a set of points are clustered or uniformly spaced.

### 5. Flask

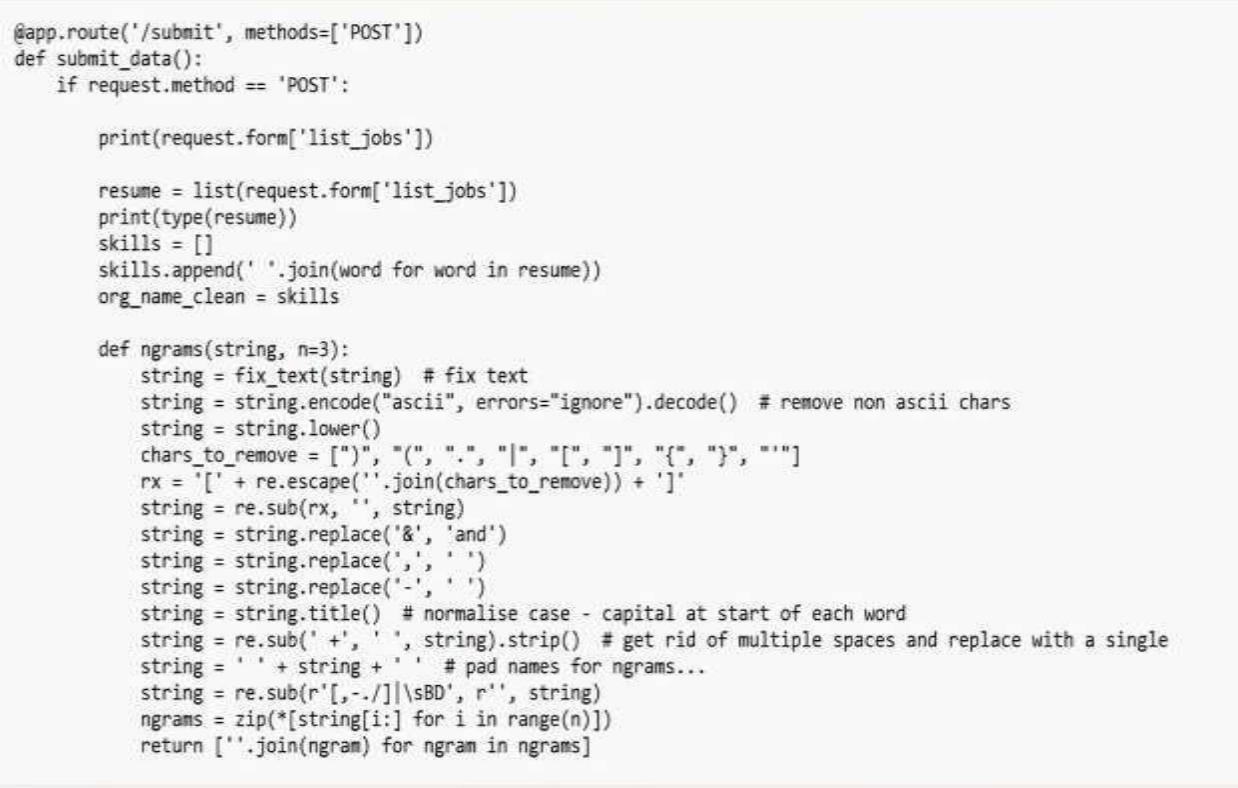
Flask is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application

# CHAPTER 7

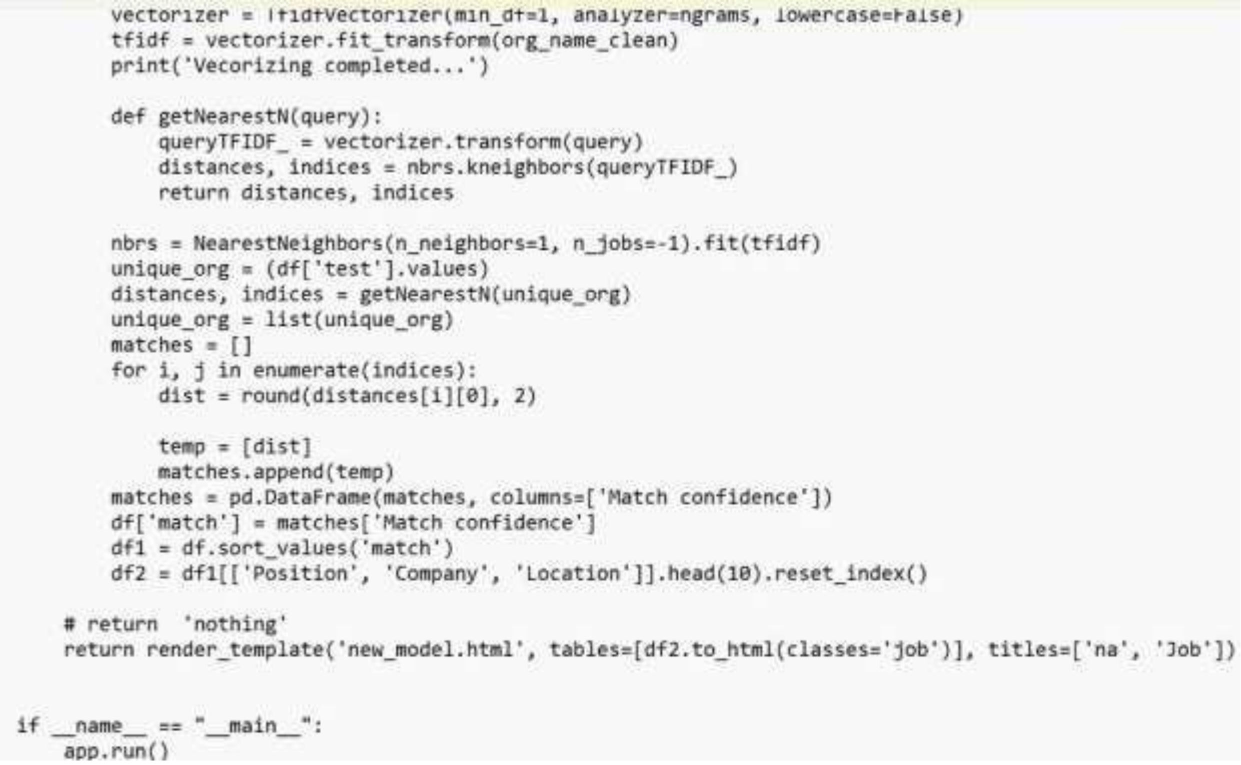
## 7.1 PROJECT CODE



```
from flask import Flask, render_template, redirect, request
import numpy as np
import pandas as pd
import re
from ftfy import fix_text
from nltk.corpus import stopwords
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.neighbors import NearestNeighbors

stopw = set(stopwords.words('english'))

df = pd.read_csv('job_final.csv')
df['test'] = df['Job_Description'].apply(
    lambda x: ' '.join([word for word in str(x).split() if len(word) > 2 and word not in (stopw)])

app = Flask(__name__)


@app.route('/')
def hello():
    return render_template("new_model.html")


@app.route("/home")
def home():
    return redirect('/')
```



```
@app.route('/submit', methods=['POST'])
def submit_data():
    if request.method == 'POST':

        print(request.form['list_jobs'])

        resume = list(request.form['list_jobs'])
        print(type(resume))
        skills = []
        skills.append(' '.join(word for word in resume))
        org_name_clean = skills

        def ngrams(string, n=3):
            string = fix_text(string)  # fix text
            string = string.encode("ascii", errors="ignore").decode()  # remove non ascii chars
            string = string.lower()
            chars_to_remove = [")", "(", ".", "|", "[", "]", "{", "}", "'"]
            rx = '[' + re.escape(''.join(chars_to_remove)) + ']'
            string = re.sub(rx, '', string)
            string = string.replace('&', 'and')
            string = string.replace(',', ' ')
            string = string.replace('-', ' ')
            string = string.title()  # normalise case - capital at start of each word
            string = re.sub(' +', ' ', string).strip()  # get rid of multiple spaces and replace with a single
            string = ' ' + string + ' '  # pad names for ngrams...
            string = re.sub(r'[,-./]|\sBD', r'', string)
            ngrams = zip(*[string[i:] for i in range(n)])
            return [''.join(ngram) for ngram in ngrams]
```

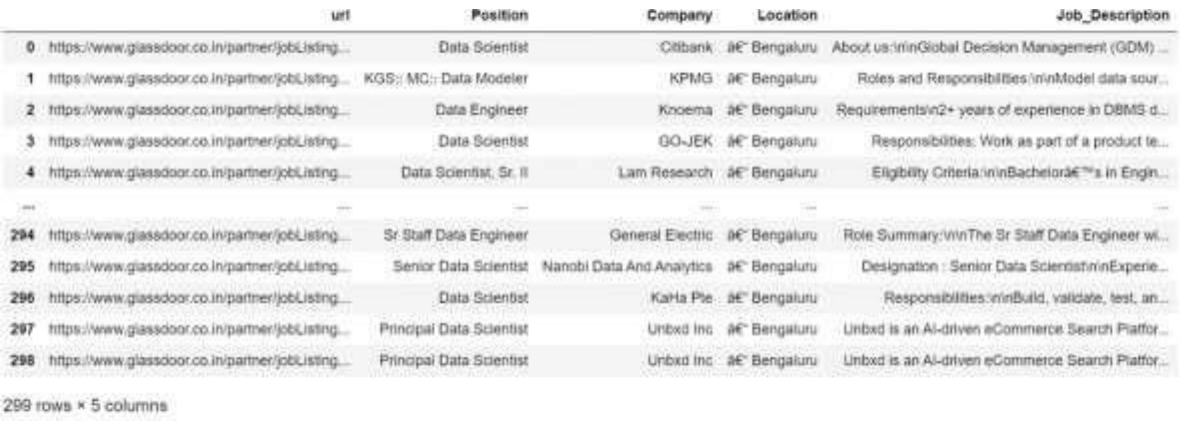


```
        vectorizer = TfidfVectorizer(min_df=1, analyzer=ngrams, lowercase=False)
        tfidf = vectorizer.fit_transform(org_name_clean)
        print('Vecorizing completed...')

        def getNearestN(query):
            queryTFIDF_ = vectorizer.transform(query)
            distances, indices = nbrs.kneighbors(queryTFIDF_)
            return distances, indices

        nbrs = NearestNeighbors(n_neighbors=1, n_jobs=-1).fit(tfidf)
        unique_org = (df['test'].values)
        distances, indices = getNearestN(unique_org)
        unique_org = list(unique_org)
        matches = []
        for i, j in enumerate(indices):
            dist = round(distances[i][0], 2)

            temp = [dist]
            matches.append(temp)
        matches = pd.DataFrame(matches, columns=['Match confidence'])
        df['match'] = matches['Match confidence']
        df1 = df.sort_values('match')
        df2 = df1[['Position', 'Company', 'Location']].head(10).reset_index()

    # return  'nothing'
    return render_template('new_model.html', tables=[df2.to_html(classes='job')], titles=['na', 'Job'])


if __name__ == "__main__":
    app.run()
```

# CHAPTER 8

# IMPLEMENTED RESULTS

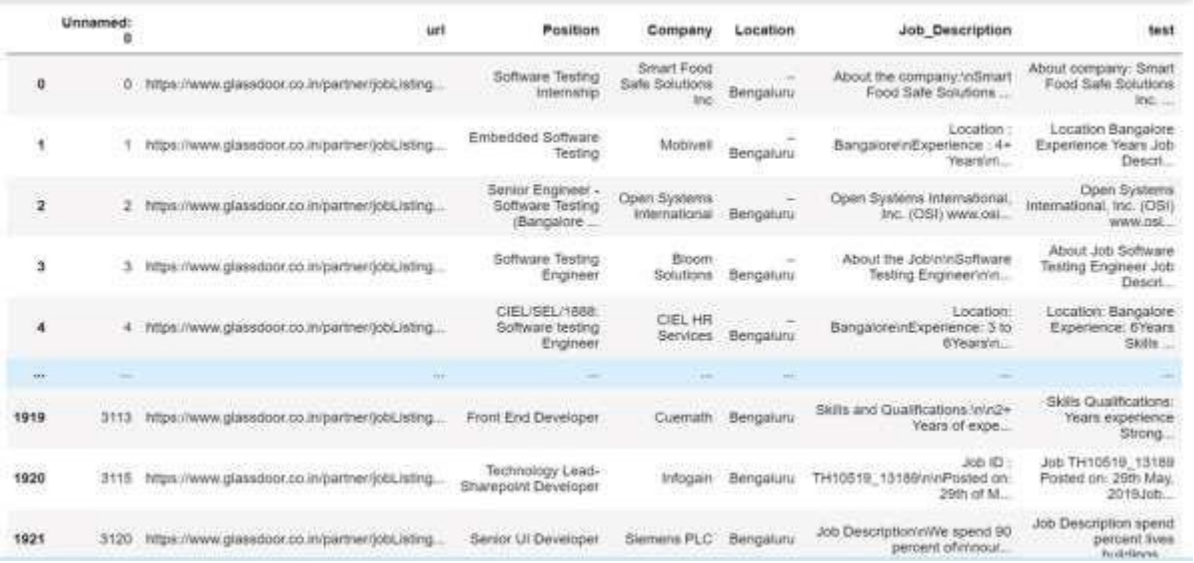
## 8.1 WEBSCRAPING



| | url | Position | Company | Location | Job_Description |
|---|---|---|---|---|---|
| 0 | https://www.glassdoor.co.in/partner/jobListing... | Data Scientist | Citibank | â€" Bengaluru | About us:\n\nGlobal Decision Management (GDM) ... |
| 1 | https://www.glassdoor.co.in/partner/jobListing... | KGS:: MC:: Data Modeler | KPMG | â€" Bengaluru | Roles and Responsibilities:\n\nModel data sour... |
| 2 | https://www.glassdoor.co.in/partner/jobListing... | Data Engineer | Knoema | â€" Bengaluru | Requirements\n2+ years of experience in DBMS d... |
| 3 | https://www.glassdoor.co.in/partner/jobListing... | Data Scientist | GO-JEK | â€" Bengaluru | Responsibilities: Work as part of a product te... |
| 4 | https://www.glassdoor.co.in/partner/jobListing... | Data Scientist, Sr. II | Lam Research | â€" Bengaluru | Eligibility Criteria:\n\nBachelorâ€™s in Engin... |
| ... | ... | ... | ... | ... | ... |
| 294 | https://www.glassdoor.co.in/partner/jobListing... | Sr Staff Data Engineer | General Electric | â€" Bengaluru | Role Summary:\n\nThe Sr Staff Data Engineer wi... |
| 295 | https://www.glassdoor.co.in/partner/jobListing... | Senior Data Scientist | Nanobi Data And Analytics | â€" Bengaluru | Designation : Senior Data Scientist\n\nExperie... |
| 296 | https://www.glassdoor.co.in/partner/jobListing... | Data Scientist | KaHa Pte | â€" Bengaluru | Responsibilities:\n\nBuild, validate, test, an... |
| 297 | https://www.glassdoor.co.in/partner/jobListing... | Principal Data Scientist | Unbxd Inc | â€" Bengaluru | Unbxd is an AI-driven eCommerce Search Platfor... |
| 298 | https://www.glassdoor.co.in/partner/jobListing... | Principal Data Scientist | Unbxd Inc | â€" Bengaluru | Unbxd is an AI-driven eCommerce Search Platfor... |

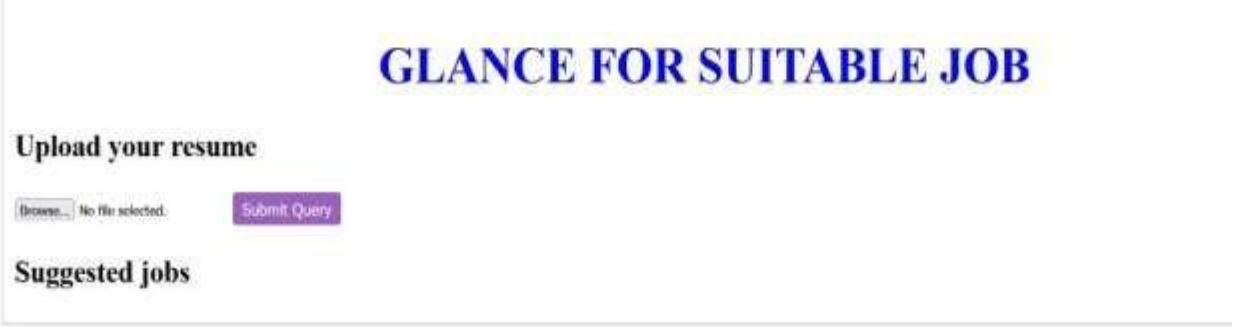
299 rows × 5 columns

## 8.2 DATA CLEANING



| | Unnamed: 0 | url | Position | Company | Location | Job_Description | test |
|---|---|---|---|---|---|---|---|
| 0 | 0 | https://www.glassdoor.co.in/partner/jobListing... | Software Testing Internship | Smart Food Safe Solutions Inc | – Bengaluru | About the company:\nSmart Food Safe Solutions ... | About company: Smart Food Safe Solutions Inc. ... |
| 1 | 1 | https://www.glassdoor.co.in/partner/jobListing... | Embedded Software Testing | Mobivell | – Bengaluru | Location : Bangalore\nExperience : 4+ Years\n... | Location Bangalore Experience Years Job Descri... |
| 2 | 2 | https://www.glassdoor.co.in/partner/jobListing... | Senior Engineer - Software Testing (Bangalore ... | Open Systems International | – Bengaluru | Open Systems International, Inc. (OSI) www.osi... | Open Systems International, Inc. (OSI) www.osi... |
| 3 | 3 | https://www.glassdoor.co.in/partner/jobListing... | Software Testing Engineer | Bloom Solutions | – Bengaluru | About the Job\n\nSoftware Testing Engineer\n\n... | About Job Software Testing Engineer Job Descri... |
| 4 | 4 | https://www.glassdoor.co.in/partner/jobListing... | CIEL/SEL/1888: Software testing Engineer | CIEL HR Services | – Bengaluru | Location: Bangalore\nExperience: 3 to 6Years\n... | Location: Bangalore Experience: 6Years Skills ... |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 1919 | 3113 | https://www.glassdoor.co.in/partner/jobListing... | Front End Developer | Cuemath | Bengaluru | Skills and Qualifications:\n\n2+ Years of expe... | Skills Qualifications: Years experience Strong... |
| 1920 | 3115 | https://www.glassdoor.co.in/partner/jobListing... | Technology Lead-Sharepoint Developer | Infogain | Bengaluru | Job ID : TH10519_13189\n\nPosted on: 29th of M... | Job TH10519_13189 Posted on: 29th May, 2019Job... |
| 1921 | 3120 | https://www.glassdoor.co.in/partner/jobListing... | Senior UI Developer | Siemens PLC | Bengaluru | Job Description\nWe spend 90 percent of\nour... | Job Description spend percent lives buildings ... |

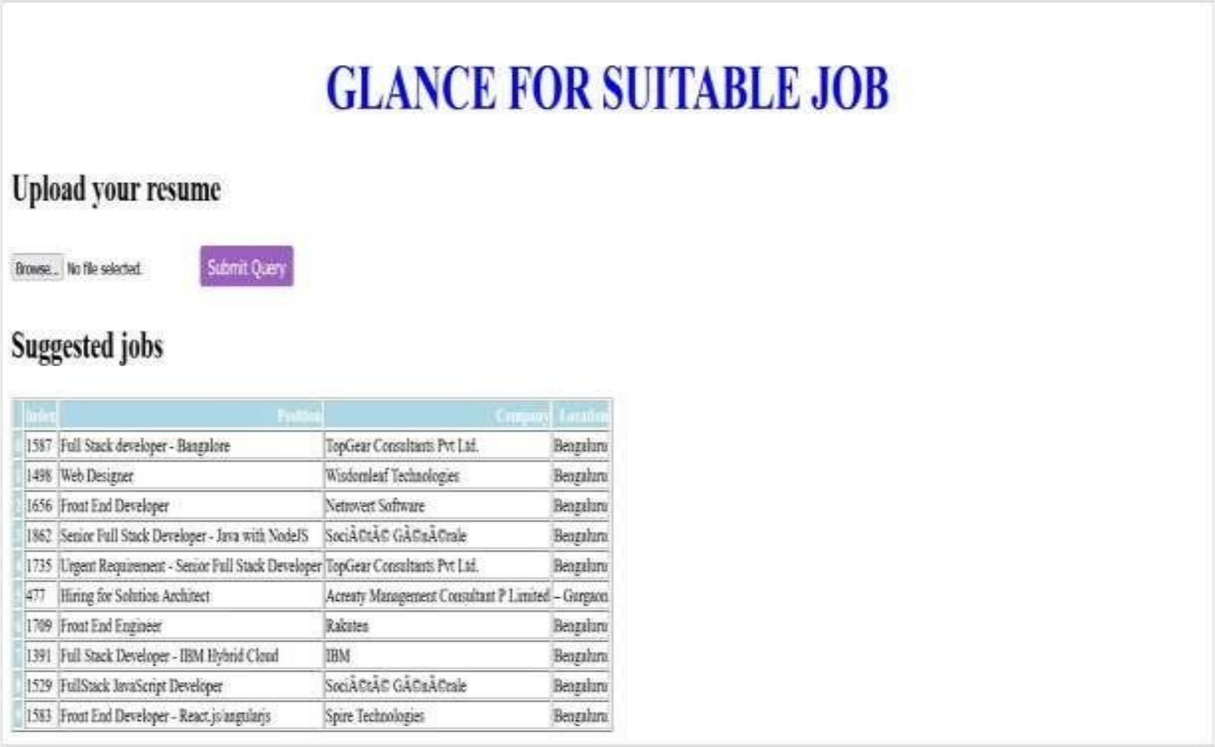
## 8.3 VERSION 1



GLANCE FOR SUITABLE JOB

Upload your resume

Suggested jobs



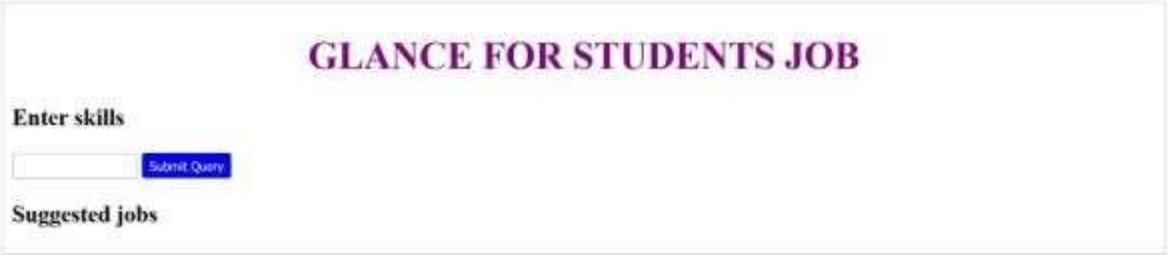
GLANCE FOR SUITABLE JOB

Upload your resume

Suggested jobs

| | Index | Position | Company | Location |
|---|---|---|---|---|
| 0 | 1587 | Full Stack developer - Bangalore | TopGear Consultants Pvt Ltd. | Bengaluru |
| 1 | 1498 | Web Designer | Wisdomleaf Technologies | Bengaluru |
| 2 | 1656 | Front End Developer | Netrovert Software | Bengaluru |
| 3 | 1862 | Senior Full Stack Developer - Java with NodeJS | SociÃ©tÃ© GÃ©nÃ©rale | Bengaluru |
| 4 | 1735 | Urgent Requirement - Senior Full Stack Developer | TopGear Consultants Pvt Ltd. | Bengaluru |
| 5 | 477 | Hiring for Solution Architect | Acreaty Management Consultant P Limited | – Gurgaon |
| 6 | 1709 | Front End Engineer | Rakuten | Bengaluru |
| 7 | 1391 | Full Stack Developer - IBM Hybrid Cloud | IBM | Bengaluru |
| 8 | 1529 | FullStack JavaScript Developer | SociÃ©tÃ© GÃ©nÃ©rale | Bengaluru |
| 9 | 1583 | Front End Developer - React.js/angularjs | Spire Technologies | Bengaluru |

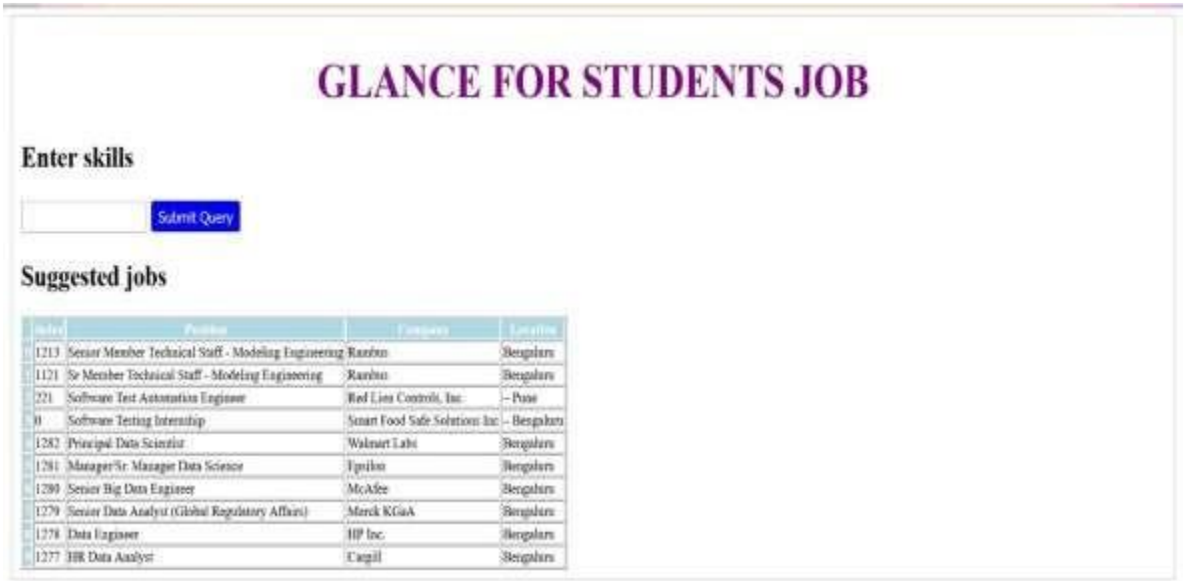
## 8.4 VERSION 2



GLANCE FOR STUDENTS JOB

Enter skills

Suggested jobs



GLANCE FOR STUDENTS JOB

Enter skills

Suggested jobs

| | Index | Position | Company | Location |
|---|---|---|---|---|
| | 1213 | Senior Member Technical Staff - Modeling Engineering | Rambus | Bengaluru |
| | 1121 | Sr Member Technical Staff - Modeling Engineering | Rambus | Bengaluru |
| | 221 | Software Test Automation Engineer | Red Lion Controls, Inc | – Pune |
| | 0 | Software Testing Internship | Smart Food Safe Solutions Inc | – Bengaluru |
| | 1282 | Principal Data Scientist | Walmart Labs | Bengaluru |
| | 1281 | Manager/Sr. Manager Data Science | Epsilon | Bengaluru |
| | 1280 | Senior Big Data Engineer | McAfee | Bengaluru |
| | 1279 | Senior Data Analyst (Global Regulatory Affairs) | Merck KGaA | Bengaluru |
| | 1278 | Data Engineer | HP Inc. | Bengaluru |
| | 1277 | HR Data Analyst | Cargill | Bengaluru |

# CONCLUSION

The main purpose of this dissertation was to analyse and predict job prospect features in the capital city of Kerala, Trivandrum using machine learning techniques and to link theminto a web application. The web application created for this study using angular platform provides an interactive interface for the job aspirants based on azure machine learning neural models. After training and testing five algorithms, an interactive session based on the prediction models from each algorithm was incorporated in the web application of the thesis. Multilayer perceptron neural network was observed to be the best algorithm meeting the requirements of this study.

This Application can be a valuable tool for connecting real people with real jobs, in real time, especially for Job Seekers , who are more likely to search for jobs on the Internet.

The Application allow job seekers to customize their job search according to their choice and preference. It improves Job seekers visibility on the landscape of the recruitment spectrum. It also increases the chances of getting opportunities from number of companies. This application will be a user friendly for the students which make easier forthe students in search of jobs.

The Application allow job seekers to customize their job search according to their choice and preference. It improves Job seekers visibility on the landscape of the recruitment spectrum. It also increases the chances of getting opportunities from number of companies. This application will be a user friendly for the students which make easier forthe students in search of jobs.

# REFERENCE

- ❖ Google
- ❖ Wikipedia
- ❖ Chrome
- ❖ k-Nearest_neighbour_classifiers
- ❖ msc_saju_a_g_2018.pdf
- ❖ https://www.rchilli.com/blog/resume-parsing-101/
- ❖ https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/#:~:text=Web%20scraping%20is%20an%20automatic,be%20used%20in%20various%20applications
- ❖ https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- ❖ https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
- ❖ https://www.dataquest.io/blog/regular-expressions-data-scientists/