RESUME SUMMARIZER AND JOB DESCRIPTION MATCHER USING NATURAL LANGUAGE PROCESSING AND SPACY

Sai Tarun Boddu
121910317026@gitam.in
India,GITAM University
(deemed to be)

Sujeeth Desu
121910317001@gitam.in
India,GITAM University
(deemed to be)

Sreekanth Puli
spuli@gitam.edu
India, GITAM University
(deemed to be)

Abstract

In the HR recruitment process, one pivotal stage is "Summarizing & Screening." At this juncture, recruiters grapple with the laborious task of manually sifting through a myriad of resumes to shortlist suitable candidates. This project aims to streamline and automate this process using two innovative tools. Firstly, the "Resume Summarizer" leverages advanced natural language processing techniques to swiftly extract pertinent details from candidates' resumes, offering a concise summary that underscores their skills, experience, and qualifications. This summary can be tailored to match specific job requirements, facilitating recruiters in the comparative evaluation of multiple candidates. Secondly, the "Job Description Matcher" assists recruiters in expeditiously and accurately aligning job descriptions with candidate resumes. By employing natural language processing, this tool identifies keywords and phrases that align with the job opening's prerequisites, generating a compatibility score to rank resumes according to their alignment with the job description. This innovation greatly expedites candidate selection and enhances recruiters' efficiency. Importantly, this project employs a unique approach, incorporating regular expressions to extract essential information from resumes, setting it apart from existing NLP-based models and systems.

Keywords: Natural Language Processing, Text Analysis, Data Extraction, Keyword matching, Regular Expressions, Optical Character Recognition (OCR), N-Gram Model, Name Entity Recognition (NER), Unsupervised NER.

Introduction

In the rapidly evolving landscape of today's job market, recruiters face the formidable challenge of sifting through an overwhelming volume of resumes to identify the most suitable and qualified candidates for various job opportunities. Constrained by limited time and resources, the process of screening and pinpointing the ideal profiles from the pool of outsourced resumes becomes a daunting and time-consuming task. In response to this challenge, we have undertaken a project designed to simplify and optimize the recruitment process by introducing two essential tools: The Resume Summarizer and the Job Description Matcher.

The Resume Summarizer is a powerful tool that empowers recruiters to swiftly and efficiently condense extensive resumes into concise summaries. This tool enables recruiters to compare and evaluate potential candidates for job vacancies by extracting critical information from their resumes and providing a brief yet comprehensive overview. The Resume Summarizer not only saves valuable time and effort for recruiters but also ensures a comprehensive understanding of each candidate's skills, qualifications, and experience.

Complementing this, the Job Description Matcher is engineered to assist recruiters in aligning job descriptions with candidate resumes. It calculates a compatibility score that ranks resumes based on their alignment with the job description, identifying relevant keywords and phrases. This innovative tool equips recruiters with the means to identify the most promising prospects for a given job and confirms that they are selecting the most suitable candidate for the position.

These two tools, the Resume Summarizer and Job Description Matcher, represent integral components of our project, dedicated to enhancing and streamlining the recruitment process. When used in conjunction, they offer recruiters a powerful solution for simplifying the hiring process. Our project endeavors to provide recruiters with the tools and insights needed to make well-informed decisions about candidate selection, leveraging sophisticated Natural Language Processing techniques to result in more efficient recruiting processes and ultimately better hiring outcomes. As Natural Language Processing (NLP) continues to gain prominence across diverse applications, from virtual assistants like Siri and Alexa to chatbots and language translation tools, its significance as a resource for organizations and industries dealing with vast amounts of language-based data cannot be overstated, facilitating improved decision-making and operational efficiency.

Literature study

The literature review reveals a comprehensive landscape of research efforts related to resume summarization, job description matching, and the application of Natural Language Processing (NLP).

According to [Narendra et al.] the Proposed system uses the spaCy library for NER and tokenization, the system can extract all the required details from the resumes. Then, the extracted information is used to generate a summary of the candidate's profile. The system was tested on a dataset of 100 resumes and achieved an accuracy of 92% in entity extraction and 87% in summary generation.

Another work in [Suresh et al.] the proposed methodology makes an important contribution to the field of resume analytics and NLP, providing a novel approach that uses contextual information to improve the accuracy of information extraction of resumes.The model also incorporates contextual information such as job descriptions and industry-specific terminologies to improve the accuracy of the extraction process.

[Jiang et al.] The proposed approach uses a conventional neural network to perform local detection of named entities, and then uses conditional random field model to combine thee local detection into a final output. Overall, the research paper presents an innovative approach to named entity recognition that addresses some limitations of existing methods.

Another process proposed in [Sadiq et al.] The model preprocesses the resumes and performs tokenization, stemming, stop-word removal, Part-Of-Speech (POS) tagging. The model extract features from the resumes and used a SVM classifier to rank the resumes with respect to the Job description. The model is evaluated with a dataset of 500 resumes and achieves an accuracy of 85% in ranking the resumes.

These studies collectively enrich the understanding of NLP's significance in streamlining recruitment processes and offer valuable insights for our project.

Problem Identification and Objective

During the course of our research and project development, several potential challenges emerged, each of which our project was designed to address effectively. These challenges included the inherent variability in resume templates and data positioning, missing or incomplete information in some resumes, unstructured and poorly formatted data, and the diverse formats in which resumes can be uploaded. Additionally, existing resume parsers primarily catered to PDF and Word formats, limiting their compatibility. Extracting contact details using machine learning methods posed computational and time-consuming challenges.

To tackle these issues, our objectives were clear. We aimed to develop a precise resume summarizer capable of extracting key details such as work experience, education, and skills. Our job description matcher leveraged natural language processing techniques to align job requirements with candidate qualifications. For image-based resumes, we standardized formats before applying Optical Character Recognition using 'pytesseract.' To efficiently extract contact details, we employed regular expressions. Subsequently, textual data analysis with NLTK allowed us to identify technical skills by cross-referencing them with a comprehensive database sourced from LinkedIn, offering a holistic solution to these common recruitment challenges.

Proposed Methodology

With an advancement in Textual Analysis and NLP techniques, this study of Resume Summarizing is so focused on ease of implementation and accuracy. The model is an automated system for evaluating resumes using NLP techniques, the system uses a NLP techniques and SpaCy for performing resume screening and classification, the system has a matching algorithm to match the skills and qualifications of the job description with the candidate's resume.

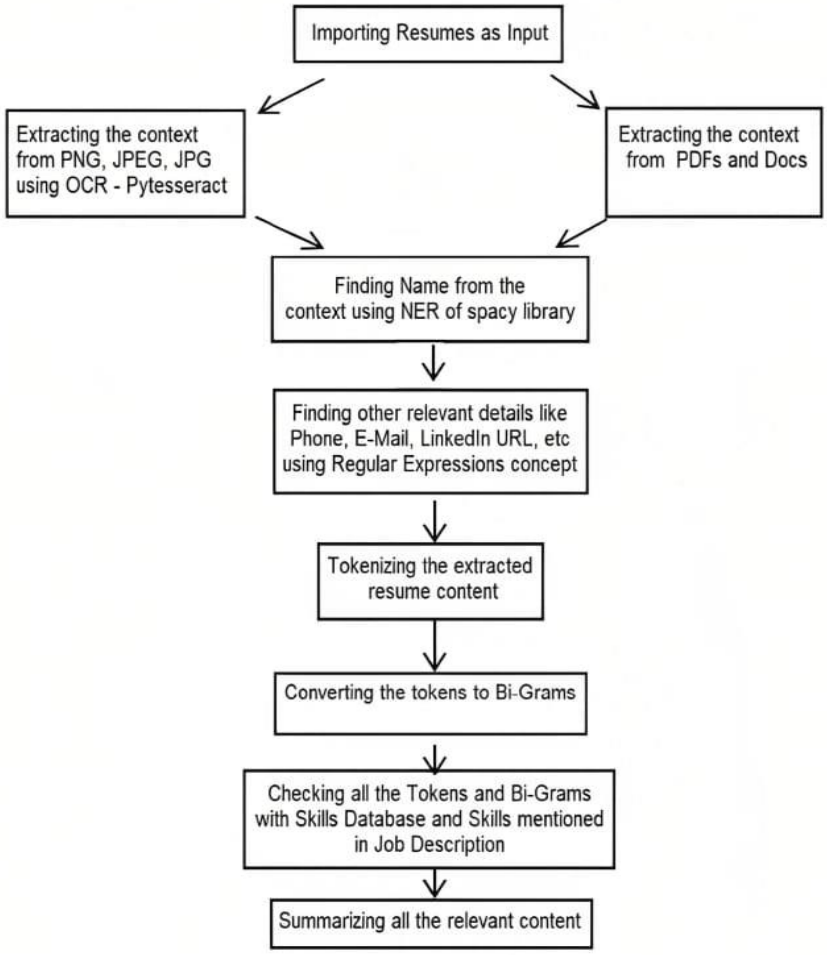

Importing Resumes as Input
Extracting the context from PNG, JPEG, JPG using OCR - Pytesseract
Extracting the context from PDFs and Docs
Finding Name from the context using NER of spacy library
Finding other relevant details like Phone, E-Mail, LinkedIn URL, etc using Regular Expressions concept
Tokenizing the extracted resume content
Converting the tokens to Bi-Grams
Checking all the Tokens and Bi-Grams with Skills Database and Skills mentioned in Job Description
Summarizing all the relevant content

Figure 1. System Workflow

Figure 2. shows the how the Resume Summarizer works, where the input can be PDF, DOCX, PNG, JPG, JPEG format and after the parsing the summary consists of the Candidate Name, Candidate Phone number, Candidate Email ID, Candidate LinkedIn profile link if mentioned, Skills set of the candidate.

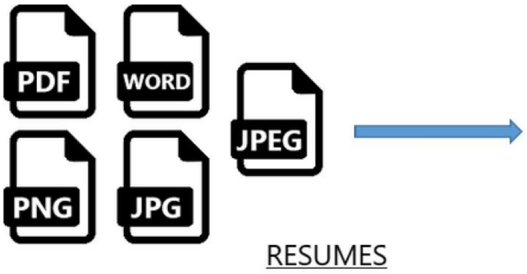
- Using **"pdfminer"** library and its functions we can extract the data from files which are in PDF format.
- Using **"docx2txt"** library and its functions we can extract the data from files which are in Docx format.
- For images, we have 3 different formats, so firstly we need to convert them into one particular common format such that it will be easy to apply Optical Character Recognition (OCR) using **"pytesseract"**.

After extracting the data from the documents, the model look for related details like Phone Number, Email ID, and LinkedIn Url using Regular Expressions.

- Regular Expression for Phone numbers - **'([+(]?\d+[)\-]?[\t\r\f\v]*[(]?\d{2,}[()\-]?[\t\r\f\v]*\d{2,}[()\-]?[\t\r\f\v]*\d*[\t\r\f\v]*\d*[\t\r\f\v]*)'**
- Regular Expression for Email - **'[a-zA-Z0-9\.\-+_]+@[a-zA-Z0-9\.\-+_]*\.(com|edu|net|in)'**
- Regular Expression for LinkedIn Profile link - **'((http(s?)://)*([www])*\.|[linkedin])[linkedin/~\-]+\.[a-za-z0-9/~\-_,&=\?\.;]+[^\.,\s<]'**

The Skillset from candidate resume will be finalized by using N-Gram Model, Unigram and Bigram concept, where each word will be considered as a token and the tokens can form bigrams too. Now the tokens and bigrams will undergoes keyword matching with a Database of 30000 skills scraped from Linkedin.

To extract the candidate name from the resume, we can use either Parts of Speech (POS) tagging or Name Entity Recognition (NER).

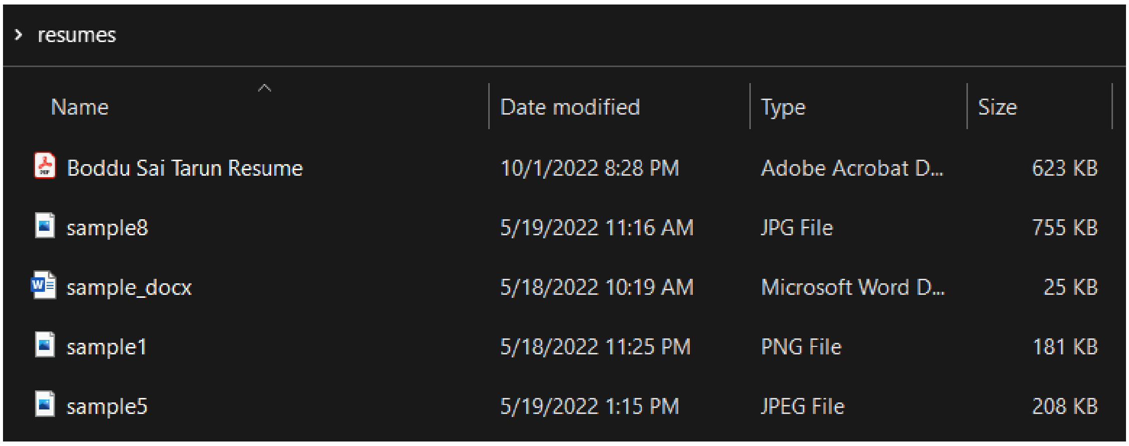


```
[{
'Name': 'Desu Sujeeth',
'Phone_no': '8374965202',
'Email': 'sujeethdesu01@gmail.com',
'Linkedin': None,
'Skills': ['OpenCV', 'and Management',
'learning', 'research project', 'Machine
Learning', 'Operating System', 'Data',
'Windows', 'Python', 'Java', 'data',
'Learning', 'Data visualization', 'HTML',
'linkedin', 'projects', 'CSS', 'research',
'Technology', 'Excel', 'python', 'excel',
'Management', 'C', 'Computer Science',
'Javascript', 'deep learning', 'trading']
}]
```

Figure 2. Input and Output of Resume Summarizer

Results

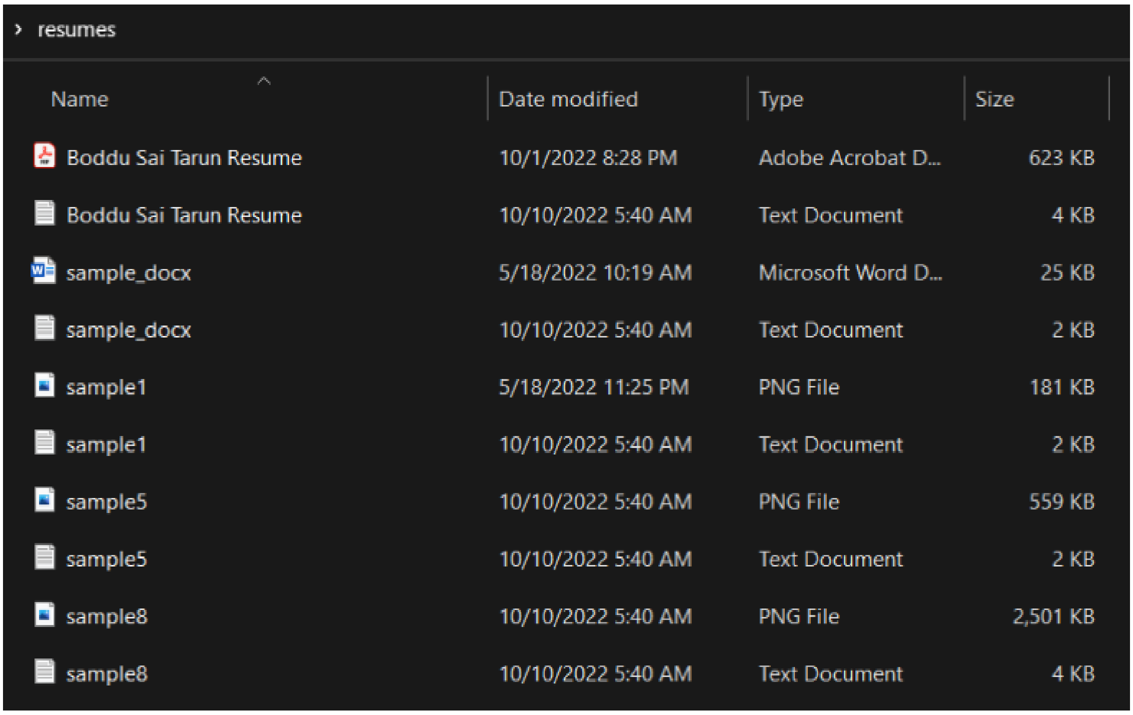
When the tool window opens we need to select the folder which consists of candidate resumes.
Figure 3. shows the resumes directory has 5 resumes of different file formats (PDF, JPG, DOCX, PNG, JPEG).



Name	Date modified	Type	Size
Boddu Sai Tarun Resume	10/1/2022 8:28 PM	Adobe Acrobat D...	623 KB
sample8	5/19/2022 11:16 AM	JPG File	755 KB
sample_docx	5/18/2022 10:19 AM	Microsoft Word D...	25 KB
sample1	5/18/2022 11:25 PM	PNG File	181 KB
sample5	5/19/2022 1:15 PM	JPEG File	208 KB

Figure 3. Format of Sample Inputs

Figure 4. shows that all resumes were converted into text files for further process parsing, and the 3 different formats of image resumes got converted to single image format (PNG format).

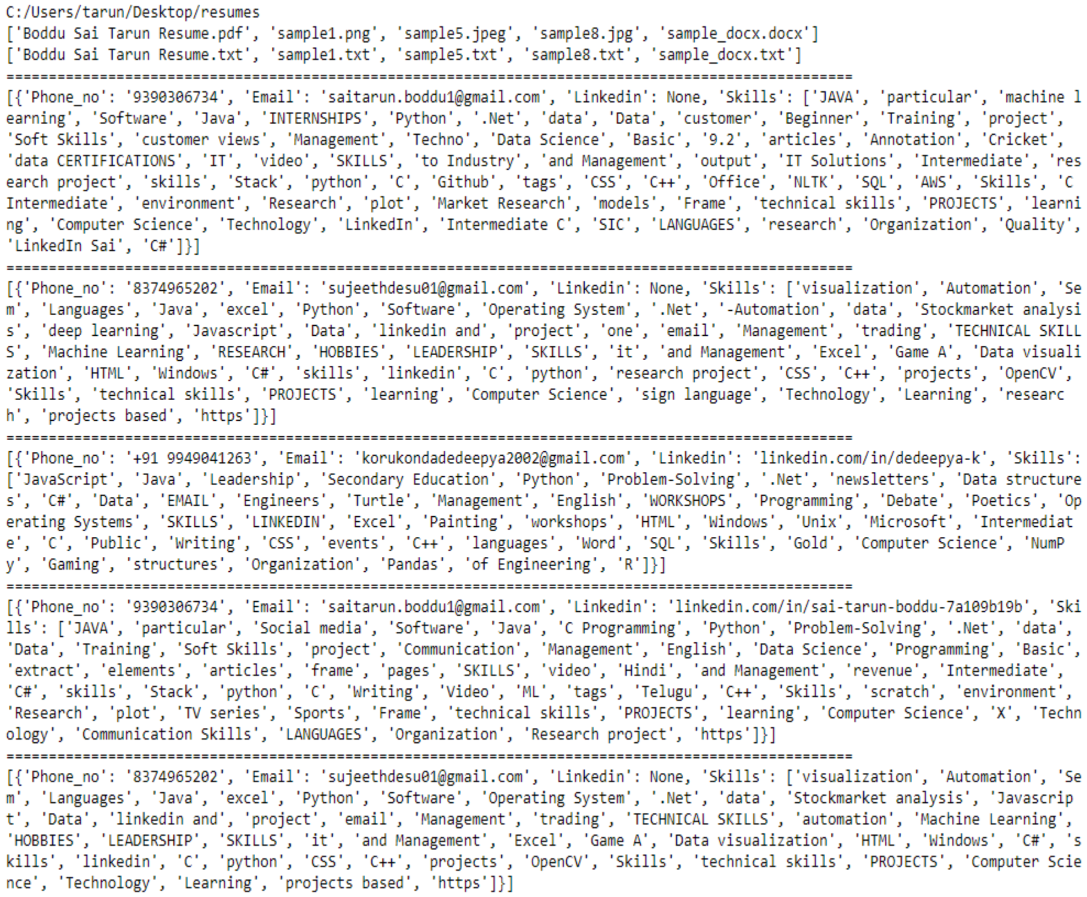


› resumes

Name	Date modified	Type	Size
Boddu Sai Tarun Resume	10/1/2022 8:28 PM	Adobe Acrobat D...	623 KB
Boddu Sai Tarun Resume	10/10/2022 5:40 AM	Text Document	4 KB
sample_docx	5/18/2022 10:19 AM	Microsoft Word D...	25 KB
sample_docx	10/10/2022 5:40 AM	Text Document	2 KB
sample1	5/18/2022 11:25 PM	PNG File	181 KB
sample1	10/10/2022 5:40 AM	Text Document	2 KB
sample5	10/10/2022 5:40 AM	PNG File	559 KB
sample5	10/10/2022 5:40 AM	Text Document	2 KB
sample8	10/10/2022 5:40 AM	PNG File	2,501 KB
sample8	10/10/2022 5:40 AM	Text Document	4 KB

Figure 4. Format of Sample Outputs

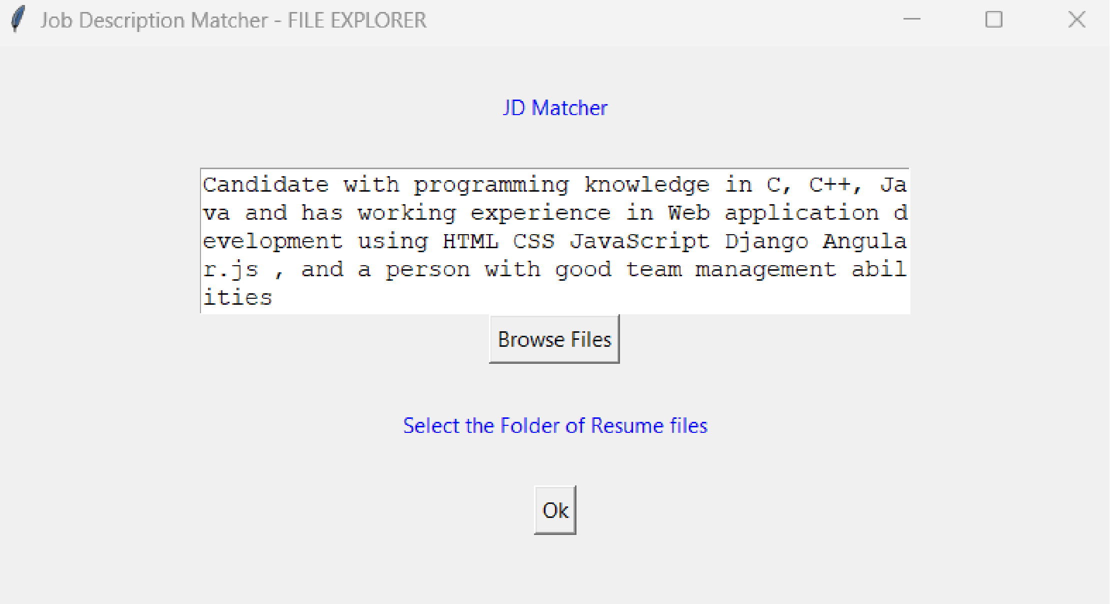
Figure 5. will show the parsed summary of each individual resume



```
C:/Users/tarun/Desktop/resumes
['Boddu Sai Tarun Resume.pdf', 'sample1.png', 'sample5.jpeg', 'sample8.jpg', 'sample_docx.docx']
['Boddu Sai Tarun Resume.txt', 'sample1.txt', 'sample5.txt', 'sample8.txt', 'sample_docx.txt']
==========================================================================================
[{'Phone_no': '9390306734', 'Email': 'saitarun.boddu1@gmail.com', 'Linkedin': None, 'Skills': ['JAVA', 'particular', 'machine l
earning', 'Software', 'Java', 'INTERNSHIPS', 'Python', '.Net', 'data', 'Data', 'customer', 'Beginner', 'Training', 'project',
'Soft Skills', 'customer views', 'Management', 'Techno', 'Data Science', 'Basic', '9.2', 'articles', 'Annotation', 'Cricket',
'data CERTIFICATIONS', 'IT', 'video', 'SKILLS', 'to Industry', 'and Management', 'output', 'IT Solutions', 'Intermediate', 'res
earch project', 'skills', 'Stack', 'python', 'C', 'Github', 'tags', 'CSS', 'C++', 'Office', 'NLTK', 'SQL', 'AWS', 'Skills', 'C
Intermediate', 'environment', 'Research', 'plot', 'Market Research', 'models', 'Frame', 'technical skills', 'PROJECTS', 'learni
ng', 'Computer Science', 'Technology', 'LinkedIn', 'Intermediate C', 'SIC', 'LANGUAGES', 'research', 'Organization', 'Quality',
'LinkedIn Sai', 'C#']}]
==========================================================================================
[{'Phone_no': '8374965202', 'Email': 'sujeethdesu01@gmail.com', 'Linkedin': None, 'Skills': ['visualization', 'Automation', 'Se
m', 'Languages', 'Java', 'excel', 'Python', 'Software', 'Operating System', '.Net', '-Automation', 'data', 'Stockmarket analysi
s', 'deep learning', 'Javascript', 'Data', 'linkedin and', 'project', 'one', 'email', 'Management', 'trading', 'TECHNICAL SKILL
S', 'Machine Learning', 'RESEARCH', 'HOBBIES', 'LEADERSHIP', 'SKILLS', 'it', 'and Management', 'Excel', 'Game A', 'Data visuali
zation', 'HTML', 'Windows', 'C#', 'skills', 'linkedin', 'C', 'python', 'research project', 'CSS', 'C++', 'projects', 'OpenCV',
'Skills', 'technical skills', 'PROJECTS', 'learning', 'Computer Science', 'sign language', 'Technology', 'Learning', 'researc
h', 'projects based', 'https']}]
==========================================================================================
[{'Phone_no': '+91 9949041263', 'Email': 'korukondadedeepya2002@gmail.com', 'Linkedin': 'linkedin.com/in/dedeepya-k', 'Skills':
['JavaScript', 'Java', 'Leadership', 'Secondary Education', 'Python', 'Problem-Solving', '.Net', 'newsletters', 'Data structure
s', 'C#', 'Data', 'EMAIL', 'Engineers', 'Turtle', 'Management', 'English', 'WORKSHOPS', 'Programming', 'Debate', 'Poetics', 'Op
erating Systems', 'SKILLS', 'LINKEDIN', 'Excel', 'Painting', 'workshops', 'HTML', 'Windows', 'Unix', 'Microsoft', 'Intermediat
e', 'C', 'Public', 'Writing', 'CSS', 'events', 'C++', 'languages', 'Word', 'SQL', 'Skills', 'Gold', 'Computer Science', 'NumP
y', 'Gaming', 'structures', 'Organization', 'Pandas', 'of Engineering', 'R']}]
==========================================================================================
[{'Phone_no': '9390306734', 'Email': 'saitarun.boddu1@gmail.com', 'Linkedin': 'linkedin.com/in/sai-tarun-boddu-7a109b19b', 'Ski
lls': ['JAVA', 'particular', 'Social media', 'Software', 'Java', 'C Programming', 'Python', 'Problem-Solving', '.Net', 'data',
'Data', 'Training', 'Soft Skills', 'project', 'Communication', 'Management', 'English', 'Data Science', 'Programming', 'Basic',
'extract', 'elements', 'articles', 'frame', 'pages', 'SKILLS', 'video', 'Hindi', 'and Management', 'revenue', 'Intermediate',
'C#', 'skills', 'Stack', 'python', 'C', 'Writing', 'Video', 'ML', 'tags', 'Telugu', 'C++', 'Skills', 'scratch', 'environment',
'Research', 'plot', 'TV series', 'Sports', 'Frame', 'technical skills', 'PROJECTS', 'learning', 'Computer Science', 'X', 'Techn
ology', 'Communication Skills', 'LANGUAGES', 'Organization', 'Research project', 'https']}]
==========================================================================================
[{'Phone_no': '8374965202', 'Email': 'sujeethdesu01@gmail.com', 'Linkedin': None, 'Skills': ['visualization', 'Automation', 'Se
m', 'Languages', 'Java', 'excel', 'Python', 'Software', 'Operating System', '.Net', 'data', 'Stockmarket analysis', 'Javascrip
t', 'Data', 'linkedin and', 'project', 'email', 'Management', 'trading', 'TECHNICAL SKILLS', 'automation', 'Machine Learning',
'HOBBIES', 'LEADERSHIP', 'SKILLS', 'it', 'and Management', 'Excel', 'Game A', 'Data visualization', 'HTML', 'Windows', 'C#', 's
kills', 'linkedin', 'C', 'python', 'CSS', 'C++', 'projects', 'OpenCV', 'Skills', 'technical skills', 'PROJECTS', 'Computer Scie
nce', 'Technology', 'Learning', 'projects based', 'https']}]
```

Figure 5. Output of Resume Summarizer for Sample Input

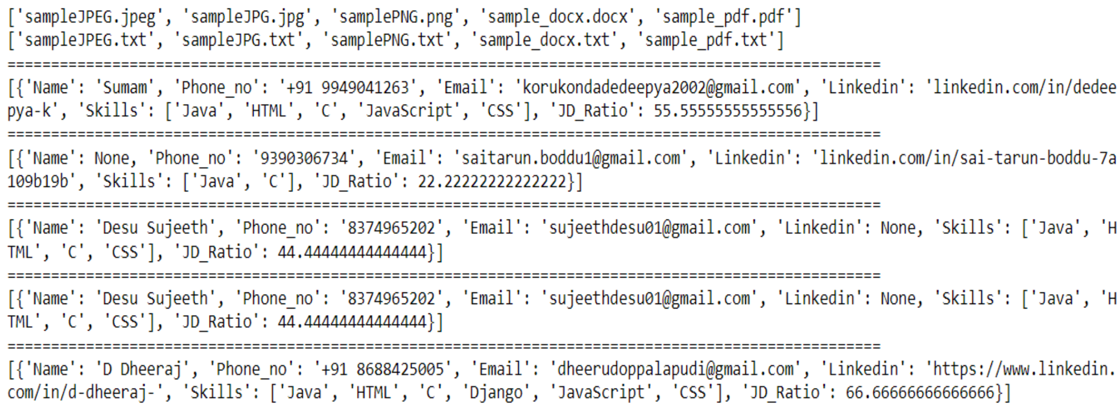
Figure 4. The Job Description Matcher, when the tool window opens we need to select the folder which consists of candidate resumes by clicking the button of ‘Browse Files’ button but for this tool we must give another input that is Job Description which is a textual input that will be given by recruiters.
Figure 3. shows the resumes directory which has 5 resumes of different file formats (PDF, JPG, DOCX, PNG, JPEG), which is the same directory of input files that is tested for the previous tool Resume summarizer.


Job Description Matcher - FILE EXPLORER
JD Matcher
Browse Files
Select the Folder of Resume files
Ok

```
Candidate with programming knowledge in C, C++, Ja
va and has working experience in Web application d
evelopment using HTML CSS JavaScript Django Angula
r.js , and a person with good team management abil
ities
```

Figure 4. Input way of Job Description Matcher tool

Now the tool will rank the resumes by looking for essential keywords of the Job Description in the folder of the resumes which can be of any format by giving a parameter called JD_Ratio as output, which means the tool is assigning a compatibility score for each resume with respect to the Job Description which is given as input.



```
['sampleJPEG.jpeg', 'sampleJPG.jpg', 'samplePNG.png', 'sample_docx.docx', 'sample_pdf.pdf']
['sampleJPEG.txt', 'sampleJPG.txt', 'samplePNG.txt', 'sample_docx.txt', 'sample_pdf.txt']
=========================================================================================
[{'Name': 'Sumam', 'Phone_no': '+91 9949041263', 'Email': 'korukondadedeepya2002@gmail.com', 'Linkedin': 'linkedin.com/in/dedee
pya-k', 'Skills': ['Java', 'HTML', 'C', 'JavaScript', 'CSS'], 'JD_Ratio': 55.55555555555556}]
=========================================================================================
[{'Name': None, 'Phone_no': '9390306734', 'Email': 'saitarun.boddu1@gmail.com', 'Linkedin': 'linkedin.com/in/sai-tarun-boddu-7a
109b19b', 'Skills': ['Java', 'C'], 'JD_Ratio': 22.22222222222222}]
=========================================================================================
[{'Name': 'Desu Sujeeth', 'Phone_no': '8374965202', 'Email': 'sujeethdesu01@gmail.com', 'Linkedin': None, 'Skills': ['Java', 'H
TML', 'C', 'CSS'], 'JD_Ratio': 44.44444444444444}]
=========================================================================================
[{'Name': 'Desu Sujeeth', 'Phone_no': '8374965202', 'Email': 'sujeethdesu01@gmail.com', 'Linkedin': None, 'Skills': ['Java', 'H
TML', 'C', 'CSS'], 'JD_Ratio': 44.44444444444444}]
=========================================================================================
[{'Name': 'D Dheeraj', 'Phone_no': '+91 8688425005', 'Email': 'dheerudoppalapudi@gmail.com', 'Linkedin': 'https://www.linkedin.
com/in/d-dheeraj-', 'Skills': ['Java', 'HTML', 'C', 'Django', 'JavaScript', 'CSS'], 'JD_Ratio': 66.66666666666666}]
```

Conclusion

In summary, our project successfully aids recruiters in streamlining the resume shortlisting process through the integration of the RESUME SUMMARIZER and JOB DESCRIPTION MATCHER tools. Leveraging cutting-edge technology, we have seamlessly combined Natural Language Processing techniques and Python libraries, including spaCy and NLTK, to achieve this goal. Our approach involved OCR and image processing for text extraction, followed by preprocessing steps. We utilized spaCy for keyword and phrase identification, employing bigrams and trigrams for n-gram analysis. The tools, employing advanced algorithms, offer efficiency and accuracy in recruiting, reducing screening time and increasing the potential for identifying qualified candidates, making it a versatile solution for various industries and job types. This project highlights the effectiveness of NLP and machine learning in recruitment and hiring, enabling comprehensive text analysis and valuable insights for recruiters.

References

[Narendra et al.] Narendra.G.O and Hashwanth.S. Named Entity Recognition based Resume Parser and Summarizer. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 2, Issue1, March 2022

[Vrinda et al.] Vrinda Mittal, Priyanshu Mehta, Devanjali Relan, Goldie Gabrani. (2020) Methodology for resume parsing and job domain prediction. Journal of Statistics and Management Systems 23:7, pages 1265-1274**.**

[Jiang et al.] Xu, M., Jiang, H., & Watcharawittayakul, S. (2017, July). A local detection approach for named entity recognition and mention detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1237-1247).

[Sanyal et al.] Sanyal, S., Hazra, S., Adhikary, S., & Ghosh, N. (2017). Resume parser with natural language processing. International Journal of Engineering Science, 4484**.**

[Suresh et al.] Suresh, Y., & Manusha Reddy, A. (2021). A contextual model for information extraction in resume analytics using NLP's spacy. In Inventive Computation and Information Technologies: Proceedings of ICICIT 2020 (pp. 395-404). Singapore: Springer Singapore**.**

[Bhor et al.] Bhor, S., Gupta, V., Nair, V., Shinde, H., & Kulkarni, M. S. (2021). Resume parser using natural language processing techniques. Int. J. Res. Eng. Sci, 9(6).

[Nimbekar et al.] Nimbekar, R., Patil, Y., Prabhu, R., & Mulla, S. (2019, December). Automated resume evaluation system using NLP. In 2019 International Conference on Advances in Computing, Communication and Control (ICAC3) (pp. 1-4). IEEE.

[Sadiq et al.] Sadiq, S. Z. A. M., Ayub, J. A., Narsayya, G. R., Ayyas, M. A., & Tahir, K. T. M. (2016). Intelligent hiring with resume parser and ranking using natural language processing and machine learning. International Journal of Innovative Research in Computer and Communication Engineering, 4(4), 7437-7444.