NLP-Based Resume Parsing and Skills Extraction

Ranjith James
Satvik Kundargi
Andria Grace
Sathwik kuchana

Introduction

Streamline recruitment by automating resume processing, skill extraction, and candidate-job matching.

Objective:

- Automate Resume Processing
- Extract and Analyze Key Skills
- Match Candidates to Job Descriptions

Main Components

1. **Resume Parsing:** Extract text from various formats (PDF, DOCX)
2. **Skill Extraction:** Identify key skills using NLP techniques
3. **Matching Algorithm:** Align skills with job descriptions and rank candidates
4. **User Interface:** Easy-to-use platform for HR professionals to manage resumes and job descriptions.

Timeline and Assumptions

Project milestones: Data collection, model training, testing, deployment

Assumptions: Availability of quality resumes, accuracy of skill extraction

Timeline: 7 weeks from start to finish

Week 1: Planning & Data Collection
Week 2: Data Preprocessing
Week 3: Model Development
Week 4: Integration & UI Development
Week 5: Testing
Week 6: UI Enhancement
Week 7: Final Testing & Deployment

Dataset

Dataset URL: Kaggle Resume Dataset

Dataset Composition:

Total Resumes: More than 2,400 resumes
Formats: Available as both text within a CSV file and as PDF documents

CSV File: Includes resumes in text format and associated metadata

- **ID:** A unique identifier for each resume, corresponding to the PDF filename
- **Resume_str:** The text content of the resume
- **Resume_html:** Resume data in HTML format, derived from web scraping
- **Category:** The job category associated with each resume

PDF Documents:

- **Storage:** Organized in a data folder, divided into subfolders by job category
- **Naming Convention:** Each PDF is named using its unique ID for easy reference

This dataset provides a rich resource for developing and testing NLP models for tasks such as resume parsing, job matching, and recruitment analytics.

Design Summary

- **Data Ingestion:** Handles input resumes
- **Resume Parsing:** Extracts structured data from resumes
- **Skill Extraction:** Identifies and extracts skills from the parsed data
- **Model Training:** Trains the model for job matching
- **User Interface:** Provides an interactive interface for users

Methodology

Data Ingestion: The system ingests resumes from various formats into a structured format for further processing

- Reading Resumes: Utilizes PyPDF2 to process PDF resumes
- Conversion: Converts PDF content into a structured format for ease of parsing

Resume Parsing: Parses resumes to extract text and relevant sections such as experience, education, and skills.

- NLP Tasks: Uses spaCy for tokenization, Named Entity Recognition (NER)
- Section Extraction: Identifies and extracts key sections from the resume text for further analysis

Methodology

Skill Extraction
Extracts skills from parsed text using predefined skill patterns.

- Pattern Matching: Utilizes a JSON file containing patterns to identify skills and compares text against skill patterns defined in Skill_patterns.json
- Normalization: Extracts and normalizes skill names for consistency

Model Training

Training a machine learning model to match resumes with job descriptions.

- Data Preparation: Cleans and prepares data for model training
- Feature Extraction: Extracts relevant features from resumes and job descriptions
- Model Training: Employs various algorithms to train the matching model
- Evaluation: Assesses model performance using metrics such as precision, recall, and F1-score

User Interface

User Interface - HR Application

An interactive HR application for managing resumes and job matches.

- **Interface:** Built using Gradio for a user-friendly experience
- **Features:** Allows uploading resumes, viewing parsed data, and matching candidates to jobs

User Interface - User Application

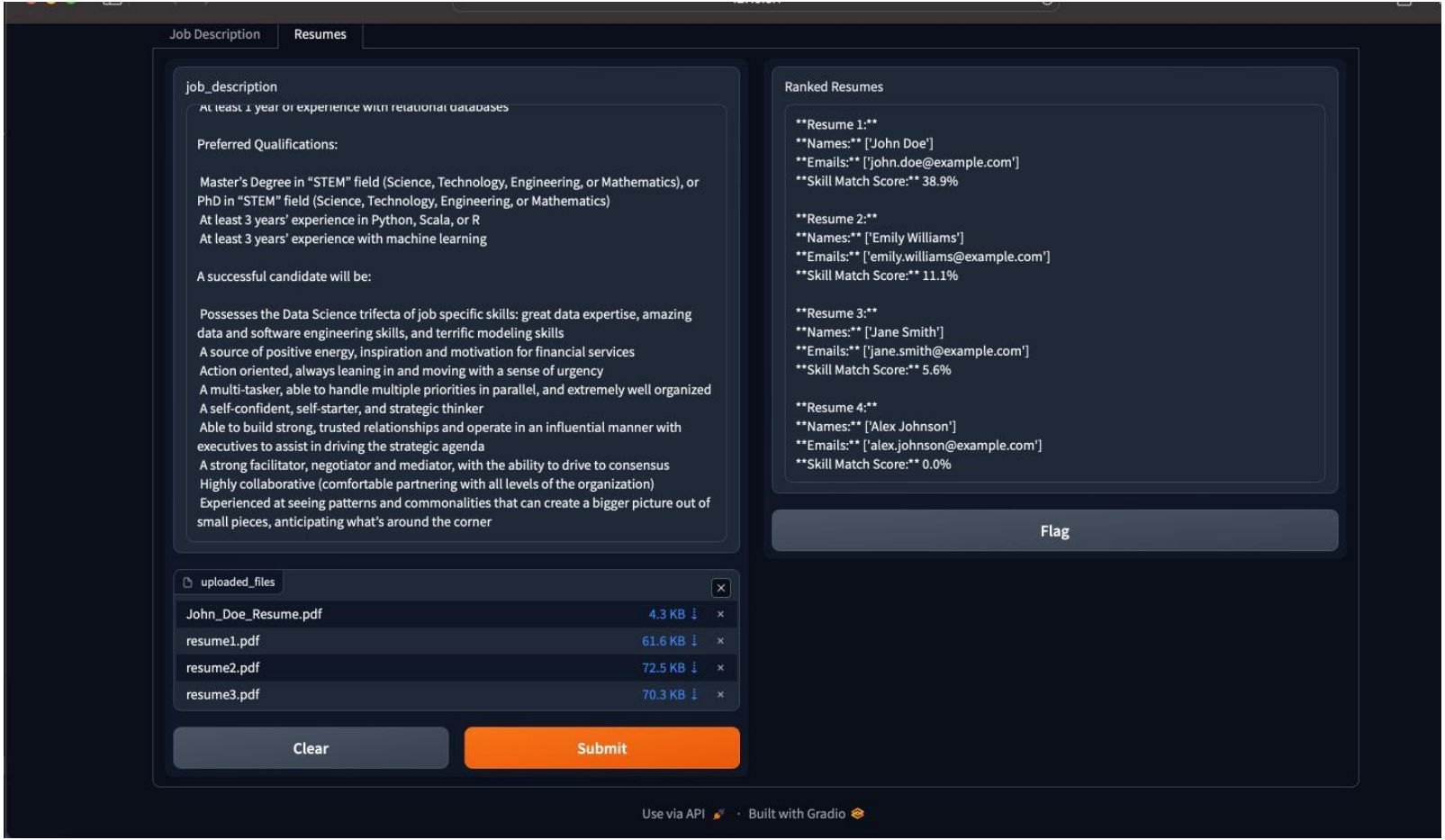
A user-facing application that allows candidates to upload their resumes and receive feedback.

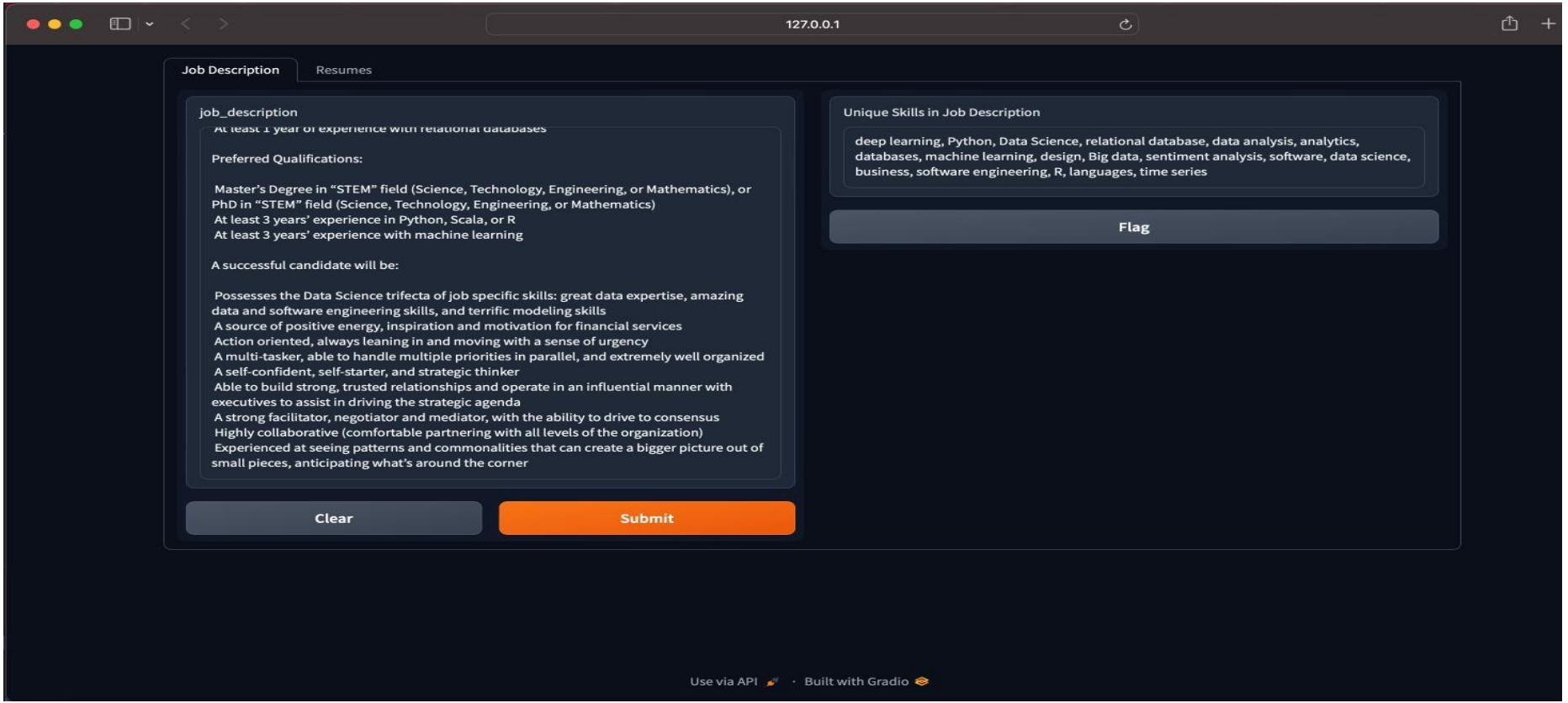
- **Resume Upload:** Enables candidates to upload their resumes
- **Feedback:** Provides feedback on skills and job matches based on the parsed data

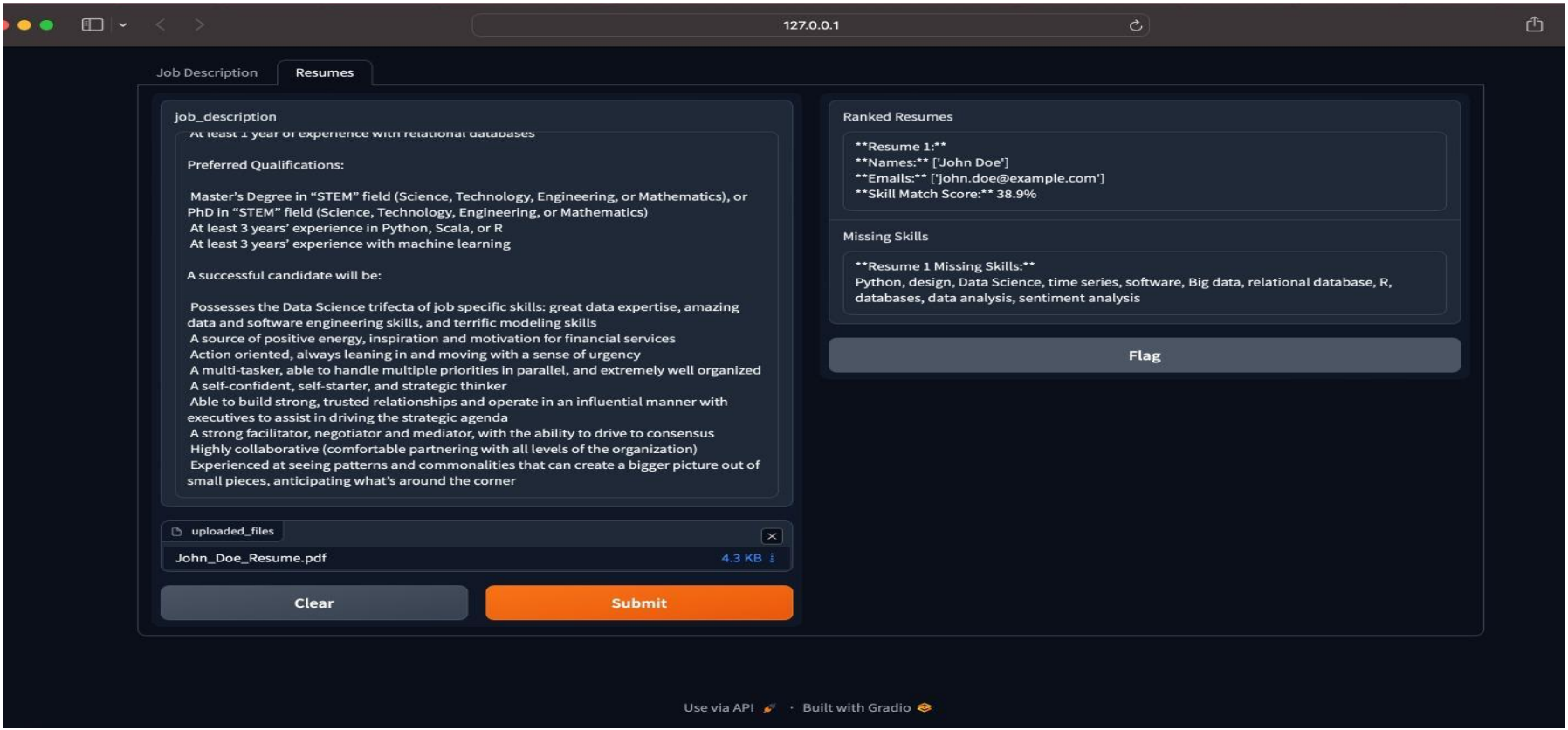
Requirements and Dependencies

- **spaCy:** For NLP tasks
- **nltk:** For additional NLP functionality
- **jsonlines:** For handling JSON Lines files
- **requests:** For making HTTP requests
- **PyPDF2:** For PDF processing
- **gradio:** For building interactive ML Interface

Solution Demo (Live)




127.0.0.1
Job Description
Resumes
job_description
At least 1 year of experience with relational databases
Preferred Qualifications:
Master's Degree in "STEM" field (Science, Technology, Engineering, or Mathematics), or PhD in "STEM" field (Science, Technology, Engineering, or Mathematics)
At least 3 years' experience in Python, Scala, or R
At least 3 years' experience with machine learning
A successful candidate will be:
Possesses the Data Science trifecta of job specific skills: great data expertise, amazing data and software engineering skills, and terrific modeling skills
A source of positive energy, inspiration and motivation for financial services
Action oriented, always leaning in and moving with a sense of urgency
A multi-tasker, able to handle multiple priorities in parallel, and extremely well organized
A self-confident, self-starter, and strategic thinker
Able to build strong, trusted relationships and operate in an influential manner with executives to assist in driving the strategic agenda
A strong facilitator, negotiator and mediator, with the ability to drive to consensus
Highly collaborative (comfortable partnering with all levels of the organization)
Experienced at seeing patterns and commonalities that can create a bigger picture out of small pieces, anticipating what's around the corner
Clear
Submit
Unique Skills in Job Description
deep learning, Python, Data Science, relational database, data analysis, analytics, databases, machine learning, design, Big data, sentiment analysis, software, data science, business, software engineering, R, languages, time series
Flag
Use via API
Built with Gradio


127.0.0.1
Job Description
Resumes
job_description
At least 1 year of experience with relational databases
Preferred Qualifications:
Master's Degree in "STEM" field (Science, Technology, Engineering, or Mathematics), or PhD in "STEM" field (Science, Technology, Engineering, or Mathematics)
At least 3 years' experience in Python, Scala, or R
At least 3 years' experience with machine learning
A successful candidate will be:
Possesses the Data Science trifecta of job specific skills: great data expertise, amazing data and software engineering skills, and terrific modeling skills
A source of positive energy, inspiration and motivation for financial services
Action oriented, always leaning in and moving with a sense of urgency
A multi-tasker, able to handle multiple priorities in parallel, and extremely well organized
A self-confident, self-starter, and strategic thinker
Able to build strong, trusted relationships and operate in an influential manner with executives to assist in driving the strategic agenda
A strong facilitator, negotiator and mediator, with the ability to drive to consensus
Highly collaborative (comfortable partnering with all levels of the organization)
Experienced at seeing patterns and commonalities that can create a bigger picture out of small pieces, anticipating what's around the corner
uploaded_files
John_Doe_Resume.pdf
4.3 KB
Clear
Submit
Ranked Resumes
Resume 1:
Names: ['John Doe']
Emails: ['john.doe@example.com']
Skill Match Score: 38.9%
Missing Skills
Resume 1 Missing Skills:
Python, design, Data Science, time series, software, Big data, relational database, R, databases, data analysis, sentiment analysis
Flag
Use via API
Built with Gradio

Findings

- **Performance Metrics:** Evaluated on precision (98%), recall (91%), and F1-score (95%)
- **Resume Parsing:** High accuracy in extracting key information (e.g., names, skills, experience), with slightly lower performance in context-dependent fields
- **Job Matching:** Good accuracy in ranking candidates, with semantic analysis providing a nuanced understanding of qualifications
- **Probable Impact:** Reduce candidate selection time by upto 50%, improves accuracy, and reduces human biases, which can lead to more efficient and fair recruitment

Model Improvements

Model	Precision	Recall	F1-Score
en_core_web_md (Base Model)	95%	84%	89%
en_core_web_md (Fine Tuned Model)	98%	91%	95%

Conclusion

System Purpose:

- Enhances recruitment by automating and optimizing resume processing
- Improves candidate-job matching using NLP techniques

Benefits:

- Addresses biases in traditional hiring
- Increases efficiency in the recruitment process

Challenges:

- Diverse resume formats.
- Industry-specific needs.
- Ethical concerns around AI fairness.

Future Work

- **Expand Training Data:** Incorporate more diverse resume formats and industry-specific data to enhance model accuracy
- **Adaptability:** Improve the system's handling of non-traditional resumes and niche job roles
- **Ethical Considerations:** Continuously monitor and address bias, ensuring fairness and transparency in AI-driven recruitment
- **User-Centric Enhancements:** Tailor resume suggestions more closely to specific job markets and employer preferences
- **Integrate LLM's:** To suggest Resume Enhacements automatically based on missing skills

THANK YOU