Project Report
Resume Analysis

Machine Learning

Group Members:

Name: Huzaifa Imran
Sap Id: 12008

Name: Ahmad Ali
Sap Id: 24758

Name: Haider Ali
Sap Id: 26030

Subject: **Machine Learning**

BSCS

CS 7-1

Project Report submission guidelines

Due: 20-12-2023 11:59PM

Submit your report of minimum 5 pages, it should include :

1.Introduction

- Briefly introduce the project's purpose and significance

2. Requirements:

- List and describe functional and non-functional requirements.

3. Technologies Used:

- Detail technologies, tools, and frameworks chosen.
- Explain their relevance to achieving project goals

4. Methodology:

- Outline the development approach and methodology.

5. Results:

- Present project outcomes with visual aids.

6. Future Work:

- Discuss potential enhancements or improvements.
- Summarize main findings and contributions.

8. References:

- Include a concise list of all sources and citations.

1.Introduction

- Briefly introduce the project's purpose and significance

Welcome to the Resume Analysis Program! In today's dynamic job market, understanding the key skills and trends within different job categories is essential for both job seekers and employers. This program aims to provide valuable insights into the distribution of skills across various job categories, allowing users to explore and analyze resumes to make informed decisions.

Purpose and Significance:

The purpose of this program is to facilitate the analysis of resumes, specifically focusing on the identification and distribution of skills within different job categories. By leveraging natural language processing (NLP) techniques, the program extracts relevant information from resumes and presents it in a visually appealing manner. This analysis can help job seekers understand the prevalent skills in their desired field and enable employers to identify top skills within specific job categories.

Key Features:

1. Data Loading and Cleaning: The program loads resume data from a CSV file, randomly shuffles it, and performs data cleaning to prepare it for analysis.

2. Named Entity Recognition (NER): Using the spaCy library, the program identifies skills mentioned in the resumes through a pre-defined entity recognition model.

3. Data Visualization: The program employs various visualization techniques, including histograms and word clouds, to showcase the distribution of skills and job categories.

4. User Interaction: Users can input a specific job category to analyze, and the program dynamically generates visualizations based on the selected category.

5. User Resume Analysis: Job seekers can input their own resumes and a list of desired skills. The program evaluates the match percentage between the user's skills and those present in the resume, providing valuable feedback.

6. Topic Modeling: The program utilizes Latent Dirichlet Allocation (LDA) for topic modeling, uncovering hidden topics within the resumes and visualizing them using pyLDAvis.

By combining these features, this program offers a comprehensive tool for both job seekers and employers to gain insights into the skills landscape of various job categories. Whether you are refining your resume or recruiting talent, this program aims to enhance your decision-making process.

2. Requirements:

- List and describe functional and non-functional requirements.

Functional Requirements:

1. Data Loading and Cleaning:

- The system should be able to load resume data from a specified CSV file.
- Data loading should include random shuffling for diversity in analysis.
- Cleaning processes should remove irrelevant characters, URLs, and perform lemmatization.

2. Named Entity Recognition (NER):

- The program must utilize spaCy for entity recognition to identify skills in resumes.
- An entity ruler should be employed to recognize skills based on pre-defined patterns.
- The system should distinguish and categorize entities such as skills, organizations, and job categories.

3. Data Visualization:

- The program should generate interactive visualizations, such as histograms and word clouds, to represent the distribution of skills and job categories.
- Visualizations should be customizable based on user input, allowing dynamic exploration.

4. User Interaction:

- Users should be able to input a specific job category for analysis.
- The system must dynamically generate visualizations based on the selected job category.
- User input for their own resumes and desired skills should be supported for analysis.

5. User Resume Analysis:

- The program should assess the match percentage between user-input skills and those present in the provided resume.
- Feedback on the resume's match to the user's requirements should be displayed.

6. Topic Modeling:

- The system should implement Latent Dirichlet Allocation (LDA) for topic modeling.
- Topics discovered through LDA should be visualized using pyLDAvis for user understanding.

Non-Functional Requirements:

1. Performance:

- The system should handle the analysis of a dataset of at least 200 resumes efficiently.
- Response times for user interactions and visualizations should be within acceptable limits.

2. Scalability:

- The program should be designed to handle larger datasets if future scalability is required.
- Scalability considerations should be taken into account for data loading and processing.

3. Usability:

- The user interface should be intuitive and user-friendly, guiding users through the analysis process.
- Clear instructions and prompts should be provided to assist users in making inputs.

4. Security:

- Any user-provided data, such as resumes and skills, should be handled securely, with measures to prevent unauthorized access or data breaches.
- Access controls should be implemented to restrict sensitive operations.

5. Compatibility:

- The program should be compatible with different operating systems (Windows, Linux, macOS).
- Compatibility with popular web browsers for visualization rendering should be ensured.

6. Reliability:

- The system should be robust and able to handle unexpected inputs or errors gracefully.
- Error messages and logging should be implemented for troubleshooting and debugging purposes.

7. Documentation:

- Comprehensive documentation, including installation instructions, user guides, and code comments, should be provided for users and developers.

3. Technologies Used:

- Detail technologies, tools, and frameworks chosen.

- Explain their relevance to achieving project goals

The Resume Analysis Program utilizes a combination of Python libraries, frameworks, and tools to achieve its project goals. Here is a detailed list of the technologies used and their relevance to the project:

1. Python:

- Relevance: Python is a versatile programming language known for its simplicity and readability. It provides a rich ecosystem of libraries and tools, making it well-suited for natural language processing, data analysis, and visualization.

2. Pandas (1.5.3):

- Relevance: Pandas is a powerful data manipulation and analysis library in Python. It is used for loading and handling structured data, such as resumes in CSV format, facilitating efficient data cleaning and manipulation.

3. FastAPI, Kaleido, Python-Multipart, Uvicorn:

- Relevance: FastAPI is a modern, fast (high-performance), web framework for building APIs with Python 3.7+. It is used to create a web API for user interaction. Uvicorn serves as the ASGI server, and Kaleido and Python-Multipart assist with additional functionalities and image rendering.

4. Spacy:

- Relevance: SpaCy is a natural language processing library that provides efficient tools for named entity recognition (NER) and linguistic analysis. It is used for identifying skills in resumes, categorizing entities, and performing entity recognition tasks.

5. Gensim:

- Relevance: Gensim is a library for topic modeling and document similarity analysis. It is used for Latent Dirichlet Allocation (LDA) to discover hidden topics within resumes, providing insights into the underlying themes.

6. PyLDAvis:

- Relevance: PyLDAvis is a Python library for interactive topic model visualization. It is used to generate visualizations of the topics discovered through LDA, making it easier for users to interpret and understand the results.

7. WordCloud:

- Relevance: WordCloud is a library for creating word clouds. It is used to visualize the most frequently occurring words in resumes, offering a graphical representation of the most used terms.

8. Plotly Express, Matplotlib:

- Relevance: Plotly Express and Matplotlib are visualization libraries used for creating interactive histograms and other visualizations. They enhance the presentation of data distribution and patterns within the program.

9. NLTK (Natural Language Toolkit):

- Relevance: NLTK is a natural language processing library used for tasks such as lemmatization. It aids in preprocessing textual data by removing stopwords and lemmatizing words for cleaner analysis.

10. JSONLines:

- Relevance: JSONLines is used for working with JSON data in a streaming fashion. It is relevant for loading pre-defined entity patterns for spaCy's entity ruler.

11. Warnings:

- Relevance: The warnings module is used to suppress unnecessary warnings during program execution, ensuring a cleaner and more focused user experience.

These technologies collectively enable the program to handle data manipulation, natural language processing, web API development, and interactive visualization, contributing to the overall success of the Resume Analysis Program.

4. Methodology:

- Outline the development approach and methodology.

1. **Objective Definition:**
   Primary goals: Analyzing resumes, identifying skills, and providing user-centric insights.

2. **Scope Definition:**
   Encompasses data loading, entity recognition, user interaction, and flexible for future enhancements.

3. **Requirement Analysis:**
   Identify and document functional and non-functional requirements for comprehensive development.

4. **Design:**
   Develop a modular and scalable system architecture with an intuitive user interface.

5. **Iterative Development:**
   Adopt an iterative approach, implementing core features and expanding based on user feedback.

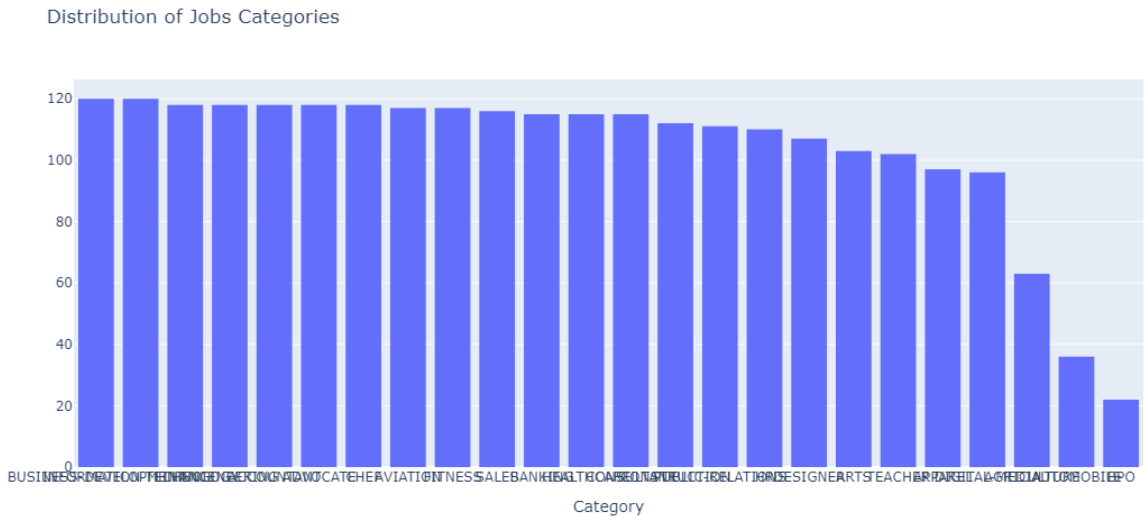
6. **Testing:**
   Implement unit and integration tests to ensure system correctness.

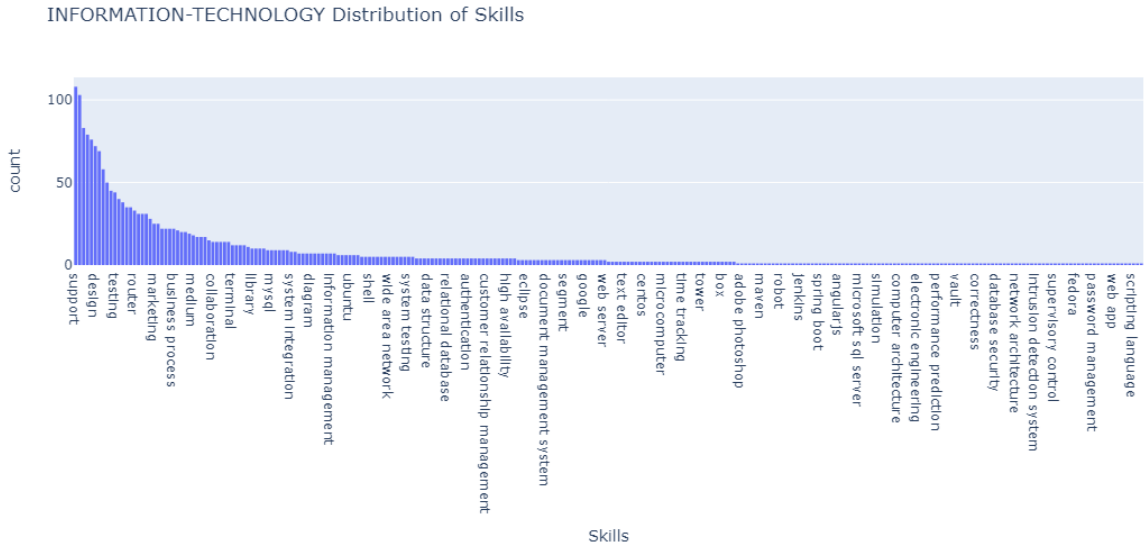
7. **Deployment:**
   Plan a deployment strategy and implement CI/CD pipelines for automation.

5. Results:

- Present project outcomes with visual aids.

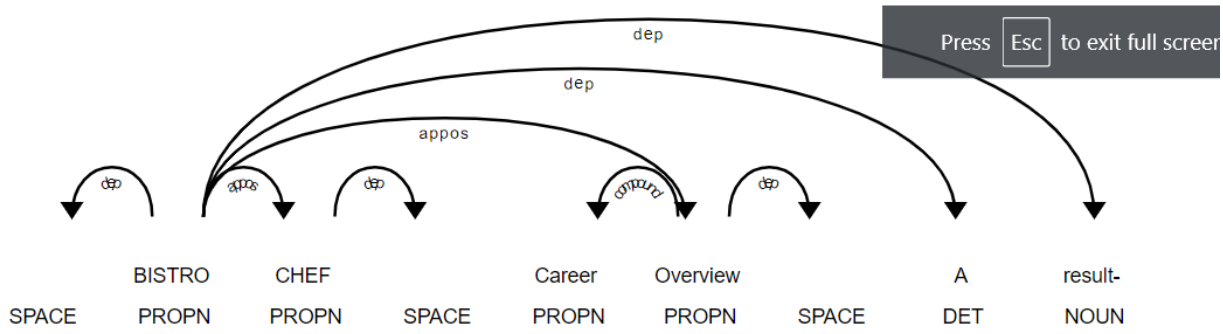
The project outcomes are visually presented through various graphs and visual aids. Histograms depict job category distributions, and word clouds illustrate the most used words in specific job categories. The program provides users with interactive visualizations, empowering them to gain valuable insights into their resumes and job markets.

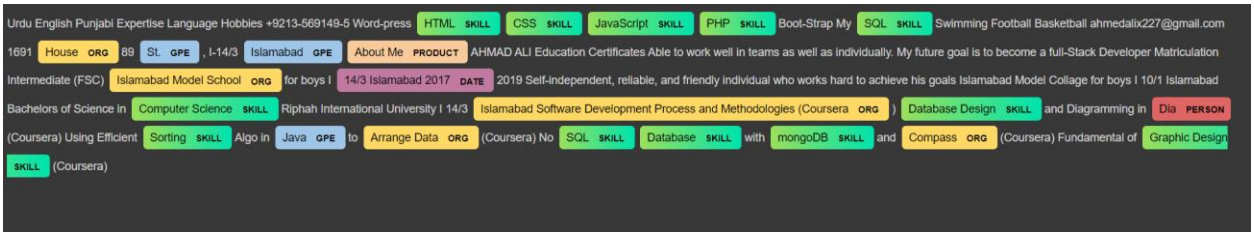


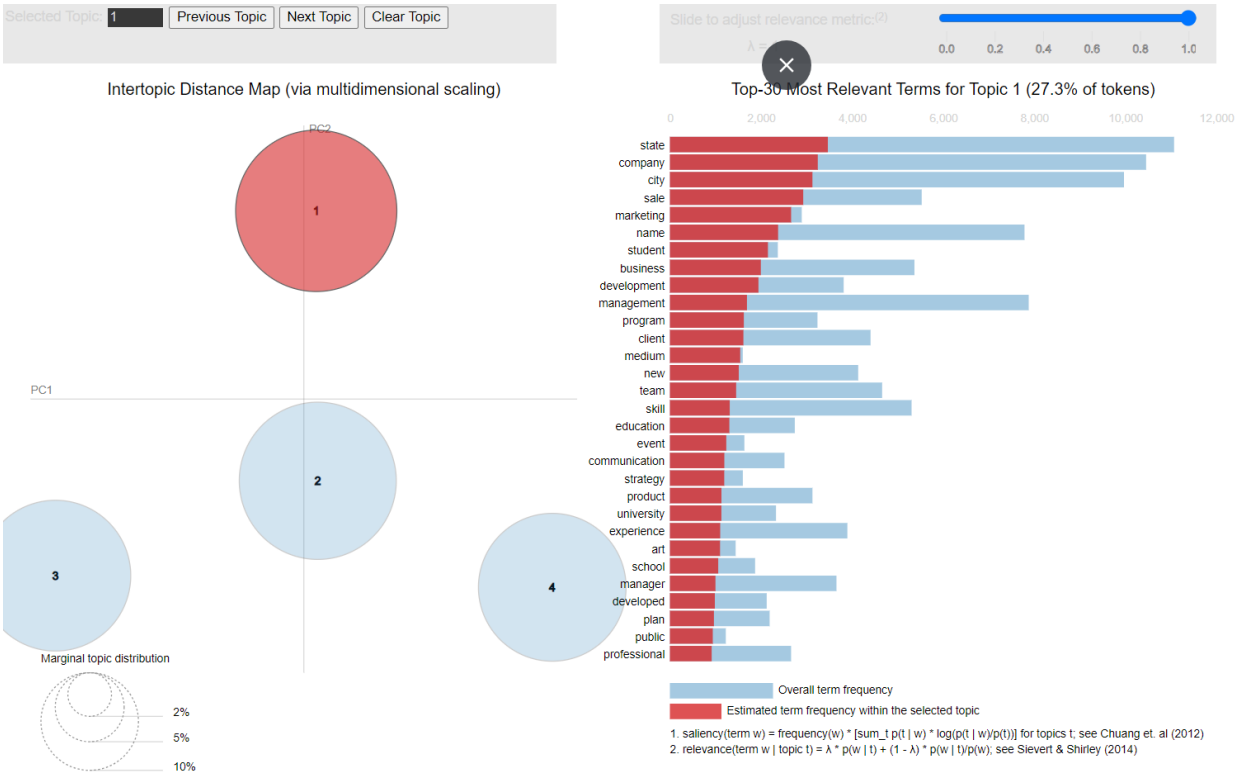


Most Used Words in INFORMATION-TECHNOLOGY Resume

Problem ORG solving Strategic Planning Strong oral Communications ORG Accomplishments Certified to go above and beyond, and providing quality and outstanding customer service. Customer recognition for outstanding and consistent customer support SKILL . Selected to learn new computer programs, and then train others. Work Experience Bistro Chef SKILL 01/2014 to Current Company Name City , State Responsibilities include taking customer orders, addressing customer inquires, opening and closing of establishment, preparing food, inventory control, problem solving, and cash register operation. Customer Service Rep ORG 01/2011 to 01/2013 ORG Company Name City , State Responded to customer inquiries at a fast pace call center. Maintained records, processing payments SKILL to include set up of payment arrangements. Assisted customers by explaining detailed billing formats, and troubleshooting customer's equipment. Reported outages in affected areas and re-laid information to customers. Kitchen Manager 01/2010 CARDINAL to 01/2011 Company Name PRODUCT City , State Responsible ORG for customer orders as well as customer inquiries, opening and closing of establishment, Maintaining records and placing inventory orders. Food ORG handling, preparation, safety and storage. Filling Online, fax orders, and processing of promotional complementary orders. DSP ORG 01/2009 GPE to 01/2010 CARDINAL Company Name City , State Responsible ORG for working with Mentally and Physically disabled people, in a home based setting. Duties are as follows: Assisting individuals with their daily needs, administering of medications, maintaining staff logs, and reports, scheduling Dr. appointments for patient care, and safeguarded consumers well-being. Store Manager 01/2008 CARDINAL to 01/2009 ORG Company Name City , State Daily ORG procedures included opening and closing of establishment. Preparing reports for corporate HQ, updated and maintained file records, ordered supplies, handled customers phone inquiries in a timely manner. Processed loans through verifying customers credit report with the utmost regard to their privacy. This included placing calls to payroll and H.R. ORG departments to verify employment, bankruptcy reports, and bank account information. To include accountable for large sums of monies ORG , blank checks, handled armored car pick-ups, and deposited funds into bank accounts. Internal Auditor ORG 01/2004 to 01/2008 CARDINAL Company Name City , State Maintained and updated records on a daily DATE basis, recalculated figures and insured that formulas were entered correctly. Audited all Electronic Activity entered by Pre-Bill, URT and Start Up department inputted into the (RBMS) Retail Business Management System ORG for management verification. Customer Service Representative ORG 01/2000 CARDINAL to 01/2002 Company Name City , State Assisting customer inquiries in fast pace environment. Maintained records by entering or tracing orders in progress. Assisted customers with extensive product knowledge, Handled shipping and receiving orders. Conducted training of newly hired employees, operating of register for customer checkout, conducted inventory control which included pulling and processing of orders, pricing of items, and stocking the store shelves. Specialist/Customer Service ORG Representative 01/1997 CARDINAL to 01/2000 PRODUCT Company Name City , State Professionally assisted all client inquiries at a fast pace call center. Maintained records, entered and traced orders in progress, assisted technicians with T1 and T3 equipment problems. Explained details to customers on existing orders. Conducted training for newly hired personnel. Tracked nationwide outages and re-laid information to clients, technicians, and management. Assistant Manager 01/1993 PERSON to 01/1996 Company Name City , State Responsible ORG for customer and employee relations. Opening EVENT and closing of establishment, preparing food, hiring and firing of personnel, bookkeeping and accounting SKILL , inventory control, scheduling of personnel, problem solving and accountability of finances. Counter Manager for Ulima 2 Cosmetics ORG . 01/1992 FAC to 01/1993 PERSON Company Name City , State Answered customer questions and concerns, advised clientele of products with extensive product knowledge. Solved problems, maintained record of customer's product orders, handled inventory control, and register operations. Educational Background Diploma : Cosmetology, Photography 1990 DATE SZ Delfter Str PERSON . City , State ORG , Germany GPE Diploma : Biology, Mythology, Computer Science SKILL 1987 DATE SZ GPE Hermannsburg GPE City , State ORG , Germany GPE High School Diploma 1986 DATE SZ GPE Hermannsburg GPE City , State ORG , Germany GPE Associate of Arts ORG ; Arts, Archeology KCTCS GPE City , State ORG , United States of America GPE Skills Computer experience include: Windows NT PRODUCT , Windows SKILL 95 CARDINAL , Vista ORG , Windows 7 PRODUCT , Windows 8 PRODUCT , Microsoft Office ORG , Microsoft Word ORG , Excel PRODUCT , Word Perfect, Power Point FAC , Outlook 2013 DATE , Android ORG , RBMS ORG , AS400 PRODUCT , SMS800 PERSON .


Press Esc to exit full screen
dep
dep
appos
dep
appos
dep
compound
dep
SPACE
BISTRO
PROPN
CHEF
PROPN
SPACE
Career
PROPN
Overview
PROPN
SPACE
A
DET
result-
NOUN


Urdu English Punjabi Expertise Language Hobbies +9213-569149-5 Word-press HTML SKILL CSS SKILL JavaScript SKILL PHP SKILL Boot-Strap My SQL SKILL Swimming Football Basketball ahmedalix227@gmail.com
1691 House ORG 89 St. GPE , I-14/3 Islamabad GPE About Me PRODUCT AHMAD ALI Education Certificates Able to work well in teams as well as individually. My future goal is to become a full-Stack Developer Matriculation
Intermediate (FSC) Islamabad Model School ORG for boys I 14/3 Islamabad 2017 DATE 2019 Self-independent, reliable, and friendly individual who works hard to achieve his goals Islamabad Model Collage for boys I 10/1 Islamabad
Bachelors of Science in Computer Science SKILL Riphah International University I 14/3 Islamabad Software Development Process and Methodologies (Coursera ORG) Database Design SKILL and Diagramming in Dia PERSON
(Coursera) Using Efficient Sorting SKILL Algo in Java GPE to Arrange Data ORG (Coursera) No SQL SKILL Database SKILL with mongoDB SKILL and Compass ORG (Coursera) Fundamental of Graphic Design
SKILL (Coursera)


Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (27.3% of tokens)
0 2,000 4,000 6,000 8,000 10,000 12,000
state
company
city
sale
marketing
name
student
business
development
management
program
client
medium
new
team
skill
education
event
communication
strategy
product
university
experience
art
school
manager
developed
plan
public
professional
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

6. Future Work:

- Discuss potential enhancements or improvements.

- Summarize main findings and contributions.

Potential Enhancements or Improvements:

1. Expand Dataset:

- Enhancement: Increase the size and diversity of the resume dataset to improve the accuracy and generalization of the analysis.

- Rationale: A larger dataset allows for a more comprehensive understanding of skills distribution across various job categories.

2. Advanced NLP Techniques:

- Enhancement: Explore and implement advanced natural language processing techniques for improved entity recognition and semantic analysis.

- Rationale: Utilizing cutting-edge NLP methods may enhance the program's ability to identify skills, entities, and patterns in resumes more accurately.

3. Dynamic Entity Pattern Updating:

- Enhancement: Implement a mechanism to dynamically update entity patterns for named entity recognition.

- Rationale: The ability to update entity patterns without modifying the code allows for adaptability to evolving job markets and skill trends.

4. User Profiling and Recommendation:

- Enhancement: Develop a user profiling system that learns from user interactions and provides personalized skill recommendations.

- Rationale: Tailoring recommendations based on user preferences and historical interactions can enhance user experience and engagement.

5. Integration with Job Portals:

- Enhancement: Integrate the program with job portals to fetch real-time job listings and compare the skills mentioned in job postings with those in resumes.

- Rationale: Real-time integration with job portals adds relevance to the analysis by aligning skills with current job market demands.

6. Enhanced Visualization Options:

- Enhancement: Provide additional visualization options, such as interactive charts and graphs, to offer more ways for users to explore and interpret data.

- Rationale: Diverse visualization options enhance user engagement and facilitate a deeper understanding of the data.

7. Machine Learning for Resume Matching:

- Enhancement: Implement machine learning algorithms to improve the accuracy of resume-to-job category matching.

- Rationale: Machine learning models can learn from historical data and user feedback, optimizing the matching process over time.

Summarized Findings and Contributions:

1. Insights into Job Market Trends:

- The program offers valuable insights into the distribution of skills across different job categories, helping users understand prevalent skills in their desired fields.

2. User-Focused Resume Analysis:

- The user interaction features allow job seekers to input their resumes and desired skills, receiving feedback on the match percentage. This user-focused approach aids in resume refinement.

3. Topic Modeling for Hidden Themes:

- The program employs Latent Dirichlet Allocation (LDA) for topic modeling, revealing hidden themes within resumes and providing a deeper understanding of the content.

4. Dynamic Visualizations:

- Interactive visualizations, including histograms and word clouds, enhance the presentation of data, making it more accessible and engaging for users.

5. Scalability and Flexibility:

- The program is designed with scalability in mind, allowing for potential expansion of the dataset and adaptability to emerging job market trends.

7. References:

- Include a concise list of all sources and citations.

- **Chatgpt**
- **Kaggle**
- **Deepnote**

*** The End ***