CSE3024- Web Mining

Digital Assignment - I

CURRICULUM VITAE PARSER

By

20BCE1808 Arunima Agarwal
20BCE1832 Harini S

B.Tech CSE

Submitted to

Dr.A.Bhuvaneswari,
Assistant Professor Senior,
SCOPE, VIT, Chennai

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

December 2023

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computing Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

FALL SEM 23-24

Worklet details

Programme	B.Tech CSE	
Course Name / Code	Web Mining – CSE3024	
Slot	C1+TC1	
Faculty Name	Dr. A. Bhuvaneswari	
J Component Title	Curriculum Vitae Parser	
Team Members Name | Reg. No	Arunima Agarwal	20BCE1808
	Harini S	20BCE1832

Team Members(s) Contributions – Tentatively planned for implementation:

Worklet Tasks	*Contributor's Names*
Database connection and integration using Python	Arunima
Preprocessing	Harini
Model building	Arunima
Visualization	Harini
Technical Report writing	Arunima, Harini
Presentation preparation	Arunima, Harini

ABSTRACT

Due to the increase in population, the overall growth of a country reduces. To stabilize this growth, increasing the employment rate is one of the major solutions. With many candidates applying for jobs, recruiters spend a lot of time manually reviewing resumes or curriculum vitae. Searching through the resumes of thousands of job applicants has become a challenging task. To overcome this difficulty, recent research has focused on building machine-learning models that can be adapted to different resume styles and formats. This work proposes a model that can create resumes for each candidate based on the information they provide, helping recruiters intelligently sift through thousands of resumes so they can select the right candidate for a job interview using the resume extraction technique. Resume extracting or parsing is used to find structured information from unstructured data in the resume, to facilitate its storage, analysis, and management. It is a very useful tool for recruiters, and human resource professionals, as it allows them to quickly and efficiently identify key information about job candidates and make it easier to shortlist the candidates.

Keywords—Curriculum Vitae, Machine Learning, Python

1. Introduction

a. Problem Statement

The current landscape of job searching is characterized by information overload and inefficiencies, with job seekers struggling to find relevant opportunities and employers grappling with an inundation of unqualified applicants. To address these challenges, there is a pressing need for an Automated Recommendation System for Job Search Analysis that leverages Stochastic Gradient Descent (SGD) models and enhances the keyword search platform. The primary problem to be solved is the optimization of job matching and recommendation processes to facilitate a seamless and personalized job search experience for both job seekers and employers.

b. Objectives

User Personalization:
To create a recommendation system that tailors job suggestions to individual users based on their skills, experience, preferences, and historical job search behavior.

Enhanced Keyword Search:
To improve the accuracy and relevance of keyword-based job search results by incorporating advanced semantic understanding and contextual information.

Reduced Information Overload:
To help job seekers by reducing information overload, ensuring they receive a manageable number of highly relevant job recommendations.

Improved Candidate Screening for Employers:
To assist employers in finding qualified candidates by providing them with a refined pool of job applicants through more accurate matching.

c. Challenges Faced

Developing an Automated Recommendation System for Job Search Analysis poses several challenges. Personalization requires extensive user data, raising privacy concerns. Enhancing keyword search necessitates complex natural language processing. Optimizing the SGD model demands computational resources and expertise. Scaling the infrastructure for large datasets is resource-intensive. Balancing reduced information overload with relevant recommendations requires fine-tuning algorithms. Ensuring data security and complying with regulations adds complexity. Integrating with existing platforms requires seamless compatibility. Continuous improvement relies on feedback and iterative development. User training and support are essential for effective utilization. Managing costs while delivering value presents a financial challenge. Lastly, market adoption necessitates marketing efforts and overcoming user resistance to change.

2. Literature Survey

Sl. No	Title	Author(s) / Journal Name / Year	Technique	Result
1.	"Personalized Job Recommendation System"	Yu et al. / Expert Systems / 2017	Collaborative Filtering	Improved job recommendations based on user preferences and behavior.
2.	"Semantic Matching for Job Recruitment: A Dataset and Two Baselines"	S. Moosavi, et al. / arXiv / 2019	Semantic Matching (Word Embeddings)	Introduces a dataset and baseline models for semantic matching in job recruitment.
3.	"Context-Aware Event Recommendation in Event-based Social Networks"	M. Macedo, et al. / ACM Transactions on Social Computing / 2016	Context-Aware Recommendation	Explores context-aware recommendation techniques that can be applied to job recommendation systems.
4.	"Improving Content-Based and Hybrid Music Recommendation Using Deep Learning"	Lee et al. / Expert Systems with Applications / 2019	Deep Learning (CNN)	Enhanced music recommendation through deep learning techniques.
5.	"A survey of content-based recommendation systems in e-commerce"	Sharma et al. / Journal of Computer Science and Technology / 2020	Content-Based Filtering	A survey of content-based recommendation systems in e-commerce, relevant to content-based enhancements.
6.	"Stochastic Gradient Descent Tricks"	Bottou et al. / Neural Networks:	Stochastic Gradient Descent	Discussion of various technique

		Tricks of the Trade / 2012		improving SGD optimization.
7.	"A Review on Job Recommendation Algorithms"	V. Aggarwal, et al. / International Journal of Computer Applications / 2014	Collaborative Filtering, Content-Based Filtering	Offers a comprehensive review of various job recommendation algorithms.
8.	"Enhancing Job Recommendations by Analyzing User Click Behavior"	Y. Zhang, et al. / IEEE International Conference on Data Mining (ICDM) / 2013	Click Behavior Analysis	Discusses how user click behavior analysis can improve job recommendations.
9.	"Keyword-based Recommendations in Job Markets"	J. Lee, et al. / International Conference on Data Engineering (ICDE) / 2013	Keyword-Based Recommendation	Explores keyword-based recommendation approaches in job markets.
10.	"Improving Job Recommendation in E-recruitment with Multiple Rankers"	A. Brzezinski, et al. / Expert Systems with Applications / 2018	Multiple Rankers Ensemble	Presents a method for improving job recommendations using multiple rankers.

3. Dataset and Tool to be used (Details)

CV parser project involves training a machine learning model to automatically extract relevant information from resumes or CVs in various formats, such as PDF, DOC, DOCX, and TXT. To do this, a dataset of resumes is needed, which can be obtained from various sources, such as online job boards, company websites, and public resume databases. The dataset is preprocessed and annotated to identify the relevant information and labels, such as job titles, company names, dates, and skills. Preprocessing the dataset involves cleaning and standardizing the resumes to remove irrelevant or inconsistent information, such as headers, footers, and logos. This can be done using text extraction and cleaning tools, such as PDFMiner, Tesseract, or regular expressions. Annotation of the dataset involves labeling the relevant information and fields in the resumes to provide a training signal for the machine learning model. This can be done manually by human annotators or through automated tools, such as named entity recognition (NER) algorithms, which can identify and extract entities such as names, dates, and locations. Once the dataset is preprocessed and annotated, it can be used to train a machine learning

model, such as a neural network or a rule-based system, to automatically extract relevant information from new resumes. The performance of the model can be evaluated using metrics such as precision, recall, and F1 score, and refined through iterative training and validation.

4. Algorithms / Techniques description

This project made use of several algorithms.

1) **Rule-based parsing:** This algorithm extracts information from resumes using a collection of rules or patterns based on particular keywords or syntax. Patterns such as "Work Experience," "Education," and "Skills" will be searched for, to extract relevant data.

2) **Named Entity Recognition (NER):** This algorithm employs machine learning techniques to recognize and extract named entities from text, such as people, dates, locations, and organizations. NER can be used to extract information from resumes such as job titles, business names, and dates.

3) **Natural Language Processing (NLP):** This algorithm analyses and understands human language by employing statistical and computational models. Based on semantic and syntactic patterns, NLP techniques can be used to recognize and extract pertinent information from resumes.

4) **Machine Learning (ML):** This algorithm learns trends and relationships in data and makes predictions based on new data using statistical and computational models. Based on annotated data, ML techniques can be used to train a model to automatically extract relevant data from resumes. Stochastic Gradient Descent (SGD) was one prominent ML technique used for determining the learning rate of the model.

5. Github Repository Link (where your j comp project work can be seen for assessment)

Github link

REFERENCES

[1] Dav 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, 1265- 1274, DOI: 10.1080/09720510.2020.1799583.

[2] Gerard Deepak, Varun Teja & A. Santhanavijayan (2020) A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm, Journal of Discrete Mathematical Sciences and Cryptography, 23:1, 157-165, DOI: 10.1080/09720529.2020.1721879.

[3] Nirali Bhaliya, Jay Gandhi, Dheeraj Kumar Singh (2020) NLP based Extraction of Relevant Resume using Machine Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9:7, 2278-3075, DOI: 10.35940/ijitee.F4078.059720.

[4] Tejaswini K, Umadevi V, Shashank M Kadiwal, Sanjay Revanna, Design and development of machine learning based resume ranking system, Global Transitions Proceedings, Volume 3, Issue 2, 2022, Pages 371-375, ISSN 2666-285X, https://doi.org/10.1016/j.gltp.2021.10.002.

[5] Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, Rocky Bhatia, A Machine Learning approach for automation of Resume Recommendation system, Procedia Computer Science, Volume 167, 2020, Pages 2318-2327, ISSN 1877-0509, [https://doi.org/10.1016/j.procs.2020.03.284.]

[6] Agnieszka Wosiak, Automated extraction of information from Polish resume documents in the IT recruitment process, Procedia Computer Science, Volume 192, 2021, Pages 2432-2439, ISSN 1877-0509,[https://doi.org/10.1016/j.procs.2021.09.012].

[7] Shubham Bhor, Vivek Gupta, Vishak Nair, Harish Shinde, Prof. Manasi S.Kulkarni (2021), Resume Parser Using Natural Language Processing Techniques, International Journal of Research in Engineering and Science (IJRES), 9:6, 2320-9356, [https://www.ijres.org/papers/Volume-9/Issue-6/Ser8/A09060106.pdf]

[8] Vedant Bhatia, Prateek Rawat, Ajit Kumar, Rajiv Ratn Shah (2019), End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT, arXiv:1910.03089, [https://doi.org/10.48550/arXiv.1910.03089]

[9] D. Vukadin, A. S. Kurdija, G. Delač and M. Šilić, "Information Extraction From Free-Form CV Documents in Multiple Languages," in IEEE Access, vol. 9, pp. 84559-84575, 2021, DOI: 10.1109/ACCESS.2021.308791.

[10] M. F. Mridha, R. Basri, M. M. Monowar and M. A. Hamid, "A Machine Learning Approach for Screening Individual's Job Profile Using Convolutional Neural Network," 2021 International Conference on Science & Contemporary Technologies (ICSCT), Dhaka, Bangladesh, 2021, pp. 1-6, doi: 10.1109/ICSCT53883.2021.9642652.

[11] Priyavrat and N. Sharma, "Sentiment Analysis using tidytext package in R," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 577-580, doi: 10.1109/ICSCCC.2018.8703296