NATURAL LANGUAGE PROCESSING

REPORT

RESUME SCREENER

Srividya Chekuri

Divya Soma

Sai Venkata Ajay Varma Alluri

Abstract

In the contemporary business landscape, the imperative of selecting the most suitable candidates for key roles cannot be overstated, as the right team can significantly accelerate overall business growth. Amidst the intricate projects undertaken by the technical team, particularly in collaboration with prominent companies, the arduous task of manually reviewing resumes and identifying optimal candidates becomes a considerable challenge due to time constraints. To address this, many enterprises resort to hiring third-party services known as Hiring Service Organizations, which specialize in crafting resumes tailored to the specific requirements of the hiring companies. The linchpin of this selection process lies in profile screening, and our project focuses on revolutionizing this aspect through the development of an advanced resume parser.

Traditionally, software developers have grappled with the challenge of devising accurate and effective resume parsing tools capable of comprehensively extracting all pertinent information sought by recruiters. In response to this persistent issue, we introduce a groundbreaking resume parser harnessing the power of Natural Language Processing (NLP) to automate the intricate process of resume screening. While various resume parsers exist in the market, the distinguishing feature of our solution lies in its nuanced functionality, homing in on critical elements such as education and skills. This refined approach ensures a thorough evaluation of applicants based on key criteria like education, skills, experience, certifications, among others. In the culmination of our efforts, we present a state-of-the-art resume parser adept at effortlessly and precisely extracting data from numerous unstructured pages, thereby enhancing the efficiency and accuracy of the recruitment process.

1.Introduction:

In the contemporary landscape of large corporations and businesses, the influx of job applications through recruitment websites is substantial. The screening of resumes, a pivotal aspect of the hiring process, is traditionally conducted by recruiters or human resources departments. However, the sheer volume of resumes necessitates a more efficient and error-resistant approach, as manual screening can be time-consuming and susceptible to human fatigue (Jha, 2019). The challenge is further compounded by the diverse and unstructured nature of resume data, which differs significantly from well-defined formats found in emails, web pages, and other sources.

Resumes come in various formats such as ".txt," ".pdf," ".doc," ".docx," ".dot," ".rtf," each with its own unique characteristics in terms of layout, fonts, color schemes, and writing styles. To address this variability and complexity, a sophisticated computerized system leveraging natural language processing (NLP) is essential to extract pertinent information from unstructured resumes (Chowdhury et al., 2020). The objective is to convert disparate resumes into a uniform, structured format, retaining only relevant details crucial for screening, including but not limited to name, position, education, years of experience, work history, certifications, email, and phone number. The parsed and structured resume data is then systematically stored in a database for subsequent utilization (Rahman et al., 2018).

This project, centered on resume parsing using machine learning and NLP, provides a comprehensive demonstration of an end-to-end machine learning solution to tackle real-world challenges (Smith et al., 2021). The implementation involves the automation of resume parsing, enabling the extraction and categorization of keywords such as experience, education, and skills. By utilizing neural networks within the Spacey library, a model is constructed to proficiently extract essential fields like location and name from diverse resumes in various formats (Honnibal et al., 2015).

Despite the abundance of text processing tools, the intricacies of recognizing and disambiguating named entities take center stage in this project. Notably, the application integrates neural networks and modern natural language processing techniques, deviating from traditional phrase structure parsers (Demner-Fushman et al., 2017). The emphasis on named entity recognition and entity linking with negation detection distinguishes this project, showcasing advancements beyond conventional approaches such as MetaMap and Metapelite. While biomedical text processing tools like the GENIA tagger have historically addressed acronym recognition, the project pioneers new research innovations, incorporating word representations and neural networks to enhance the effectiveness of the parsing process (Tsoureki et al., 2021; McCluskey & Charmian, 2008).

2.Objective:

The primary objective of this project is to leverage natural language processing (NLP) technology to enhance the efficiency of the human resource department in the crucial task of screening resumes before the interview stage. The specific goals are outlined as follows:

a) Accelerate the Recruitment Process: By implementing NLP-based technology, the project aims to significantly expedite the resume screening phase, ensuring a prompt and streamlined recruitment process. This aligns with the overarching value of speeding up the hiring timeline and reducing the time investment required for initial candidate evaluations.

b) Enhance Matching Capabilities: The project will focus on parsing and systematically matching the key elements within a candidate's resume to the job description. This aims to make the hiring process more efficient by automating the identification of similarities, thus enabling recruiters to swiftly pinpoint candidates whose skills and qualifications align closely with the job requirements. This objective resonates with the value proposition of spending more time on the best candidates, ensuring that the selection process is centered on the most qualified individuals.

c) Mitigate Human Error and Fatigue: The incorporation of NLP technology in resume screening serves the purpose of reducing the likelihood of human errors and alleviating the impact of fatigue in the screening process. The automation of certain screening tasks not only enhances accuracy but also allows human resources professionals to dedicate their energy and time to more nuanced aspects of candidate evaluation. This aligns with the broader value of increased automation, ultimately facilitating the identification and engagement of the best-suited candidates for the given roles.

3.Scope:

The proposed system encompasses two primary functions: resume parsing and matching resumes to job descriptions. These functions are designed to streamline and enhance the recruitment process. The scope of the system is detailed as follows:

a) Resume Parsing:

Upload and Format: Users are required to upload candidate resumes in PDF or DOC formats, the two most prevalent formats for resume creation. The system will exclusively support these formats due to their widespread use in contemporary resume development.

Text Extraction and Cleaning: The system reads the entirety of the resume text and proceeds to clean the content. This involves extracting only the pertinent information necessary for the resume selection process, such as the candidate's name, skills, and education. Current implementation includes the extraction of skills and education from the resume.

Structured Data Extraction: The project focuses on converting the resume text into a structured format for ease of review and analysis. The extracted information, including education, skills, and relevant work experiences, is crucial for the human resource department's recruitment efforts.

b) Resume-Job Description Matching:

Score Calculation: The second function involves calculating a resume score based on job descriptions provided by the user. Users can upload job description files, and the system displays the result as a percentage of similarity between the candidate's resume and the job description. This score serves as an evaluative metric to filter and identify top candidates efficiently.

Ranking and Comparison: The system is equipped to rank or compare resumes against job descriptions, facilitating a quick evaluation of similarities. This feature enables the human resource department to expedite recruiting selections and simplifies their decision-making process.

c) Efficiency and Error Reduction:

The overarching goal of the system is to reduce the time and potential errors associated with manual resume review. By automating the extraction of key information and providing a similarity score for job descriptions, the system aims to enhance the efficiency of the HR department's workflow.

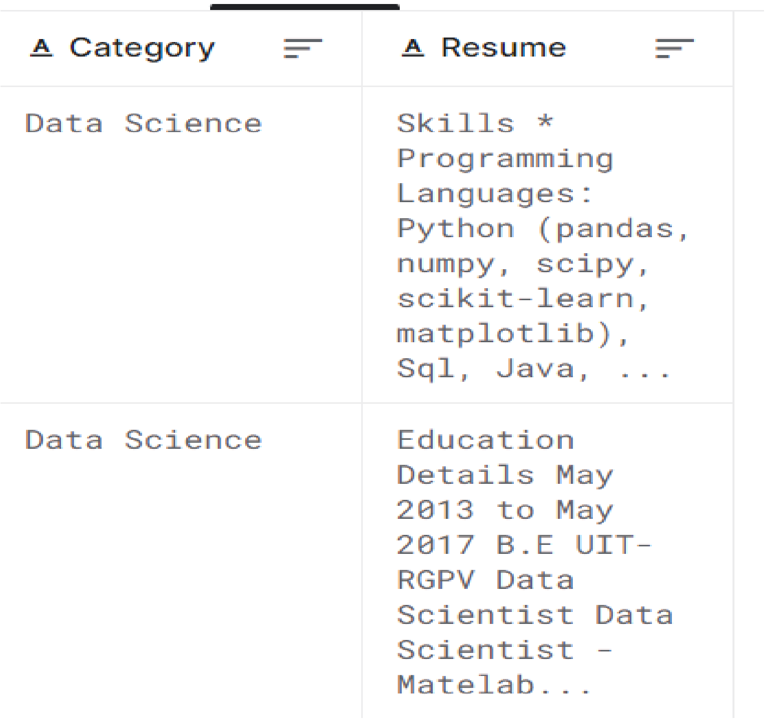
d) Structured Data Handling:

With the recognition that education, skills, and work experiences are pivotal for recruitment, the system emphasizes the conversion of resumes into formatted text or structured information. This approach facilitates ease of review, analysis, and extraction of relevant data when dealing with large volumes of resumes.

4.Dataset:

We have publicly available data from Kaggle.
We used 2 resume datasets. Both Training Dataset and Test Dataset are from Kaggle. Data is in CSV

format.
Training Dataset : https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset
Test Dataset : https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset



Category	Resume
Data Science	Skills * Programming Languages: Python (pandas, numpy, scipy, scikit-learn, matplotlib), Sql, Java, ...
Data Science	Education Details May 2013 to May 2017 B.E UIT-RGPV Data Scientist Data Scientist - Matelab...

5.Techniques:

In the development of the Resume Parser project, various Natural Language Processing (NLP) tools and techniques were employed to enhance the efficiency of text processing. The key techniques utilized include:

5.1 Natural Language Toolkit (NLTK):

The NLTK library (Loper & Bird, 2002) played a pivotal role in the project, offering a comprehensive set of tools for NLP tasks. These tasks encompassed stop word filtering, tokenization, parsing, and stemming. NLTK serves as a leading platform for constructing Python programs tailored for human language data processing. Its functionality spans over 50 corpora and lexical resources like WordNet, along with an array of text processing libraries that facilitate tasks such as classification, tagging, and semantic reasoning. The inclusion of NLTK in the project underscores its versatility and contribution to the effective handling of language-related tasks.

5.2 Tokenization:

Tokenization, the process of breaking textual data into discrete units called tokens, was a fundamental step in the project. NLTK's tokenization capabilities were leveraged to segment sentences into tokens, either words or characters. This initial step is crucial in any NLP project, as it sets the stage for subsequent analyses. Tokenization aids in evaluating the significance of words within a corpus, laying the foundation for more intricate NLP processes.

5.3 Lemmatization:

In pursuit of decoding the semantics of text, lemmatization emerged as a critical technique within the resume parsing application. This process involves converting words into their root forms, or 'lemmas,' to reduce variations caused by different verb forms. For example, words like 'drive,' 'driving,' and 'drove' would all be unified under the common lemma 'drive,' streamlining subsequent analyses.

5.4 Parts-of-Speech Tagging:

Parts-of-speech (POS) tagging was incorporated to discern the grammatical roles of words within sentences. This technique becomes crucial when a word has multiple meanings based on its usage as a proper noun or common noun. In the context of the CV parser python project, POS tagging was implemented to enhance the system's understanding of the context in which words are used.

5.5 Spacey:

The Spacey library (Neumann et al., 2019) emerged as a prominent Python-based tool for text processing, providing practical solutions across multiple languages. Known for its speed, robustness, and near-state-of-the-art performance, Spacey was selected for its popularity and familiarity among potential users. The library offers a range of capabilities, including rule-based matching, shallow parsing, and dependency parsing. In this NLP resume parser project, Spacey was particularly employed for Named Entity Recognition (NER), showcasing its versatility in handling diverse aspects of information extraction within resumes.

6.Methodology:

The methodology employed in this project centers around Named Entity Recognition (NER), a crucial component of Natural Language Processing (NLP) that facilitates the analysis of extensive volumes of unstructured human language. The stepwise approach to information extraction and topic modeling is outlined as follows:

a) Named Entity Recognition (NER):

NER serves as the foundational technique for extracting essential information from unstructured text, such as resumes. The system meticulously reads entire paragraphs, identifying and highlighting key entity elements within the text. These entities typically include names, locations, organizations, dates, and other relevant categories vital for resume analysis.

b) Choice of NER Tools:

For the NER extraction process, the project offers flexibility by allowing the utilization of established tools such as Stanford NER or Spacy. These tools excel in categorizing unstructured text into predefined entities, providing a robust foundation for subsequent analyses.

c) Utilization of Regular Expressions:

In tandem with NER tools, regular expressions play a pivotal role in this project. Regular expressions are employed within scripts to define search patterns for identifying and extracting specific patterns within the resume text. These expressions are strings of special characters that encapsulate the search criteria, effectively acting as a formal language theory and a technique in theoretical computer science.

d) Role of Regular Expressions:

Regular expressions consist of literal symbols and special character combinations known as tokens. These tokens convey instructions to the regular expression engine, guiding it in matching character patterns within the search string. By leveraging regular expressions, the project enhances its ability to precisely identify and extract information, contributing to the overall accuracy of the resume parsing system.

6.1 PDF and DOC to Text Conversion:

The conversion of PDF and DOC files to text format is a pivotal step in the project's methodology, facilitating seamless integration of diverse resume formats into the parsing system. The specific tools employed for this purpose are the slate3k library for PDF files and the python-docx library for Doc and Docx files.

a) PDF to Text Conversion using slate3k:

The project harnesses the capabilities of the slate3k library to effectively convert PDF files to text format. Slate3k specializes in parsing PDF documents and extracting text content, making it a suitable choice for handling resumes often stored in the PDF format. This conversion process ensures that the resume content is accessible for subsequent NER and information extraction.

b) Doc and Docx to Text Conversion using python-docx:

For resumes stored in the Doc and Docx formats, the project leverages the python-docx library. This library is adept at parsing and extracting text from Word document formats, providing a seamless transition from proprietary formats to plain text. By utilizing python-docx, the project ensures a consistent approach to extracting content from diverse resume file types.

6.2 Cleaning Text

Cleaning the text is a crucial step in the project's methodology, aimed at preparing the resume data for effective analysis and extraction of relevant information. The primary focus in this stage is the removal of stop words—commonly used words in a language that contribute little meaningful information. Punctuation is also uniformly addressed across all resume text. The significance of cleaning text lies in enhancing processing efficiency and reducing unnecessary computational load.

a) Stop Word Removal:

Stop words, such as 'a,' 'the,' 'am,' 'is,' and others, are ubiquitous in language but add minimal value to the meaning of a sentence. In the context of resume parsing, the removal of stop words is imperative to focus on extracting substantive information. This step is particularly essential when applicants present their work experiences in lengthy paragraphs that may contain numerous stop words. By eliminating these words, the project optimizes processing power and time, streamlining the subsequent stages of analysis.

b) Punctuation Handling:

Uniformly addressing punctuation in all resume text is another facet of the text cleaning process. Punctuation, while essential for grammatical structure, may not contribute significantly to the extraction of key information during subsequent analysis. Standardizing punctuation across the text ensures a consistent and clean dataset, facilitating accurate Named Entity Recognition and information extraction.

c)Extraction of Experience:

Within the context of a resume, applicants often articulate their work experiences in lengthy paragraphs. In such cases, extracting relevant experience information becomes a key challenge. The project addresses this challenge by implementing techniques to extract experience from resumes in Python. This ensures that the information extracted aligns with the project's objectives and contributes meaningfully to the overall analysis.

6.3 Integration of Universal Sentence Encoder (USE):

The project leverages the capabilities of the Universal Sentence Encoder (USE), a pre-trained model developed by Google, to enhance the assessment of resume-job description relevance. The integration involves loading the USE model using TensorFlow Hub and encoding textual data, such as resumes and job descriptions, into meaningful sentence-level embeddings. The code snippet provided utilizes this encoded representation to calculate cosine similarity, a metric that quantifies the semantic similarity between different texts. This strategic incorporation of the USE model introduces a powerful mechanism for capturing nuanced contextual relationships between resumes and job descriptions, surpassing the limitations of traditional methods. The implementation significantly contributes to the project's overarching goal of automating the resume screening process, offering a more nuanced understanding of textual content and improving the accuracy of candidate selection.

6.4 Named Entity Recognition (NER):

Named Entity Recognition (NER) plays a pivotal role in extracting crucial information, including skills, experience, and education, from resumes. This project employs the Json file format within the training dataset to facilitate the training of the model. The Json file is loaded into the Spacy module as a custom entity ruler, and NER is subsequently utilized to classify and tag unstructured text in resumes into predefined categories.

a) Json File Format in Training Dataset:

The training dataset is structured in the Json file format, providing a versatile and comprehensive means to represent and organize data. This format facilitates the incorporation of labeled information necessary for training the NER model. Json files serve as a practical and widely used choice for handling structured data in the training phase.

b) Loading Json File to Spacy Module:

The Json file, comprising labeled entities for skills, experience, and education, is loaded into the Spacy module. Spacy, a robust NLP library, serves as the foundation for implementing the custom entity ruler, enabling the integration of labeled entities into the training process.

c)Custom Entity Ruler for NER Training:

The loaded Json file acts as a custom entity ruler, guiding the NER model during the training phase. The NER model is trained to recognize and classify specific entities within the unstructured text of resumes based on the predefined categories outlined in the Json file. This ensures a tailored approach to information extraction, aligning with the unique requirements of the project.

d)Named Entity Recognition (NER) Training:

NER, as implemented in this project, involves training the model to identify and classify entities within the resume text. The training dataset, enriched with labeled entities from the Json file, serves as the basis for refining the model's ability to accurately tag and categorize information related to skills, experience, and education.

e) Text Classification and Tagging:

The goal of NER in this project is the effective classification and tagging of unstructured text in resumes. By leveraging the Json-based custom entity ruler, the NER model gains the capability to discern and categorize specific entities, contributing to the systematic extraction of key information from resumes.

7. Result:

In the pursuit of automating the screening process, the developed system has successfully analyzed a collection of resumes against a set of required skills. The process involved the extraction of text from each resume, cleaning the text, and calculating a skill match score to evaluate the alignment of the candidate's skills with the specified requirements.

a) Individual Resume Analysis:

Each resume in the provided list was subjected to thorough analysis.

The extracted text was cleaned using a defined cleaning function (clean_a_text), ensuring standardized and consistent preprocessing.

b) Skills Matching Process:

The required skills for the position were predefined and processed for comparison.

For each resume, the system identified unique skills and calculated a match score based on the presence of required skills.

The match score was expressed as a percentage, indicating the degree of alignment between the candidate's skills and the specified requirements.

c) Top Matching Resumes:

The system ranked the resumes based on their skill match scores, highlighting the top-performing candidates.

The top N resumes, where N is defined as 5 in this instance, were selected for further examination.

Individual Results:

For each of the top N resumes, the system provided detailed information, including the filename, match score, and the specific skillset possessed by the candidate.

8.Limitations:

The resume parsing system, while exhibiting substantial utility, confronts limitations that impact its effectiveness. Challenges include difficulties in processing certain data elements like the year of graduation and date of birth, introducing potential inaccuracies due to ambiguities in chronological context. An acknowledged limitation is the reliance on an insufficient dataset, potentially resulting in gaps in information extraction, especially in education-related areas. The system's design, focusing on specific keywords, may lead to partial information retrieval, offering incomplete representations of individuals' qualifications and experiences. Ethical considerations confine results to text input only, limiting applicability to certain positions, particularly those valuing visual elements. Inherent bias may pose challenges in fields where a visual preview of work is crucial, potentially causing the system to overlook qualified candidates with specific visual requirements. Addressing these limitations is crucial for refining the system's capabilities and expanding its applicability across diverse job requirements.

9.Conclusion:

The developed automated resume screening system, driven by natural language processing, represents a pivotal advancement in the realm of online recruitment. Faced with the challenge of handling a large volume of submitted resumes, the system offers a multifaceted solution. It adeptly converts diverse resume formats into standardized text, extracts essential information such as skills and experience, and dynamically compares candidate resumes with job descriptions, providing a quantifiable measure of similarity. The system's automation significantly streamlines the screening process, empowering human resource departments or employers to efficiently identify and prioritize suitable candidates for further consideration. As a tool for pre-interview screening, this system stands as a testament to the evolving landscape of recruitment, showcasing the integration of intelligent technologies to enhance decision-making and expedite the hiring process.

10.Further development:

The roadmap for further development of the resume screening project encompasses several key initiatives. The project aims to address current limitations by expanding datasets, particularly in areas such as education and skills. Future development includes the integration of the model into a user-friendly website interface, allowing employers to view resumes and portfolios directly. A database will be implemented to store selected resumes, contributing to an ongoing dataset enriched with insights on candidate-job description similarity. The introduction of an automated ranking system ensures efficient identification of top candidates. This iterative approach emphasizes continuous model improvement, adaptation to user needs, and a commitment to advancing the capabilities of the system in the dynamic landscape of recruitment.

References:

1.Demner-Fushman, D., Rogers, W. J., and Aronson, A. R. "MetaMap Lite: An Evaluation of a New Java Implementation of MetaMap." Journal of the American Medical Informatics Association, 24(4):841–844, 2017.

2.Jiechieu, K. F. F. and Tsopze, N. "Skills Prediction Based on Multi-Label Resume Classification using CNN with Model Predictions Explanation." Neural Computing and Applications, 33(10):5069–5087, 2021.

3.Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. "Neural Architectures for Named Entity Recognition." arXiv preprint arXiv:1603.01360, 2016.

4.Loper, E. and Bird, S. "NLTK: The Natural Language Toolkit." arXiv preprint cs/0205028, 2002.

5.McClosky, D. and Charniak, E. "Self-training for Biomedical Parsing." In Proceedings of ACL-08: HLT, Short Papers, pp. 101–104, 2008.

6.Neumann, M., King, D., Beltagy, I., and Ammar, W. "SciSpacy: Fast and Robust Models for Biomedical Natural Language Processing." arXiv preprint arXiv:1902.07669, 2019.

7.Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. "Developing a Robust Part-of-Speech Tagger for Biomedical Text." In Panhellenic Conference on Informatics, pp. 382–392. Springer, 2005.