PROJECT INTIALIZATION DOCUMENT

1. DETAILS

Team Details: Team consist of three people, and they are **Milind Sai (xxxxxx@gmail.com), Zubair Siddique (xxxx@gmail.com), Harshit Mehta (xxxxxxxxxx@gmail.com)**

Project Name: Resume Parser

2. DEFINING PROJECT AND ITS SCOPE

Understanding of the project: By ranking a resume, the employer can find the best candidates in no time. However, there is a need to find the factors that affect the ranking of the resume for a certain job description or employer. Some of the factors that we came across are as follows: i) Job description, ii) Number of experiences required, iii) Required skills, iv) Preferred location, v) Able to travel, vi) Minimum qualifications, vii) Past roles, viii) Projects, ix) Internship or Past work history, x) Workshop or Open-source contribution, xi) Certifications, xii) Honours and Awards, xiii) Volunteer experience and more. Current corporate companies and recruitment agencies process numerous resumes daily which is no human task. Since most of the time resumes are parsed manually, many times good candidates are not picked. Even though there are many elementary techniques for parsing structured documents, they are not suitable for parsing unstructured documents like resumes. We are going to overcome these problems using a Machine Learning model connected to a centralized database server where user can upload their resume on the web platform. The parser parses all the necessary information from the resume and auto-fills a form for the user to proofread. Once the user confirms, the resume is saved into our NoSQL database ready to show itself to the employers. Also, the user gets their resume in both JSON format and pdf.

Reason for choosing this project: We chose this problem for the reason of technology and interest. We are also quite interested in the concept of machine learning and text processing to rank the resume for effective selection of the right candidate. We have previously done few small projects on machine learning, and it would be interesting to use it in real-life applications. Also, we all have a good grasp of python technologies which we are going to use in our project. Each member had their way of solving this problem using different machine learning algorithms and we finally ended up with a common solution. The ideas didn’t stop flowing when we came across the problem.

Most challenging aspect of the project statement: Solving a problem using machine learning concepts is quite challenging. Also, the time complexity for text processing of the resumes needs to be considered. The decision of selecting the most suitable algorithm based on the most effective output is challenging. The conversion of resumes from different formats (like .pdf, .docx, .doc, .odt, etc) into text format is a very intricate task.

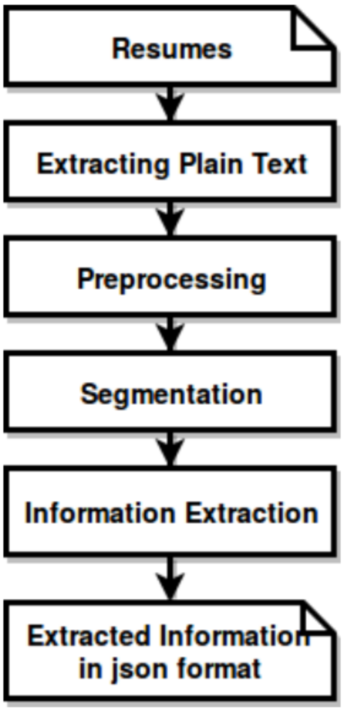
3. APPROACH OF PROBLEM CHOSEN

Approach: Below is the approach that we are going to follow.

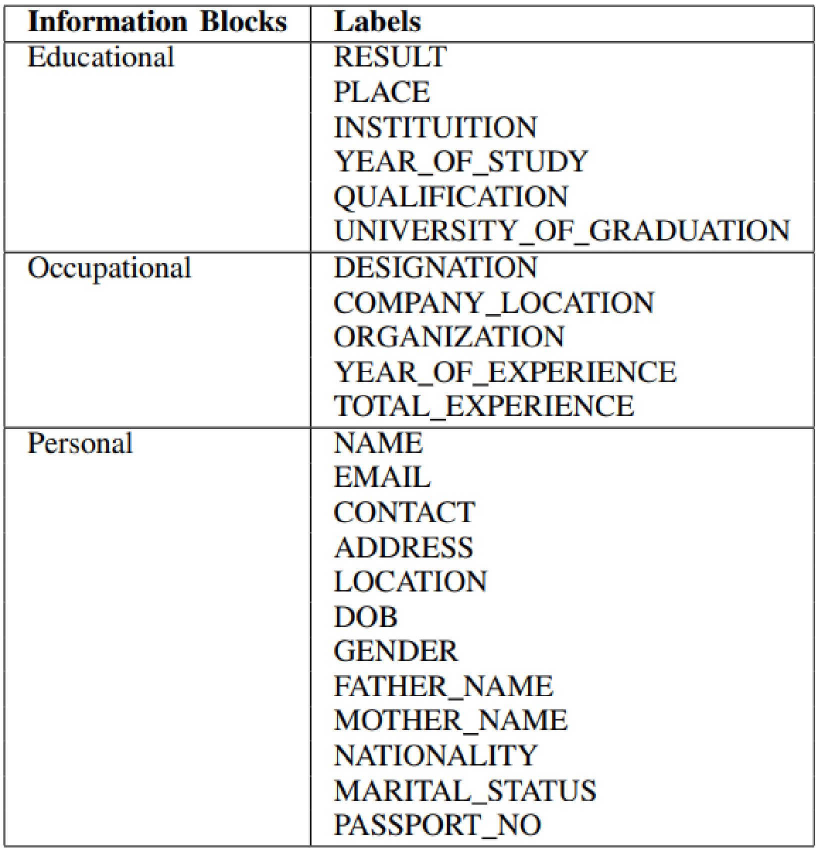
1. We are going to divide our model into two major parts which are The Uploader and The Parser. The Uploader takes the resumes from the client-side. The algorithm will block any file having extension other than '.pdf', '.doc', '.docx', '.odt', ‘.ods’, '.txt'. After the client has successfully uploaded the file, the algorithm takes the file, reads the contents, and writes the content into a text file before passing on the data to the parser. The Parser parses all the relevant data from the uploaded resume including name, emails, etc through natural language processing.
2. Initially, we will collect 1000s of resumes and will perform data pre-processing on them. It includes steps like Data Cleaning, Data Integration, Data Transformation, Data Reduction, Data Discretization.
3. After that, we will perform EDA on the dataset to understand the relationship between the parameters. Here we can use Linear/Lasso Regression and with the help of visualization tools like Matplotlib to visualize the data to better understand.
4. The above step will help us to identify the outliers or the most important features for the resume ranking.
5. After that, we will perform Tokenization and Lemmatization on the dataset to ignore the common words like ‘I’, ‘We’, ‘is’, ‘are’, etc.
6. Then we will perform Stemming and Segmentation to give a label to each information we have like “Python:Skills” whereas “IBM:Company”.
7. We will be using extensive Natural Language Processing to find accurate results.
8. Cosine Similarity using Manhattan Distance will be used to find the similarity between the Job description and resume to rank the resume.
9. To improve the accuracy, we will be performing mathematical calculations to make it more efficient and flexible using python libraries.
10. We will be using Python libraries like NumPy, Sci-kit learn, TensorFlow, matplotlib for machine learning using Jupyter notebook. NLP includes CNN, LSTM, Bi-LSTM, CRF based models.
11. After training the model we will test the model, if any error is produced, then the data will be sent back to the cleaning process to remove the error so that the model can produce a more accurate result. If there is any data leakage, we will hold back a validated dataset for a final check on our developed model.
12. If the accuracy is as wanted, then the model will be deployed and predicted ranking will be displayed. The output will contain all the fields being categorized into a different bucket.

Diagram/Flowchart:

Below is the basic architecture for the Resume Parser.

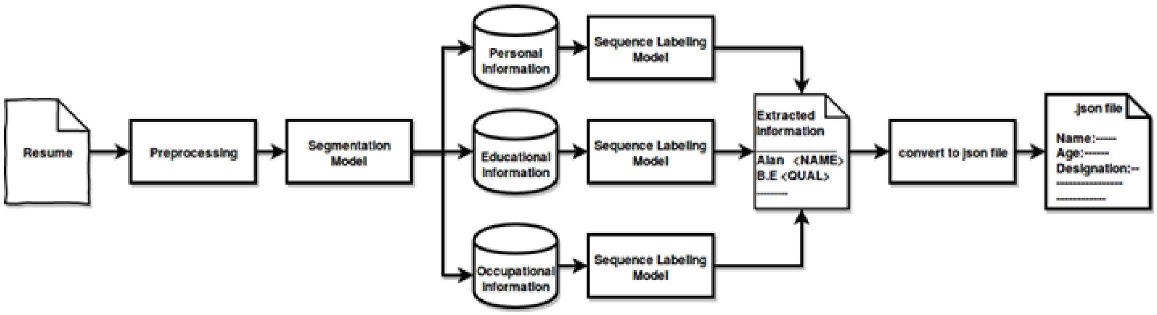


The extracted information can be categorized into a different bucket as shown below.



Information Blocks	**Labels**
Educational	RESULT
	PLACE
	INSTITUITION
	YEAR_OF_STUDY
	QUALIFICATION
	UNIVERSITY_OF_GRADUATION
Occupational	DESIGNATION
	COMPANY_LOCATION
	ORGANIZATION
	YEAR_OF_EXPERIENCE
	TOTAL_EXPERIENCE
Personal	NAME
	EMAIL
	CONTACT
	ADDRESS
	LOCATION
	DOB
	GENDER
	FATHER_NAME
	MOTHER_NAME
	NATIONALITY
	MARITAL_STATUS
	PASSPORT_NO

Detailed System Architecture for Resume Parser.



Platform/Coding Language/Frameworks: Jupyter Notebook, Google Colab, Visual studio Code, Python, Node.js, HTML, CSS, JavaScript, Pandas, NumPy, Ski-kit learn, Convolutional Neural Network, Flask/Django

Database/Cloud/Hosting: MongoDB, VPS Hosting

External tools: We may use cloud storage for storing data if we will have to deal with an optimal amount of data to train the model. We may need to use Google Colab Pro for a large amount of text processing. We may use Google's pre-trained text blob to compare the result and accuracy. We can also use external libraries available in Python such as Spacy for NLP.

4. TEAMS ABILITY TO IMPLEMENT WINNING SOLUTION

Background of team members: Milind Sai is currently a final year student pursuing B.Tech from NIT Goa. Zubair Siddiqui has completed B.Tech in Electronic and Telecommunication Engineering from Rizvi College of Engineering. He is currently working as Data Engineer with Wipro Technologies. Harshit Mehta is currently working as a Financial Advisory Analyst at Deloitte. He wants to make a transition to Data science and therefore he enrolled in this project with us.

Major Expertise of team members: Milind Sai has the knowledge of DSA and been doing competitive programming since a long time. He also has an interest in Machine Learning and Artificial Intelligence. Zubair Siddiqui on the other hand is working as Data Engineer, so he can perform various EDA and can understand the data very well. Harshit Mehta has the knowledge of people and finance, he also has knowledge of Python, R, Excel and can help gathering datasets. We all together can help each other to solve the problem statement and to complete the project.

Roles and responsibilities of team members: Milind Sai will be working on Image Processing techniques to build the parser and get the extracted text. The dataset will be gathered by Harshit. Then both Zubair and Milind will work on NLP and build an algorithm for resume parser. Harshit and Zubair will later work on APIs and building the frontend. We all will then create the whole system and push it to the hosting and make it live.

Previous projects undertaken: We all have done various projects individually. We all are currently working on IoT based (smart streetlight, smart health monitoring system) project as a college major project with different groups. We are working on a web development project (Academic Selection Portal, Chatbot). We all have worked on projects like Air Quality Index prediction, Toxic Text Prediction, Driver Drowsiness Detection using OpenCV, etc.

Team strengths: We all have a common interest in Machine Learning. Our teammates have working experience as individuals as well as group projects. The most important thing is the understanding between our teammates which is extremely important to go ahead and reach the finale. We all have that competitive spirit in us. Our coding fundamentals are clear.

Team achievements: One of our members had taken part in Smart India Hackathon 2020. One of our members has cleared Python Level 1 exam by Cambridge Certification Authority.

Personal motivation: As our approach is to use machine learning and natural language processing, applying mathematics like statistics and linear algebra to data for visualizing and predicting analysis is very intriguing and above all applying it to a challenging real-life problem like ranking the resume and extracting the information, and giving it a label is exciting.