Project Report: Intelligent Resume Classification and Ranking System

Team Members:

- Sumaiya Waheed 22k-4201
- Muhammad Sudais 22k-4364
- Azka Sahar Shaikh 22k-4493

Objective:

The primary objective of this project is to develop an intelligent system that automatically classifies and ranks resumes (CVs) based on their relevance to predefined academic job descriptions (JDs), such as Research Assistant and Lab Instructor roles. The system leverages Information Retrieval, Machine Learning, and Natural Language Processing (NLP) techniques to assist hiring panels in screening and shortlisting candidates efficiently.

Motivation:

Manual CV screening is time-consuming, subjective, and prone to human error—especially when dealing with large applicant pools. This project aims to automate the process by:

- Extracting and analyzing relevant sections from resumes.
- Matching resumes against job requirements dynamically.
- Classifying resumes into job roles.
- Ranking resumes based on relevance using ensemble techniques.

Technologies Used:

Component	Tools/Techniques
Frontend	Streamlit
Backend	Python
NLP	TF-IDF, NLTK
ML	Naive Bayes, MultiLabelBinarizer, Scikit-learn

IR	Cosine Similarity
Version Control	Git, GitHub

Dataset Description:

- Labeled Resumes (16 CVs): Manually labeled as either "Research Assistant" or "Lab Instructor".
- Unlabeled Resumes (7 CVs): Used for prediction and ranking.
- Job Descriptions (2 JDs):
    - JD for Research Assistant
    - JD for Lab Instructor

Modules and Features:

1. Preprocessing

- Noise removal (stopwords, punctuation).
- Keyword extraction using TF-IDF.
- Dynamic constraint extraction from JDs.

2. Naive Bayes Classification

- Trained on labeled CVs.
- Predicts job role probabilities for new resumes.

3. IR-Based Ranking

- Calculates cosine similarity between JD and CV vectors.
- Scores resumes based on textual relevance.

4. Ensemble Scoring

- Combines Naive Bayes probability and IR similarity score.
- Produces a final score used to rank resumes for each JD.

5. Multi-label Classification

- Predicts multiple job roles using MultiLabelBinarizer and logistic regression.

6. Streamlit Interface

- Upload JDs and CVs.
- Visualize predicted roles and ranking metrics.
- Expandable scorecards for each CV.

Results and Evaluation:

- **Naive Bayes Accuracy**: ~63.3% on labeled data due to small and biased training data().
- **Ranking Output**: Ranked resumes matched domain expectations (e.g., RA resumes had higher scores for RA JD).
- **Multi-label predictions** aligned with single-label predictions for validation.

Achievements:

- Dynamic skill-based constraint extraction from JDs.
- Seamless combination of IR and ML techniques.
- Robust and clean UI for live demonstration.
- Modular architecture allows future extensibility (e.g., for other domains like industry hiring).

Challenges Faced:

Challenge	Solution
Inconsistent CV formats	Used robust text segmentation heuristics
Small labeled dataset	Used TF-IDF and ensemble methods to compensate
JD parsing errors	Standardized JD structure + null safety checks

Data biased towards lab instructor Applied hyperparameter tuning and cross-validation to minimize bias

Future Work:

- Integrate GPT-based resume summary/explanation.
- Add voice input for JD entry and AI-generated summaries.
- Expand to support image-based resumes (OCR support).
- Integrate a recruiter dashboard with candidate tracking.

References:

1. *Matching People and Jobs: A Bilateral Recommendation Approach*— Jochen Malinowski, Tobias Keim, Prof. Dr. Oliver Wendt, Dr. Tim Weitzel—2006
2. *Expertise Matching via Constraint-based Optimization*—Wenbin Tang, Jie Tang, and Chenhao Tan—2010
3. *Skill Finder: Automated Job-Resume Matching System*—Thimma Reddy Kalva—2013