



INTRODUCTION

1 . Introduction :-

1.1 Introduction to System

Majority of world's unstructured data is in the textual form. To make sense of it, one must, either go through it painstakingly or employ certain automated techniques to extract relevant information. Looking at the volume, variety and velocity of such textual data, it is imperative to employ Text Mining techniques to extract the relevant information, transforming unstructured data into structured form, so that further insights, processing, analysis, visualizations are possible.

This article deals with a specific domain, of applicant profiles or resumes. They, as we know, different file formats (txt, doc, pdf, etc.) but also with different contents and layouts. Such heterogeneity makes extraction of relevant information, a challenging task. Even though it may not be possible to fully extract all the relevant information from all the types of formats, one can get started with simple steps and at least extract whatever is possible from some of the known formats.

Broadly there are two approaches: linguistics based and Machine Learning based. In “linguistic” based approaches pattern searches are made to find key information, whereas in “machine learning” approaches supervised-unsupervised methods are used to extract the information. “Regular expression” (RegEx), used here, is one of the “linguistic” based pattern-matching method

1.2 ABSTRACT

Using Natural Language Processing(NLP) and (ML)Machine Learning to rank the resumes according to the given constraint, this intelligent system ranks the resume of any format according to the given constraints or the following requirements provided by the client company. We will basically take the bulk of input resume from the client company and that client company will also provided the requirement and the constraints according to which the resume shall be ranked by our system. Moreover the details acquired from the resumes, our system shall be reading the candidates social profiles (like LinkedIn, Github etc) which will the more genuine information about that candidate.



SYSTEM STUDY

2 . Business Problem : -

Making a good hiring decision not only increases employee retention but also reduces the cost associated with a bad hire. There is an ever-increasing focus on effective recruitment.

An organization invests a lot of its time and resources in search of the potential candidates and **what if the selected candidates do not join the organization** in the end. The recruiters **need to identify** the **chances** of the **potential candidates** of **joining** the organization **even before short listing** their resumes for the interview process

3 . System Study : -

3.1 Project Objective:

To save the company from short listing the candidates who might not join the company after being selected, this research study addresses the following main objectives:

- i. To determine and analyze the variables that can make a candidate back out from the job he/she is being selected for.
- ii. By analyzing the data, find out the chances or rate of the candidates backing out after the selection and make effective predictions for the future recruitment process.

Solutions:-

To deal with above business problem and project objective, I have used logical thought process for finalizing the variables which can cause for backing out from joining the organization by the selected or shortlisted candidate.

3.2 Proposed System

Develop a system which will eliminate the difficulties arising in existing system. When the resume is uploaded, it will be parsed and then transforms into a standardized format. The candidate can upload his resume in multiple formats like .doc, .pdf etc. A user interface screen will be given where the authority concerned will store the job profile for the particular post. The system will go through all the resumes and suggest the value pairs are then stored into the database and converted into a word document which can be viewed by candidate as well as the recruiters. The system keeps the copies of both original and the standardized resume.

3.3 Need for Computerization

Corporate companies and recruitment agencies process numerous resumes daily. This is no task for humans. An automated intelligent system is required which can take out all the vital information from the unstructured resumes and transform all of them to a common structured format which can then be ranked for a specific job position. Parsed information include name, email address, social profiles, personal websites, years of work experience, work experiences, years of education, education experiences, publications, certifications, volunteer experiences, keywords and finally the cluster of the resume (ex: computer science, human resource, etc.). The parsed information is then stored in a database (NoSQL in this case) for later use. Unlike other unstructured data (ex: email body, web page contents, etc.), resumes are a bit structured. Information is stored in discrete sets. Each set contains data about the person's contact, work experience or education details. In spite of this resumes are difficult to parse. This is because they vary in types of information, their order, writing style, etc. Moreover, they can be written in various formats. Some of the common ones include '.txt', '.pdf', '.doc', '.docx', '.odt', '.rtf' etc. To parse the data from different kinds of resumes effectively and efficiently, the model must not rely on the order or type of data.

3.4 Feasibility Study

Structure of Resume -

As you build your machine learning engineer resume, you should generally include the following information. The fields are ranked by importance, but the order can be tweaked depending on your personal experience and preferences.

1. **Header:** This is where you'll state your name, address, and contact information.
2. **Personal summary:** Typically around three to five sentences, a personal summary is a quick synopsis of who you are and what you've achieved. Using compelling storytelling, the summary is your opportunity to reel in the hiring manager. Show your personality and brag about your major successes. This is your opportunity to explain why you're the perfect fit for the company.
3. **Experience:** In this section, you'll include all past relevant jobs you've had, along with a brief summary of your duties and accomplishments. Incorporate the time period for which you were employed. If you have less experience, feel free to include one or two internships that are pertinent to the job you're applying for.
4. **Projects (optional):** This field highlights and briefly summarizes applicable projects you've completed and that the prospective employer may find of interest.

5. **Education/certifications:** Include degrees you've earned and other relevant professional certifications and courses you've completed. If you're earlier on in your career (and have space), feel free to add your grade point average as long as it's higher than a 3.0. (This is where you would include your completion of Springboard's AI / Machine Learning Career Track.)
6. **Skills** (optional): This segment is where you'll weave in skills and expertise that you weren't able to incorporate in other sections of your resume. For example, in a machine learning engineer resume,
7. **References** (optional): Include the contact information of two or three people who know your work ethic and would be able to speak on your behalf. It's critically important to ask these individuals for permission to list them as references on your resume before you send it out to prospective employers. It's good etiquette, and if your references are caught off guard, it reflects poorly on you.



TOOLS & ENVIRONMENT

4. Tools Used :-

4.1. Anaconda Enterprise -

Anaconda Enterprise is an enterprise-ready, secure and scalable data science platform that empowers teams to govern data science assets, collaborate and deploy data science projects.

Enterprise 5 includes these capabilities:

- Easily deploy your projects into interactive data applications, live notebooks and machine learning models with APIs.
- Share those applications with colleagues and collaborators.
- Manage your data science assets: notebooks, packages, environments and projects in an integrated data science experience.

4.2. Anaconda Distribution-

Anaconda Distribution is a free, easy-to-install package manager, environment manager and Python distribution with a collection of 1,000+ open source packages with free community support. Anaconda is platform-agnostic, so you can use it whether you are on Windows, macOS or Linux.

4.3. Anaconda Cloud-

Anaconda Cloud is a package management service that makes it easy to find, access, store and share public notebooks, and environments, as well as conda and PyPI packages.

Python :-

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site, <https://www.python.org/>, and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation.

The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications.

This tutorial introduces the reader informally to the basic concepts and features of the Python language and system. It helps to have a Python interpreter handy for hands-on experience, but all examples are self-contained, so the tutorial can be read off-line as well.

For a description of standard objects and modules, see [The Python Standard Library](#). [The Python Language Reference](#) gives a more formal definition of the language. To write extensions in C or C++, read [Extending and Embedding the Python Interpreter](#) and [Python/C API Reference Manual](#). There are also several books covering Python in depth.

This tutorial does not attempt to be comprehensive and cover every single feature, or even every commonly used feature. Instead, it introduces many of Python's most noteworthy features, and will give you a good idea of the language's flavor and style. After reading it, you will be able to read and write Python modules and programs, and you will be ready to learn more about the various Python library modules described in [The Python Standard Library](#).

NLP :-

Natural language processing (NLP) is getting very popular today, which became especially noticeable in the background of the deep learning development. NLP is a field of artificial intelligence aimed at understanding and extracting important information from text and further training based on text data. The main tasks include speech recognition and generation, text analysis, sentiment analysis, machine translation, etc.

In the past decades, only experts with appropriate philological education could be engaged in the natural language processing. Besides mathematics and machine learning, they should have been familiar with some key linguistic concepts. Now, we can just use already written NLP libraries. Their main purpose is to simplify the text preprocessing. We can focus on building machine learning models and hyper parameters fine-tuning.

There are many tools and libraries designed to solve NLP problems. Today, we want to outline and compare the most popular and helpful natural language processing libraries, based on our experience. You should understand that all the libraries we look at have only partially overlapped tasks. So, sometimes it is hard to compare them directly. We will walk around some features and compare only those libraries, for which this is possible.

General overview of Packages -

NLTK (Natural Language Toolkit) is used for such tasks as tokenization, lemmatization, stemming, parsing, POS tagging, etc. This library has tools for almost all NLP tasks.

Spacy is the main competitor of the NLTK. These two libraries can be used for the same tasks.

Scikit-learn provides a large library for machine learning. The tools for text preprocessing are also presented here.

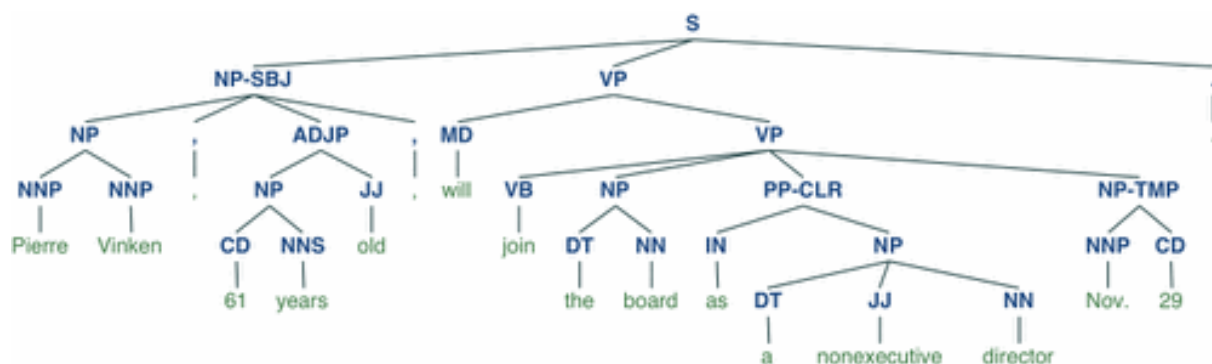
Natural Language Toolkit : -

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been updated for Python 3 and NLTK 3



5. System Requirement :-

System requirements

1. Hardware Requirements:-

- System architecture: Windows- 64-bit x86, 32-bit x86; MacOS- 64-bit x86; Linux- 64-bit x86, 64-bit Power8/Power9.
- Minimum 5 GB disk space to download and install.
- On Windows, macOS, and Linux, it is best to install Anaconda for the local user, which does not require administrator permissions and is the most robust type of installation. However, if you need to, you can install Anaconda system wide, which does require administrator permissions.

1. Software Requirements:-

- License: Free use and redistribution under the terms of the [End User License Agreement](#).
- Operating system: Windows 7 or newer, 64-bit macOS 10.10+, or Linux, including Ubuntu, RedHat, CentOS 6+, and others.
- If your operating system is older than what is currently supported, you can find older versions of the Anaconda installers in our [archive](#) that might work for you. Check our [FAQ](#) for version recommendations.



FLOWCHART

6. Flow Chart :-

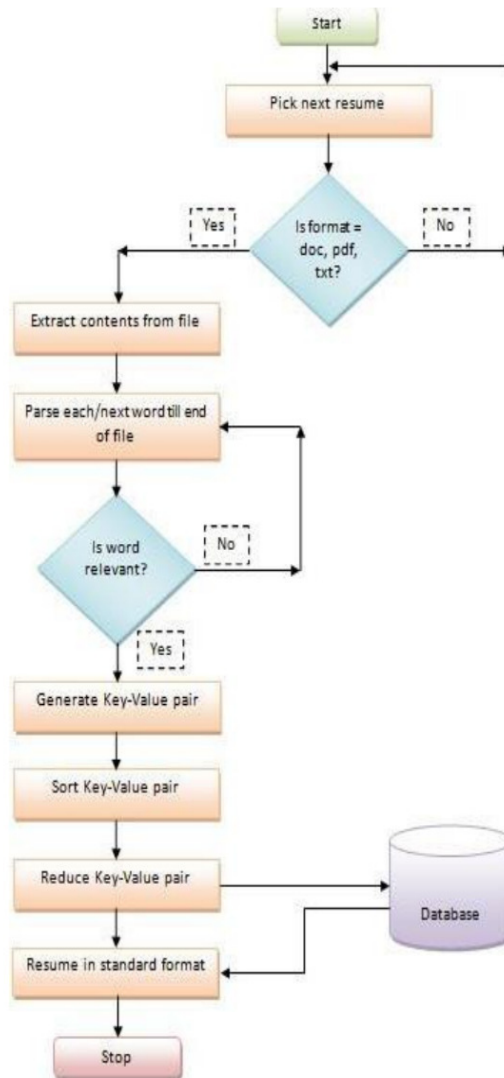


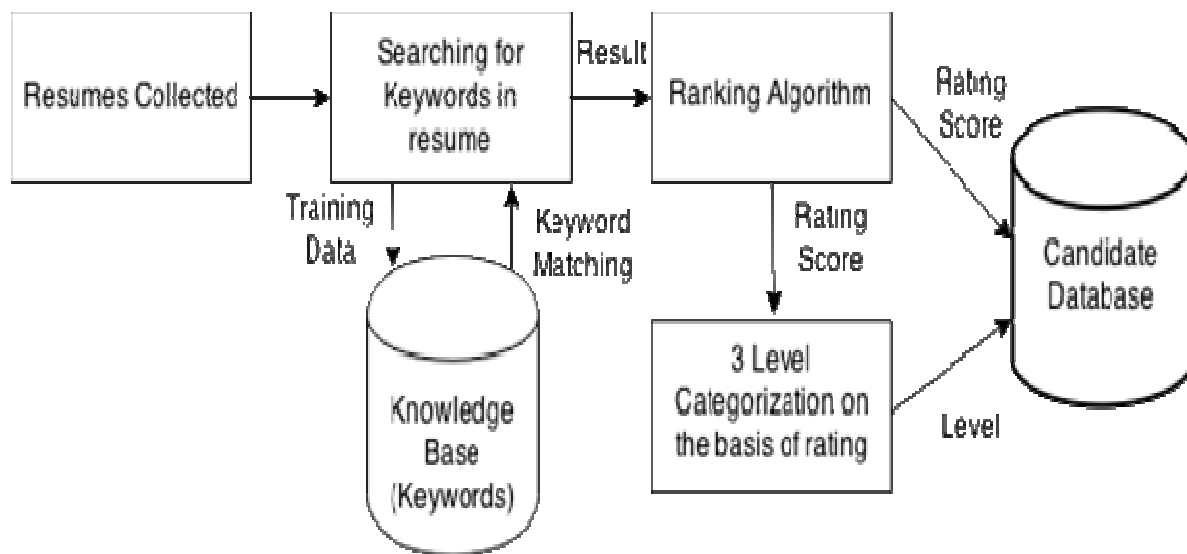
Fig.12. Flowchart of proposed resume system



SYSTEM DESIGN

7 System Design :-

To design this model, various other models and job search theories were analyzed. As a result, the whole system was easily segmented and designed properly to meet the need of both the employers and job seekers. However, the system will be more efficient with the amount of data the system gets.



Block diagram of the whole system

To build the parser it is really important to have sufficient amount of data so that the results are as accurate as possible. However, this sufficient amount should always be as much as possible so that the whole system can be trained properly. Fig. 3.1.1 describes the proposed model with a block diagram.

RESUME COLLECTION -

This step involves the collection of various resumes uploaded by the candidates. A simple web interface has been designed in our prototype model which will make the user fill a form having the fields which would be required to be filled by the job seeker. Our prototype deals with candidates for IT companies but

this can be generalized for various other sectors by using an even more extensive knowledge base.

KEYWORD SEARCHING –

This is one of the most crucial steps of our model. A knowledge base consisting of various keywords is made from the initial training data. The input text which is received needs some pre-processing before it can be used. For this purpose, we use a POS tagger and a chunker which are used to split the text into sentences, which are then analyzed by a syntactic parser which labels all the words with their part of speech information. Using a chunker helps in providing a flat structure of extracted data . Lexical Analysis can also be done to tokenize the words which can be then categorized for the purpose of parsing.

ADDITION IN KNOWLEDGE BASE –

While the keywords found in resume text will be matched, the words which are not found in knowledge base are further analyzed and if found relevant, is added to the knowledge base.

Since the data from which knowledge extraction has to be done is unstructured, we follow traditional methods of information extraction. Apart from that, Ontology based Information extraction can also be done by Semantic Annotation in which we augment the natural language text into metadata which can be represented in form of RDFa (Resource Description Framework in attributes) . The process is divided into two subtasks – Terminology Extraction and Entity Linking.

RANKING AND CATEGORIZATION -

After getting the rating score of the resume, a candidate can be ranked on the basis of his resume's score. This will be useful in comparing two candidates while short listing them. Whenever the company searches for a candidate keeping in mind certain requirements, the candidate who is ranked above will be presented to the company first which would be adding to his advantage in cases where the vacancies available may not be high.

7.1. Historical Dataset Creation for Modeling -

Finalization of variables is done in stepwise approach like:

1. First I used whatever data available on net and combined it
2. Some researched documentation available on net by SMEs
3. Then with logical thinking finalized the variables

Once my variables were ready, I used whatever data was available for the variables and for the variables added by me with my knowledge in domain, I used logical data simulation.

Now, once my data is ready, I assume it as historical data and these will be my input variables for my prediction model.

As we know our business problem is predicting the joining possibility of the candidate, the output what we expect for it is binary categorical i.e. Join or Not Join, So our output variable will be **“Join or Not Join”**, created this variable in our dataset and labelled it with the values to create the final Dataset (Assumed Historical Dataset) to be used in our prediction model building.

```
In [25]: Data_frame.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 46 entries, 0 to 45
Data columns (total 8 columns):
Education          46 non-null object
Email              46 non-null object
Last Modified      46 non-null object
Mob no             46 non-null object
Name               46 non-null object
Number of exp      46 non-null object
Resume             46 non-null object
Skills Set         46 non-null object
dtypes: object(8)
memory usage: 3.0+ KB
```

```
In [26]: Data_frame.dtypes
Out[26]:
Education          object
Email              object
Last Modified      object
Mob no             object
Name               object
Number of exp      object
Resume             object
Skills Set         object
dtype: object
```

Steps 1 : Data Collection

Data collection in multiple formats like .doc, .pdf

Step 2 : EDA Process (Text Cleaning or Preprocessing)

- **Remove Punctuations, Numbers:** Punctuations, Numbers doesn't help much in processing the given text, if included, they will just increase the size of bag of words that we will create as last step and decrease the efficiency of algorithm.
- **Stemming:** Take roots of the word
- **Convert each word into its lower case:** For example, it is useless to have same words in different cases (eg 'good' and 'GOOD').

```
Python console
Console 1/A

In [18]: fullText = docx2txt.process(path+file)

In [19]: fullText
Out[19]: 'Resume:\n\nYogesh chandan\n\n8989448856\n\nYogeshchandan12@gmail.com\n\nSenior Project/Program Manager, \n\nMS, MBA, PMP®, IIM Certification\n\nManager - AMD\n\nBengaluru, Karnataka.\n\nAspiring to manage people, business delivery & budgets with bigger teams where I can best use my skills and experience to bring in quantum improvements in projects and organizations. Have an enriching career where my skills are put to best use and I get to grow both professionally and personally. Looking to place myself at assignments in long term. Work Experience Senior Project/Program Manager Dell Technologies\n\nMay 2012 to Present\n\nResponsibilities Strategic Project Program Manager with 6+ years of experience in the Information Technology\n\nRich experience in management, delivery & growth of IT programs for key global accounts in multiple Application Development, Knowledge Management System, Training & Development and Project Delivery Management\n\nLeading extensive business & IT level interactions with C suite executives mapping customer business value chain & process and formulating strategic solutions while ensuring delivery excellence\n\nGained insightful experience in managing multiple software development projects across various geographies while meeting business needs in fast paced, dynamic and globally expanding MNC environment\n\nStrategist & implementer with proficiency in spearheading business to accomplish corporate plans and goals\n\nDrafting proposals responding to RFPs preparing and eliciting requirements\n\nExpertise in handling various software development methodologies and frameworks (DevOps, Agile, Waterfall, Kanban, Lean, Scrum)\n\nExpertise in managing offshore and onsite teams\n\nManage projects from Define to Sustain phases of Offer lifecycle process\n\nSkills\n\nAgile Methodologies (5 years), SDLC (10+ years), Software Project Management (6 years), Lean Six Sigma (6 years), ISO 9001 (Less than 1 year), Configuration Management (8 years), Project Finance (6 years), Business Analysis (6 years), Operations Management (7 years), Project Portfolio Management (6 years), Total Quality Management (TQM) (6 years), Leadership (8 years), Strategic Negotiations (8 years), Agile PLM (Less than 1 year), Docker (Less than 1 year), C#.net (Less than 1 year), Install Anywhere (Less than 1 year), Maven (Less than 1 year), Installshield (Less than 1 year), Jenkins (Less than 1 year), VC++ (Less than 1 year), DevOps (Less than 1 year), Scrum (Less than 1 year)\n\nLinks\n\nhttps://www.linkedin.com/in/yogeshchandan12/\n\nSoundar'
```

```
In [20]: fullText = re.sub("[^A-Za-z0-9.@_'\"]+", " ",fullText)

In [21]: fullText
Out[21]: 'Resume Yogesh chandan 8989448856 Yogeshchandan12@gmail.com Senior Project Program Manager MS MBA PMP IIM Certification Senior Technical Project Manager - AMD Bengaluru, Karnataka. Aspiring to manage people business delivery budgets with bigger teams where I can best use my skills and experience to bring in quantum improvements in projects and organizations. Have an enriching career where my skills are put to best use and I get to grow both professionally and personally. Looking to place myself at assignments in long term. Work Experience Senior Project Program Manager Dell Technologies May 2012 to Present Responsibilities Strategic Project Program Manager with 6+ years of experience in the Information Technology Rich experience in management delivery growth of IT programs for key global accounts in multiple Application Development Knowledge Management System Training Development and Project Delivery Management Leading extensive business IT level interactions with C suite executives mapping customer business value chain and process and formulating strategic solutions while ensuring delivery excellence Gained insightful experience in managing multiple software development projects across various geographies while meeting business needs in fast paced dynamic and globally expanding MNC environment Strategist implementer with proficiency in spearheading business to accomplish corporate plans and goals Drafting proposals responding to RFPs preparing and eliciting requirements Expertise in handling various software development methodologies and frameworks (DevOps, Agile, Waterfall, Kanban, Lean, Scrum) Expertise in managing offshore and onsite teams Manage projects from Define to Sustain phases of Offer lifecycle process Skills Agile Methodologies 5 years SDLC 10 years Software Project Management 6 years Lean Six Sigma 6 years ISO 9001 5 years Software Development 8 years Configuration Management 8 years Project Finance 6 years Business Analysis 6 years Operations Management 7 years Project Management Office PMO 6 years Project Portfolio Management 6 years Total Quality Management 6 years Leadership 8 years Strategic Negotiations 8 years Agile PLM 6 years Microsoft Office Less than 1 year Docker Less than 1 year C#.net Less than 1 year Installshield Less than 1 year Microsoft Project Less than 1 year Jenkins Less than 1 year VC++ Less than 1 year DevOps Less than 1 year Scrum Less than 1 year'
```

```

In [15]: Data_frame.columns
Out[15]:
Index(['Education', 'Email', 'Last Modified', 'Mob no', 'Name',
      'Number of exp', 'Resume', 'Skills Set'],
      dtype='object')

In [16]: Data_frame.head()
Out[16]:
   Education ... Skills Set
0 [bca, engineering] ... []
1 [engineering] ... []
2 [] ... [sql]
3 [] ... [sql]
4 [engineering] ... [sql, unix shell scripting, data warehouse, etl]

[5 rows x 8 columns]

In [17]: Data_frame.describe().head()
Out[17]:
   Education Email ... Resume Skills Set
count      46  46 ...      46      46
unique      13  31 ...      46      17
top          [] ...  harshit_Delhi_4.08_yrs.docx  []
freq        16  15 ...           1      26

[4 rows x 8 columns]

```

Steps 3 : Read Files of Resumes

output - List (46 elements)

Index	Type	Size	Value
0	dict	8	{'Name': 'Abhay Patil', 'Mob no': '', 'Email': '', 'Resume': 'Abhay Patil ...
1	dict	8	{'Name': 'HAQUE B. TECH', 'Mob no': '8688562564', 'Email': 'arshad5b2@gmail ...
2	dict	8	{'Name': 'Akmal Husan', 'Mob no': '8606845834', 'Email': 'AkmalE86@gmail ...
3	dict	8	{'Name': 'Akshay Balaso', 'Mob no': '7676857388', 'Email': 'Akshay123@gm ...
4	dict	8	{'Name': 'Amrut Gore', 'Mob no': '7899448856', 'Email': 'Rohan23hg@gmail ...
5	dict	8	{'Name': 'Ayushi Varma', 'Mob no': '9876543210', 'Email': 'varma@gmail.co ...
6	dict	8	{'Name': 'CURRICULUM VITAE', 'Mob no': '9673888932', 'Email': 'kiranshind ...
7	dict	8	{'Name': 'Gayatri Saraswat', 'Mob no': '', 'Email': '', 'Resume': 'Gayatri ...
8	dict	8	{'Name': 'Gurwinder Singh', 'Mob no': '', 'Email': '', 'Resume': 'Gurwinde ...
9	dict	8	{'Name': 'HARSHIT SAXENA', 'Mob no': '', 'Email': 'harry.saxena@yahoo.com ...
10	dict	8	{'Name': 'Kirti Chavan', 'Mob no': '', 'Email': '', 'Resume': 'Kirti Chava ...
11	dict	8	{'Name': 'Komal Desai', 'Mob no': '', 'Email': 'komalidesai7788@gmail.com' ...
12	dict	8	{'Name': 'Kuldeep Kour', 'Mob no': '', 'Email': '', 'Resume': 'Kuldeep Kou ...
13	dict	8	{'Name': 'Kuldeep Rankumar', 'Mob no': '8087408589', 'Email': 'sharmakuld ...
14	dict	8	{'Name': 'Madhu Warma', 'Mob no': '', 'Email': '', 'Resume': 'Madhu Warma. ...
15	dict	8	{'Name': 'Mahendra Bahubali', 'Mob no': '', 'Email': '', 'Resume': 'Mahend ...
16	dict	8	{'Name': 'Mahesh Ahmed', 'Mob no': '', 'Email': '', 'Resume': 'Mahesh Ahme ...
17	dict	8	{'Name': 'Manish patil', 'Mob no': '9999448844', 'Email': 'manish@gmail.c ...
18	dict	8	{'Name': 'Manisha Kour', 'Mob no': '', 'Email': '', 'Resume': 'Manisha Kou ...
19	dict	8	{'Name': 'Murali U', 'Mob no': '8999448856', 'Email': 'MuraliJadhav23@gma ...
20	dict	8	{'Name': 'Jadhav Murlil', 'Mob no': '8412960018', 'Email': 'murliljadhav007 ...
21	dict	8	{'Name': 'Nitin Gore', 'Mob no': '4545448856', 'Email': 'Nitin34gore@gmail ...
22	dict	8	{'Name': 'Poonam Mishra', 'Mob no': '', 'Email': '', 'Resume': 'Poonam Mis ...
23	dict	8	{'Name': 'Pournima Vanarase', 'Mob no': '9184089224', 'Email': 'pournimav ...
24	dict	8	{'Name': 'Rahul Mahajan', 'Mob no': '', 'Email': '', 'Resume': 'Rahul Maha ...

2 - Dictionary (8 elements)


Key	Type	Size	Value
Education	list	0	[]
Email	str	1	AkmalE86@gmail.com
Last Modified	str	1	Wed May 1 15:26:14 2019
Mob no	str	1	8606845834
Name	str	1	Akmal Husan
Number of exp	str	1	1.3 years
Resume	str	1	Akmal Eache.docx
Skills Set	list	1	['sql']

Save and Close Close

Remove Stopwords:-

20

Nouns: -

 nouns - List (6090 elements)

Index	Type	Size	
0	str	1	Yogesh
1	str	1	Yogeshchandan12
2	str	1	Senior
3	str	1	Project
4	str	1	Program
5	str	1	Manager
6	str	1	MBA
7	str	1	PMP
8	str	1	IIM
9	str	1	Certification
10	str	1	Senior
11	str	1	Technical
12	str	1	Program
13	str	1	Manager
14	str	1	AMD
15	str	1	Bengaluru
16	str	1	Karnataka
17	str	1	Aspiring
18	str	1	Executive
19	str	1	Business
20	str	1	Experience
21	str	1	Senior

Step 5: Extracted skill and Education To using Matching

skill_set - List (23 elements)

Index	Type	Size	
0	str	1	flask
1	str	1	django
2	str	1	mysql
3	str	1	c
4	str	1	cpp
5	str	1	html
6	str	1	js
7	str	1	machine learning
8	str	1	c++
9	str	1	algorithms
10	str	1	github
11	str	1	php
12	str	1	python
13	str	1	r
14	str	1	opencv
15	str	1	csharp
16	str	1	vb.net
17	str	1	ajax
18	str	1	sql
19	str	1	unix shell scripting
20	str	1	data warehouse
21	str	1	etl
22	str	1	pl/sql

edu_set - List (24 elements)

Index	Type	Size	
0	str	1	11
1	str	1	12
2	str	1	physics
3	str	1	chemisrty
4	str	1	maths
5	str	1	staticstics
6	str	1	biology
7	str	1	computer science
8	str	1	bsc
9	str	1	bca
10	str	1	msc
11	str	1	mca
12	str	1	m.phill
13	str	1	php
14	str	1	machnical
15	str	1	chemical
16	str	1	civil
17	str	1	computer
18	str	1	electrical
19	str	1	electronincs and telecommunication
20	str	1	diploma
21	str	1	degree
22	str	1	engineering
23	str	1	

Step 6: Final Structured Dataset, ready for Machine – learning Algorithm

Data_frame - DataFrame

Index	Education	Email	Last Modified	Mob no	Name	Number of exp	Resume	Skills Set
0	['bca', 'engineering']		Fri Apr 5 03:19:24 2019		Abhay Patil	Not specified	Abhay Patil.docx	[]
1	['engineering']	arshad5b2@gmail.com	Wed Apr 3 04:13:34 2019	8688562564	HAQUE B.TECH	Not specified	AHTESHAMUL HAQUE.docx	[]
2	[]	AkmalE86@gmail.com	Wed May 1 15:26:14 2019	8606845834	Akmal Husan	1.3 years	Akmal Eache.docx	['sql']
3	[]	Akshay123@gmail.com	Tue Apr 30 12:23:28 2019	7676857388	Akshay Balaso	1.3 years	Akshay.docx	['sql']
4	['engineering']	Rohan23hg@gmail.com	Tue Apr 30 12:32:24 2019	7899448856	Amrut Gore	Not specified	Amrut Gore.docx	['sql', 'unix shell scripting', 'data warehouse', 'etl']
5	['engineering']	varma@gmail.com	Wed Apr 3 04:03:52 2019	9876543210	Ayushi Varma	3 years	Ayushi Varma.docx	['python', 'sql']
6	[]	kiranshinde323@gmail...	Wed Apr 3 04:03:52 2019	9673888932	CURRICULUM VITAE	Not specified	CURRICULUM VITAE Kiran.docx	['html', 'vb.net', 'sql']
7	[]		Fri Apr 5 03:19:26 2019		Gayatri Saraswat	Not specified	Gayatri Saraswat.docx	[]
8	[]		Fri Apr 5 03:19:26 2019		Gurwinder Singh	Not specified	Gurwinder Singh.docx	[]
9	['bsc', 'engineering']	harry.saxena@yahoo.com	Wed Apr 3 04:13:36 2019		HARSHIT SAXENA	5 years	harshit_Delhi_4.08_yrs.docx	['vb.net', 'sql', 'data warehouse', 'etl']
10	[]		Fri Apr 5 03:19:26 2019		Kirti Chavan	6 months	Kirti Chavan.docx	[]
11	['mca']	komalDesai7788@gmail...	Wed Apr 3 04:18:12 2019		Komal Desai	Not specified	Komal Desai.docx	['html', 'opencv', 'vb.net', 'sql']
12	[]		Fri Apr 5 03:19:26 2019		Kuldeep Kour	Not specified	Kuldeep Kour.docx	[]
13	['bca', 'mca']	sharmakuldeep921@gmail...	Wed Apr 3 04:22:04 2019	8887408589	Kuldeep Rankumar	5 months	Kuldeep sharma.docx	[]
14	[]		Fri Apr 5 03:19:26 2019		Madhu Warma	Not specified	Madhu Warma.docx	[]
15	['mca', 'diploma']		Fri Apr 5 03:19:26 2019		Mahendra Bahubali	Not specified	Mahendra Bahubali.docx	[]
16	[]		Fri Apr 5 03:39:52 2019		Mahesh Ahmed	Not specified	Mahesh Ahmed.docx	[]
17	['engineering']	manish@gmail.com	Wed May 1 17:54:46 2019	9999448844	Manish patil	4.4 years 4 months	Manish patil.docx	['sql', 'data warehouse', 'etl']
18	[]		Fri Apr 5 03:39:54 2019		Manisha Kour	Not specified	Manisha Kour.docx	[]
19	['bsc']	MuraliJadhav23@gmail...	Tue Apr 30 12:29:56 2019	8999448856	Murali U	14 years	Murali U Jadhav.docx	['algorithms']
20	[]	murlijadhav007@gmail...	Mon Apr 29 14:57:16 2019	8412960018	Jadhav Murl	Not specified	Murli Jadhav.docx	['html', 'python', 'vb.net', 'sql']
21	['engineering']	Nitin34gore@gmail.com	Tue Apr 30 12:31:32 2019	4545448856	Nitin Gore	4.4 years 4 months	Nitin Gore.docx	['sql', 'data warehouse', 'etl']
22	[]		Fri Apr 5 03:19:26 2019		Poonam Mishra	Not specified	Poonam Mishra.docx	[]
23	['mca', 'mca']	pournimavanarase@gmail...	Wed Apr 3 04:18:12 2019	9184089224	Pournima Vanarase	3 months	Pournima.docx	['html', 'sql']
24	['engineering']		Fri Apr 5 03:19:28 2019		Rahul Mahajan	Not specified	Rahul Mahajan.docx	[]
25	['engineering']		Fri Apr 5 03:19:28 2019		Rajan Sing	Not specified	Rajan Sing.docx	[]

Format Resize ☒ Background color ☒ Column min/max Save and Close Close

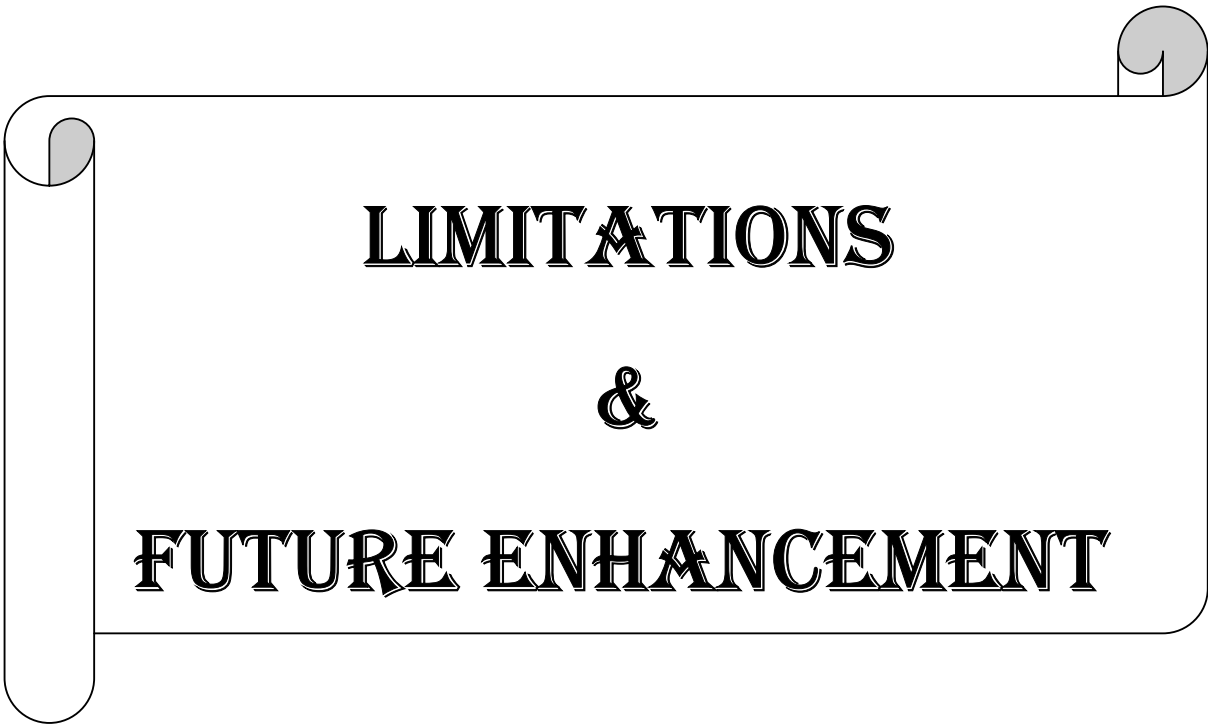


CONCLUSION

Conclusion:

Designing an automated system to extract information from unstructured resumes and transform that information to structured format. And ranking those resumes based on the information extracted, according to the skill sets of the candidate and based on the job description of the company.

The proposed model in this paper extracts necessary data from a CV/Resume and segments them based on their values. However the ranking and positive weight given for a CV/Resume might change based on different company or employers preference. As described, the whole process was segmented and each segmented was designed separately to perform its task. The segment that deals with Natural Language Processing actually worked with only the Natural Language Processing task and similarly the segments that deal with Machine Learning, completely deals with Machine Learning techniques. A different way of evaluating and analyzing the data in a CV/Resume is proposed in this paper and that was converting data into HTML code to understand different values. Finally, the model gives ranks to CVs/Resumes based on the necessary data and employers needs taking previous requirements in consideration.

A decorative scroll graphic with a light gray background and a black border. The scroll is unrolled in the center, revealing the text. The top and bottom edges of the scroll are curved, and the sides are straight. The text is centered within the unrolled portion.

LIMITATIONS

&

FUTURE ENHANCEMENT

Limitations & Future Scope :-

Limitations –

There is a lot of variability and ambiguity in the language used in CVs. There are many ways to write dates and numerous job titles and skills appear every month. Someone's name can be a company name (e.g. Harvey Nash) or even an IT skill (e.g. Cassandra). The only way a CV parser can deal with this is to "understand" the context in which words occur and the relationship between them. That is why a rule-based parser will quickly run into two big limitations: 1) the rules will get quite complex to account for exceptions and ambiguity, and 2) the coverage will be limited.

One of the issues with rule-based approaches is that no rule is 100% reliable. Take the example of finding the name in a CV. "Name: ", is not always a good left context, not all words after this context should be used, and entries in an extensive list of first names can be street names or other concepts etc.

Machine learning solves this problem by estimating the quality of these signals based on annotated data and through principled ways of combining the evidence from several signals. These signals are called features and encode information that we think is relevant for our prediction task (e.g. "is the word preceded by 'Name:', "does the word start with uppercase). In general, these features are manually written.

Future Scope -

Even though in the research one of the most feasible way to evaluate a CV/ Resume was detailed, the domain was kept restricted to the CVs/ Resumes of only engineering students and the amount of sample data versus the amount of test data was relatively small. In addition to that, CVs/ Resumes with some varied layout design is out of the scope of this paper. For the future scope of this research, the methodologies can be used for the data from CVs/ Resumes of other job departments or the whole research can be done in a much larger scope. Additionally, algorithms such as naive Bayes, logistic regression or c4.5 analysis can be performed to see if it improves the result. Therefore, the future scope is very broad.



BIBLIOGRAPHY

References –

1. elearning.excelr.com
2. <https://innodatatics.com>
3. <https://datatruks.com>
4. <https://github.com>
5. www.excelr.com
6. <https://medium.com/@divalicious.priya/information-extraction-from-cv-acec216c3f48>
7. <https://www.omkarpathak.in/projects/resume-parser/>