INTELLIGENT JOB RECOMMENDATION ENGINE

Phase 2 - Feature Engineering and Hybrid Scoring

Member 1 : Ashfaq Ali Anees Ali 40406560
Member 2 : Naveen Kumar Raju 87474061
University of Europe for Applied Sciences
Date: 31-10-2025

Objective

To transform raw, unstructured text data (Resumes and Job Descriptions) into structured, numerical features and establish a robust, initial compatibility ranking metric (Hybrid Score) for all possible Resume-Job pairs.

1. Feature Generation Methodology

The core of this phase involved extracting two distinct types of numerical feature vectors from the documents. These vectors serve as the foundational data for all subsequent machine learning steps.

1.1 Skill Feature Vector (36 Dimensions)

Using a predefined ontology of **36 skill categories** (e.g., IT, LGL, MRKT, FIN), we converted the text content of both Resumes and Job Descriptions into a numerical vector.

- **Process:** A simple binary (0/1) encoding approach was used. If a document contained keywords related to a specific skill category, the corresponding vector position was marked as **1**; otherwise, it was **0**.
- **Output:** This produced two key datasets: resume_skill_features.csv and job_skill_features.csv. These files provide a quantifiable representation of the user's skillset and the job's required skillset, respectively.

1.2 Text Similarity Feature (Vector Embeddings)

To capture contextual relevance, tone, and overall narrative alignment—elements often missed by simple keyword matching—we employed a vector-based technique on the entire document text.

- **Method: TF-IDF (Term Frequency-Inverse Document Frequency)** was used to generate sparse, high-dimensional vector embeddings for every Resume and Job Description. These embeddings capture the semantic importance of each word within the document relative to the entire corpus.
- **Purpose:** These vectors are primarily used to calculate the **Text Match Score** (**Cosine Similarity**), which measures how semantically close the two documents are.

---

2. Hybrid Compatibility Scoring

A single metric was developed to provide an initial, highly reliable compatibility ranking. This **Hybrid Score** combines two powerful and complementary similarity measures using a weighted average.

2.1 Score Components

Component	Weight	Focus	Calculation
Skill Match Score	**40% (0.4)**	Evaluates hard requirements by measuring overlap across 36 skill-category features.	Computed using cosine similarity between the 36-dimensional skill vectors.
Text Match Score	**60% (0.6)**	Assesses contextual and semantic fit between job description and résumé text.	Computed using cosine similarity on high-dimensional TF-IDF text embeddings.

2.2 Final Hybrid Score Calculation

The two scores were linearly combined to produce the final, balanced metric:

Hybrid Score = (Text Match Score × 0.6) + (Skill Match Score × 0.4)

2.3 Output Matrix

The Hybrid Score was calculated for every unique combination of Resume ID and Job ID within the data set.

- **Output File:** compatibility_score_matrix_final.csv
- **Function:** This matrix serves as the **primary feature** for the subsequent machine learning pipeline (Phase 3). It provides the supervised model with a strong, pre-engineered compatibility signal, significantly improving the model's ability to learn what constitutes a "Good Match."

---

3. Conclusion

Phase 2 successfully accomplished the critical task of transforming unstructured text into structured, actionable numerical data. By generating 36-dimensional skill vectors and combining them with text embeddings, we created the robust **Hybrid Score**.

This Hybrid Score matrix is now fully prepared for **Phase 3 (Model Training and Classification)**, where it will be used to generate training labels and power the final predictive classification engine.