**Phase 1: Resume Text Extraction**

**Objective:** Extract raw text from resumes for further processing.

**Techniques Used:**

* **OCR (Optical Character Recognition):**
  + Used pytesseract with the pillow library to extract text from image-based resumes (e.g., .jpg files).
* **Preprocessing:**
  + Applied cleaning techniques such as:
    - Lowercasing
    - Removing non-alphanumeric characters
    - Handling whitespace and formatting issues

**Evaluation Criteria:**

* **Visual validation**: Printed and manually reviewed extracted text.
* **Accuracy of text coverage**: Verified if major resume components (name, skills, education, etc.) were present.

**📌 Phase 2: Text Cleaning & Normalization**

**Objective:** Prepare raw text for structured processing.

**Techniques Used:**

* **Noise removal** using regex (e.g., stripping headers, removing non-informative lines).
* **Tokenization, stemming, and lemmatization** using NLP libraries (NLTK, spaCy).
* **Date standardization** using dateutil to ensure uniformity in employment periods.

**Evaluation Criteria:**

* **Manual inspection** of intermediate outputs.
* **Consistency** in format (e.g., all dates formatted as YYYY-MM).

**📌 Phase 3: Named Entity Recognition (NER)**

**Objective:** Identify and label key entities such as skills, organizations, job titles, etc.

**Techniques Used:**

* **Pretrained NER models** in spaCy to extract entities.
* **Custom NER training** on labeled examples using spaCy pipeline.
* **Entity categories included**: SKILL, ORG, DATE, EDU, NAME, etc.

**Evaluation Criteria:**

* **Precision / Recall / F1-score** on labeled validation data.
* **Manual validation** for small test cases.

**📌 Phase 4: Relationship Extraction**

**Objective:** Associate related entities (e.g., skill with experience, organization with job title).

**Techniques Used:**

* **Dependency parsing** using spaCy to understand sentence structure.
* **Custom rules** to link entities based on sentence proximity and patterns.

**Evaluation Criteria:**

* **Correctness of links** (e.g., matching the right job title to the correct company).
* **Validation** against known labeled relations.

**📌 Phase 5: Database Design & Storage**

**Objective:** Store parsed resume information in a structured, queryable format.

**Techniques Used:**

* **SQLite** used for lightweight storage.
* Tables designed for Candidate, Skills, Experience, Education, etc.
* **Pandas** used to insert and query data.

**Evaluation Criteria:**

* **Data integrity**: Ensured relationships via foreign keys.
* **Query output correctness**: Tested SQL joins and filters.

**📌 Phase 6: Machine Learning on Resume Data**

**Objective:** Use structured resume data to build predictive models.

**Techniques Used:**

* **Feature encoding** with MultiLabelBinarizer on skill lists.
* **Model training** using LogisticRegression and NaiveBayes.
* **Cross-validation** to prevent overfitting.
* **Synthetic data generation** with varied realism (noise, overlap).

**Evaluation Criteria:**

* **Model accuracy**, **F1 score**, and **confusion matrix**.
* Target accuracy: **~0.7–0.85** for realistic performance.