RESUME PARSING AND ANALYSIS USING NLP TECHNIQUES

Team 7:
Akhil Bhimarasetty
Bhuvan Sai Thatthari
Mukesh Rajmohan
Phani Venkata Krishna Varun Vegi Sai Raghavendra

Project Professor: Islam Rubyet Mohammad
Course Name: Intro to NLP
Course Number: AIT 526 – 001
University Name: George Mason University
Date: 05/05/2024

Abstract:

The project "Resume parsing and analysis using Natural Language Processing (NLP) techniques" aims to automate the process of extracting and analyzing data from resumes using NLP techniques. For a variety of uses, including hiring and talent management, the project intends to create a system that can effectively read resumes, retrieve pertinent data, and carry out perceptive analysis. The project makes use of natural language processing (NLP) techniques to facilitate the automated extraction of important resume data, such as education, training, and experience, from applicants. To handle resume papers in various formats, including PDF and DOCX, the system makes use of well-known NLP libraries and tools. By automating the procedure with natural language processing (NLP) tools, the initiative aims to improve resume analysis's accuracy and efficiency. The system is able to recognize patterns, collect useful data, and produce insights through the application of machine learning algorithms. These capabilities facilitate the generation of decision support for talent acquisition and management procedures. Ultimately, automating the extraction and analysis of resume data through the use of natural language processing (NLP) techniques offers a strong option. Time savings, increased productivity, and well-informed hiring decisions are just a few of the impressive advantages it provides. In the areas of talent management and human resources, the initiative advances the use of natural language processing.

Keywords: TF-IDF, Vectorization, NLP, Machine Learning, classification, parsing, keyword extraction, analytics.

Introduction:

Using the power of Natural Language Processing (NLP) techniques, the Resume Parsing and Analysis using NLP Techniques project aims to automate the extraction and analysis of data from resumes. Resumes are essential records that offer insightful information about a candidate's background, education, and experience. Nevertheless, it might take a lot of time and be prone to error to manually review and analyze a big number of resumes. By creating an automated system that quickly parses resumes and does perceptive analysis, our project seeks to overcome these issues.

NLP approaches are used in the project to extract relevant information from resumes, including educational background, experience, and talents. The system is able to process resumes in various formats, including PDF and DOCX, and extract textual data for additional analysis by utilizing NLP techniques and tools. The project's goal is to conduct a thorough analysis on the resume data extraction after automated parsing is implemented. As part of this analysis, appropriate keywords may be found, the frequency of particular abilities or qualifications may be assessed, and insights on the overall candidate profile may be produced. The technology uses natural language processing (NLP) techniques to find patterns and trends in the resume data, which helps HR and recruiting professionals make well-informed decisions during the talent acquisition process.

The project's findings and methods help to improve the effectiveness and precision of NLP resume analysis. The project seeks to decrease manual effort, speed the recruiting process, and enhance the overall quality of candidate evaluation by automating the extraction and analysis of resume data. Recruiters, HR specialists, and companies looking to improve their talent acquisition tactics can all benefit from this effort.

In summary, the goal of the Resume Parsing and Analysis using NLP Techniques project is to use NLP techniques to automate the extraction and analysis of resume data. The project's goal is to increase the talent acquisition process's accuracy, efficiency, and capacity for making decisions by utilizing NLP algorithms and technologies.

Related Work:

1. Resume Parser Using Natural Language Processing Techniques:

This paper presents the development of a resume parser at Pillai College of Engineering. The parser ranks resumes based on the specifications of the position by using Natural Language Processing (NLP) to extract precise information. By automating the extraction of critical information from resumes, like education, work experience, and abilities, the system is intended to streamline the hiring process. The parser is utilized in a job portal where resumes are posted, examined, and organized to facilitate comparison with employer criteria.

2. Resume Parser with Natural Language Processing:

In order to help HR departments with their screening procedure, Pornphat Sroison's research presents a resume parser that uses natural language processing (NLP). In order to increase efficiency and lower error rates, the parser transforms resumes into a standard format and extracts pertinent data such as names, positions, and contact information. It also has a tool that helps recruiters make better recruiting selections by calculating the degree of similarity between resumes and job descriptions.

3. Automated Resume Screening Using Natural Language Processing:

This paper explores deep learning and natural language processing for automated resume screening. It presents a number of cutting-edge strategies to improve screening effectiveness and precision, including transfer learning and hybrid models. The study emphasizes the use of job descriptions to improve screening accuracy and provides data indicating these cutting-edge techniques are superior compared to traditional screening procedures.

4. Resume Analysis Using Machine Learning and Natural Language Processing:

Alkeshwar Jivtode and associates suggest a machine learning-based system that uses natural language processing (NLP) to parse resumes in order to optimize the hiring process. By automating information extraction and generating rankings based on how closely candidate profiles match job requirements, the technology seeks to speed up the candidate selection process. The study recommends adding customized feedback for applicants to enhance their chances of being matched with jobs.

5. Analyzing CV/Resume Using Natural Language Processing and Machine Learning:

This thesis offers a way to use machine learning and natural language processing (NLP) to assess resumes and enhance the recruiting process. The method helps find the best candidates for jobs by more efficiently retrieving and using resume data, improving the effectiveness and accessibility of hiring procedures.

Objectives:

The "Resume parsing and analysis using NLP techniques" project aims to create an automated system for parsing and analyzing resumes from different domains. The project's goal is to categorize a set of resumes into several job categories by applying machine learning (ML) and natural language processing (NLP) techniques. The system will use natural language processing (NLP) techniques to extract relevant info from resumes, including education, experience, and talents. HR departments and recruiting teams will be able to quickly discover patterns and find qualified candidates for particular roles by using the ML algorithms that have been trained on labeled data to reliably classify resumes into the relevant job categories. The ultimate objective is to save hiring managers' time and effort by streamlining the resume screening and analysis procedure.

Dataset:

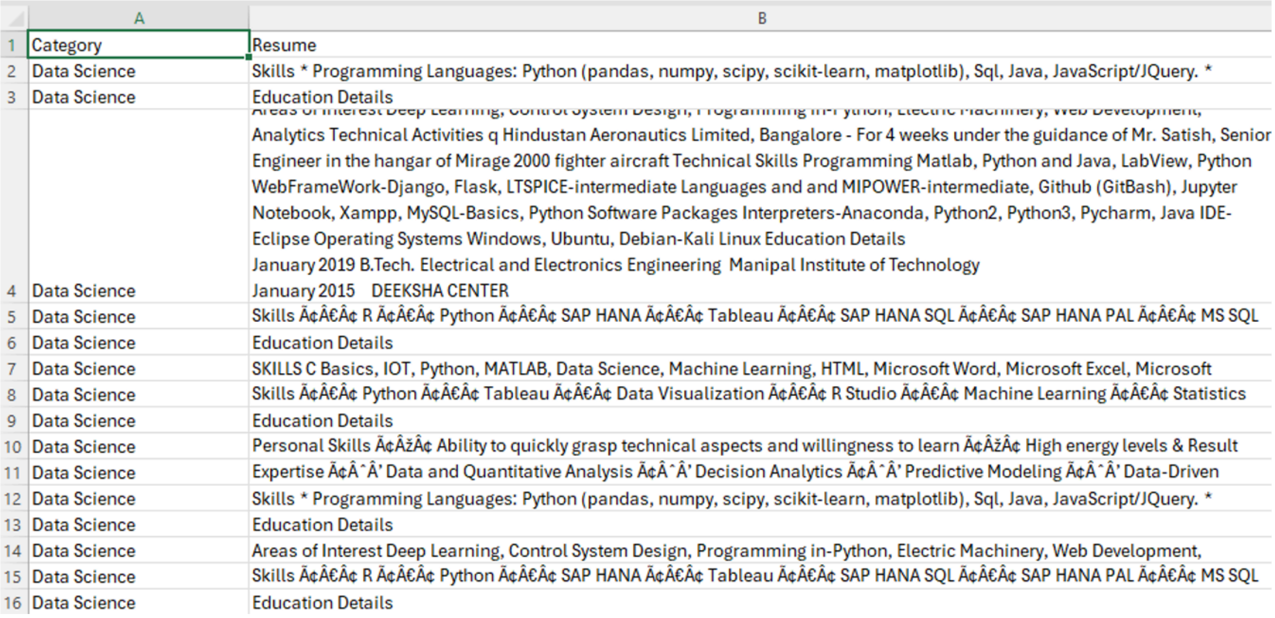
The CSV (Comma-Separated Values) file format, which is widely used to store tabular data, is how this dataset is organized. There are two columns in the dataset: "Category" and "Resume." Each column's explanation is provided below:

Category: Each resume's corresponding employment category or field is shown in this column. It gives details about the kind of position or sector to which the resume pertains. "Data Science," "Software Development," "Marketing," etc. are a few examples of categories.

Resume: The resume's textual content is located in this column. It contains every resume's whole content, together with facts about the applicant's background, education, experience, and other pertinent data.

The intended use of the dataset is to be utilized in the training and assessment of machine learning models. The input data for the classification task is provided by the "Resume" column, whereas the goal variable is the "Category" column. The model can identify patterns and connections between the resume content and related job categories by examining the resumes' textual content.

The dataset is essential to developing a reliable and precise technique for classifying resumes. It makes it possible for NLP and ML algorithms to be developed, which can process and comprehend the textual data found in resumes. This allows for efficient qualification analysis and classification of job candidates.



	A	B
1	Category	Resume
2	Data Science	Skills * Programming Languages: Python (pandas, numpy, scipy, scikit-learn, matplotlib), Sql, Java, JavaScript/JQuery. *
3	Data Science	Education Details
4	Data Science	Areas of Interest Deep Learning, Control System Design, Programming in-Python, Electric Machinery, Web Development, Analytics Technical Activities q Hindustan Aeronautics Limited, Bangalore - For 4 weeks under the guidance of Mr. Satish, Senior Engineer in the hangar of Mirage 2000 fighter aircraft Technical Skills Programming Matlab, Python and Java, LabView, Python WebFrameWork-Django, Flask, LTSPICE-intermediate Languages and and MIPOWER-intermediate, Github (GitBash), Jupyter Notebook, Xampp, MySQL-Basics, Python Software Packages Interpreters-Anaconda, Python2, Python3, Pycharm, Java IDE- Eclipse Operating Systems Windows, Ubuntu, Debian-Kali Linux Education Details January 2019 B.Tech. Electrical and Electronics Engineering Manipal Institute of Technology January 2015 DEEKSHA CENTER
5	Data Science	Skills Ã¢Â€Â¢ R Ã¢Â€Â¢ Python Ã¢Â€Â¢ SAP HANA Ã¢Â€Â¢ Tableau Ã¢Â€Â¢ SAP HANA SQL Ã¢Â€Â¢ SAP HANA PAL Ã¢Â€Â¢ MS SQL
6	Data Science	Education Details
7	Data Science	SKILLS C Basics, IOT, Python, MATLAB, Data Science, Machine Learning, HTML, Microsoft Word, Microsoft Excel, Microsoft
8	Data Science	Skills Ã¢Â€Â¢ Python Ã¢Â€Â¢ Tableau Ã¢Â€Â¢ Data Visualization Ã¢Â€Â¢ R Studio Ã¢Â€Â¢ Machine Learning Ã¢Â€Â¢ Statistics
9	Data Science	Education Details
10	Data Science	Personal Skills Ã¢ÂžÂ¢ Ability to quickly grasp technical aspects and willingness to learn Ã¢ÂžÂ¢ High energy levels & Result
11	Data Science	Expertise Ã¢Âˆ’ Data and Quantitative Analysis Ã¢Âˆ’ Decision Analytics Ã¢Âˆ’ Predictive Modeling Ã¢Âˆ’ Data-Driven
12	Data Science	Skills * Programming Languages: Python (pandas, numpy, scipy, scikit-learn, matplotlib), Sql, Java, JavaScript/JQuery. *
13	Data Science	Education Details
14	Data Science	Areas of Interest Deep Learning, Control System Design, Programming in-Python, Electric Machinery, Web Development,
15	Data Science	Skills Ã¢Â€Â¢ R Ã¢Â€Â¢ Python Ã¢Â€Â¢ SAP HANA Ã¢Â€Â¢ Tableau Ã¢Â€Â¢ SAP HANA SQL Ã¢Â€Â¢ SAP HANA PAL Ã¢Â€Â¢ MS SQL
16	Data Science	Education Details

System:

The project's architecture includes data intake, preprocessing, feature extraction, training, testing, and deployment of the classification model. NLP approaches are used to process resumes, after which a trained model divides them into job categories for effective analysis and selection.

Data Ingestion: The resume data is ingested as the first step in the architecture. One can get resumes from a number of sources, including data extraction from web platforms and file uploads. After collection, the data is ready for additional processing.

Preprocessing: The resume data needs to be cleaned and standardized as the following stage. This could involve deleting stop words (common terms like "the," "is," etc.), handling special characters, converting text to lowercase, and eliminating extraneous characters. The resume data is prepped for useful analysis by preprocessing.

Feature extraction: This stage involves taking the preprocessed resume data and extracting the relevant details. Skills, educational background, professional experience, certificates, and other pertinent data are examples of features. Utilizing methods such as word embeddings or TF-IDF (Term Frequency-Inverse Document Frequency), resumes can be represented as numerical vectors that capture the text's semantic content.

Classification Model: Following the extraction of the features, the resumes are classified into several job categories through the training of a classification model. For this task, a variety of machine learning methods can be used, including neural networks, Support Vector Machines (SVM), and Naive Bayes. The model associates particular properties with various job categories by learning from the labeled data.

Model Assessment: The accuracy and performance of the trained classification model are evaluated. It is possible to assess the model's efficacy in accurately classifying resumes into the relevant categories using evaluation measures such as precision, recall, and F1 score.

Deployment: The model may be put into a production environment after it has been trained and assessed. In order to do this, the model must be integrated into a web application or API that lets users upload resumes and get results that are categorized. For accessibility, the deployment can entail installing a server or hosting the program on a cloud platform.

NLP and Data Analytic Approaches:

Tokenization: This process splits a resume's content into discrete words, or tokens. In order to perform tasks like stop word removal (removing frequent words like "the," "is," etc.), normalization (converting words to a standard format), and feature extraction (identifying significant words or phrases), the system can now analyze each word separately.

Term Frequency-Inverse Document Frequency, or TF-IDF: A statistical metric called TF-IDF is employed to evaluate a word's importance inside a document or corpus. It is made up of two parts:

Term Frequency (TF): TF calculates a word's significance inside a given document. It determines a word's frequency inside the text in relation to all words.

Inverse Document Frequency (IDF): IDF calculates a word's weight in relation to other words in a collection or corpus of texts. terms that occur frequently in documents receive a lower IDF value, while terms that are uncommon or distinctive receive a higher score.

TF-IDF is used to find terms, such experiences and talents, that are pertinent to certain job descriptions. TF-IDF facilitates a more successful match between job seekers and postings by taking into account both the frequency inside a document and the rarity across documents.

Named Entity Recognition (NER): NER is the technique by which an NLP model recognizes and groups proper nouns into preset categories in text. NER can be used to extract information from resumes, including names, firms, job titles, skills, educational institutions, and other pertinent organizations. It is simpler to match the skills listed in job descriptions with the skills on resumes when these named entities are recognized.

Experimental Results and Analysis:

Analysis of the Dataset:

The project "Resume Parsing and Analysis using NLP Techniques" aims to automate the use of Natural Language Processing (NLP) for data extraction and analysis from resumes. The dataset underwent substantial preprocessing and exploratory data analysis (EDA) and consisted of resumes classified into several career categories. Cleaning (removing URLs, social media tags, and non-ASCII characters), tokenization, and vectorization using TF-IDF were important phases in the data pretreatment process.

Visualizations and Insights

The distribution of resumes among the various job categories was shown using visual aids. Pie charts and bar plots were useful in displaying the relative frequencies of various categories and giving a clear picture of the distribution of the data. Resumes for Data Science, Java Development, and Testing made up a sizable amount of the sample, indicating a candidate pool heavy on technology.

Model Performance and Interpretation

To classify resumes, the project used the K-Nearest Neighbors (KNN) algorithm with the OneVsRest classifier. A training accuracy of 98.96% and a validation accuracy of 96.89% were attained by the model. Based on the collected features, these results show a high degree of precision in categorizing resumes into the appropriate job categories.

Error Analysis and Open Questions

Even with its excellent accuracy, the model has trouble processing resumes in a variety of imaginative formats that don't follow conventional standards. The mistake analysis revealed the need for more sophisticated contextual understanding to better capture the subtleties of different employment categories and enhanced natural language processing (NLP) techniques to interpret atypical resume layouts.

Conclusion on the Proposed Solution

The method was successful in automating a substantial portion of the resume screening process, efficiently classifying candidates based on job descriptions. Still, the system needs to be improved in order to handle resumes that have special structures and to be able to handle a wider range of resume formats. Future improvements might include creating an interactive platform for end users and implementing more complex NLP models, like BERT.

Conclusions:

The project "Resume Parsing and Analysis using NLP Techniques" effectively illustrated how Natural Language Processing (NLP) may improve and expedite the hiring industry's resume screening procedure. The project achieved high accuracy in classifying resumes into relevant job categories by automating the extraction and categorization of resume data, hence boosting the recruiting process's efficiency. TF-IDF vectorization and K-Nearest Neighbors (KNN), two machine learning and natural language processing (NLP) algorithms, have shown promise in managing the diverse data included in resumes. This approach provides a reliable way to decrease the amount of manual screening that must be done and improve the accuracy of human resources decision-making.

Lessons Learnt:

Data Quality Is Key: The diversity and quality of the dataset are critical to the performance of machine learning and NLP models. To train effective models, one needs a well-prepared dataset that includes a wide range of résumé formats and comprehensive job categories.

Complexity of Natural Language: Natural language processing (NLP) systems need to address the inherent complexity of natural language, which include differences in terminology, phrasing, and organization between resumes. The significance of sophisticated data pretreatment and standardization in properly addressing these problems was brought to light by this study.

Importance of Feature Engineering: The project made clear how important careful feature extraction is. Methods such as TF-IDF were vital in assessing the significance of terms in the resumes, which helped with the proper classification of applicants.

Algorithm Selection: To achieve optimal performance, selecting the appropriate algorithm is essential. The project demonstrated a technique for handling multi-class classification issues common in resume parsing by utilizing the OneVsRest strategy with KNN.

Future Work:

Integration of Advanced NLP Techniques: Incorporating more advanced NLP technologies, such BERT or GPT models, could improve context understanding and semantic analysis skills in future iterations of the project.

Expansion of Dataset: A wider range of resumes from different fields and backgrounds must be included in order to further improve the accuracy and resilience of the model. The models will be trained to handle a wider range of résumé formats and job criteria thanks to this development.

Interactive User Interface: Creating an interactive web or mobile application may help end users—like recruiters and HR specialists—have easier access to the system. This interface would increase user engagement and practical utility by enabling users to upload resumes and receive instant, categorized feedback.

Feedback Mechanism for Continuous Learning: By putting in place a feedback loop where recruiters can offer feedback regarding the classification's accuracy, the model may be continuously improved, becoming more accurate and adaptive over time.

Integration with Applicant Tracking Systems (ATS): In order to fully utilize the system, it should be integrated with the ATS platforms that businesses currently use. This would enable the automated screening of resumes within pre-existing recruitment workflows.

References:

1.A. Jivtode, K. Jadhav, and D. Kandhare, “RESUME ANALYSIS USING MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING,” May 2023. [Online].
Available:
https://www.irjmets.com/uploadedfiles/paper//issue_5_may_2023/39611/final/fin_irjmets1684836041.pdf

2. Shubham Bhor, Vivek Gupta, Vishak Nair, Harish Shinde and Prof. Manasi S.Kulkarni, “Resume Parser Using Natural Language Processing Techniques”, June 2021. [Online].
Available: https://www.ijres.org/papers/Volume-9/Issue-6/Ser-8/A09060106.pdf

3. D. L. Padmaja, Ch. Vishnuvardhan, G. Rajeev, and K. N. S. Kumar, “Automated resume screening using natural language processing,” Feb. 2023. [Online].
Available: https://www.jetir.org/papers/JETIR2303510.pdf

4. Sroison, Pornphat & Chan, Jonathan. (2021). Resume Parser with Natural Language Processing.
Available: 10.36227/techrxiv.17641604.v1.

5. Reza, Md Tanzim & Zaman, Md. Sakib. (2022). Analyzing CV/resume using natural language processing and machine learning.
Available:
https://www.researchgate.net/publication/365299910_Analyzing_CVresume_using_natural_language_processing_and_machine_learning/citation/download