

Intelligent Resume Screening System

Line-by-Line Module Explanation

Module 1: Data Input Module

Purpose: Ingestion of resumes (PDF/TXT) and job descriptions.

1. Library Imports: Uses 'os' for file handling and 'PyPDF2' for extraction.
2. Directory Scan: Checks 'data/resumes' and 'data/job_descriptions' folders.
3. File Iteration: Loops through each file in the directory.
4. Text Extraction: Converts binary PDF content into readable string text.
5. Data Storage: Organizes text into a structure for bulk processing.

Module 2: Preprocessing Module

Purpose: Cleaning raw text to improve machine understanding.

1. Lowercasing: Standardizes word format (e.g., 'JAVA' to 'java').
2. Character Stripping: Removes punctuation and special characters.
3. Tokenization: Splits continuous strings into individual list elements (words).
4. Stopword Removal: Deletes non-critical words (e.g., 'and', 'the').
5. Lemmatization: Reduces words to their root form (e.g., 'coding' to 'code').

Module 3: Feature Extraction Module

Purpose: Converting text into mathematical vectors.

1. TF-IDF Init: Starts the Term Frequency-Inverse Document Frequency vectorizer.
2. Vocabulary Mapping: Learns every unique keyword in your resume dataset.
3. Weight Calculation: Assigns higher points to technical keywords over common ones.
4. Sparse Matrix: Creates a compressed table of numbers representing the resumes.

Module 4: Similarity & Ranking Module

Purpose: Matching candidates against job criteria.

1. Target Shaping: Transforms Job Description into the same vector space.
2. Cosine Similarity: Measures the 'angle' between JD and each Resume vector.
3. Percentage Calculation: Converts dot-products into a score between 0 and 1.
4. Ranking Logic: Sorts the entire database from highest to lowest score.
5. CSV Output: Export results into 'module5_resume_ranking.csv'.

Module 5: Evaluation Module

Purpose: Statistical validation and manual verification.

1. Score Distribution: Plots histograms to see the data quality.
2. Top-K View: Extracts the absolute best candidates (Top 5) for immediate hiring.
3. Thresholding: Automatically flags candidates who scored below 0.50.
4. Validation Export: Exports a table for HR to verify the AI's match quality.