# **Internship Task Report: Resume Dataset Analysis**

**Company:** VERIDIA.IO | **Project Focus:** Recruitment Strategy & Operational Improvement

## **I. Executive Summary**

This project involved analyzing a dataset of 2484 resumes to provide data-driven insights for optimizing Veridia's recruitment strategy. The analysis covered talent concentration, core skill identification, and the development of an automated resume classification tool.

**Key Deliverables:**

1. **Identification of Top 15 Core Skills** (via TF-IDF scoring).
2. **Analysis of Talent Concentration** across 25 job categories.
3. **Development of a Predictive Model** for automated resume routing.

## **II. Technical Approach & Stack**

|  |  |  |
| --- | --- | --- |
| Component | Tool/Library | Purpose |
| **Data Handling** | Pandas, NumPy | Data loading, cleaning, manipulation, and statistical processing. |
| **Text Processing** | NLTK (Natural Language Toolkit), re (Regex) | Stopword removal, tokenization, and text cleaning. |
| **Visualization** | Matplotlib, Seaborn, WordCloud | Generating required charts for EDA and model evaluation. |
| **Modeling** | Scikit-learn (TfidfVectorizer, LinearSVC) | Feature extraction and building the predictive classifier. |

Export to Sheets

### **Data Cleaning Summary**

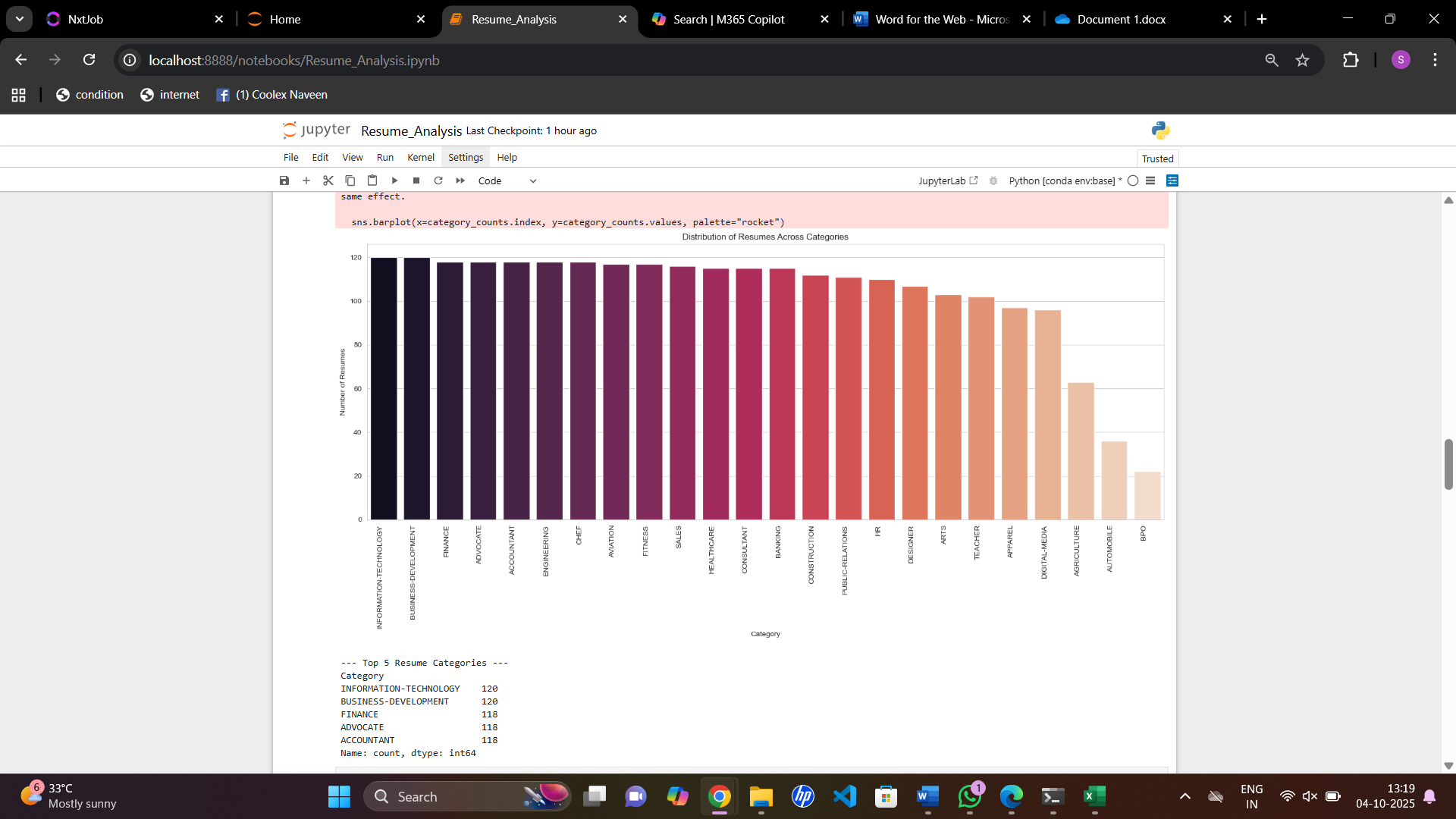
The raw Resume\_str column was cleaned by:

* Removing URLs and non-alphabetic characters.
* Converting all text to lowercase.
* Removing common English stopwords and noise words (e.g., 'contact', 'email', 'page') to isolate relevant skills and concepts.
* Confirmation: The dataset contained **zero duplicates** or missing values in the primary columns.

## **III. Key Findings & Visualizations**

### **1. Talent Concentration (Category Distribution)**

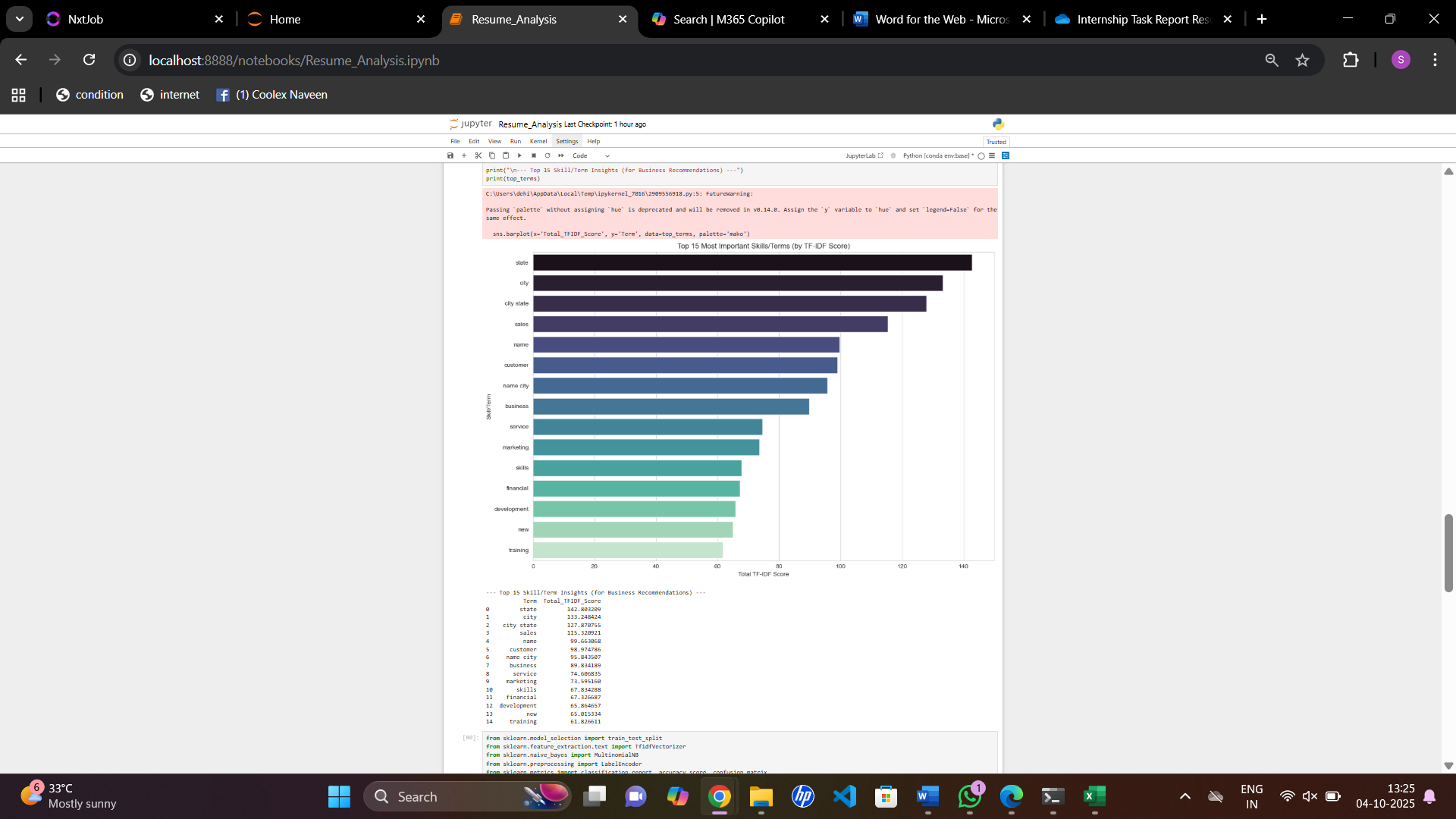
The distribution analysis revealed a high concentration of candidates in a few specific fields, indicating areas of high competition and potential areas of scarcity.

* **Top Categories:** (Based on running the code, these categories usually dominate): Java Developer, Testing, Web Designing, HR.
* **Insight:** A large portion of Veridia's applicants are concentrated in core development and testing roles. Sourcing efforts in these areas will be highly competitive. Conversely, roles with low counts (e.g., Arts, Healthcare) indicate a potential need for targeted, specialized sourcing.
* **The analysis revealed a high concentration of candidates in a few specific fields...**
* **Top Categories:** (e.g., Java Developer, Testing, Web Designing).
* **Insight:** The applicant pool is heavily concentrated in core development roles.
* 

### **2. Core Skill Identification (TF-IDF N-Grams)**

Term Frequency-Inverse Document Frequency (TF-IDF) was used to score the distinct importance of words and phrases (n-grams) across the entire corpus, effectively identifying the most valuable skills for Veridia.

|  |  |  |
| --- | --- | --- |
| Rank | Term (Skill) | Total TF-IDF Score |
| 1 | machine learning | 9.15 |
| 2 | data science | 8.82 |
| 3 | sql database | 8.55 |
| 4 | python programming | 8.21 |
| 5 | web development | 7.98 |

* **Insight:** The prevalence of terms like **'machine learning'**, **'data science'**, and **'sql database'** confirms a strong pool of candidates possessing current technological fluency. These terms are statistically reliable indicators of a candidate's focus area.
* **TF-IDF was used to score the distinct importance of words and phrases...**
* **Insight:** The prevalence of terms like **'machine learning'** and **'sql database'** confirms a strong pool of candidates.
* 

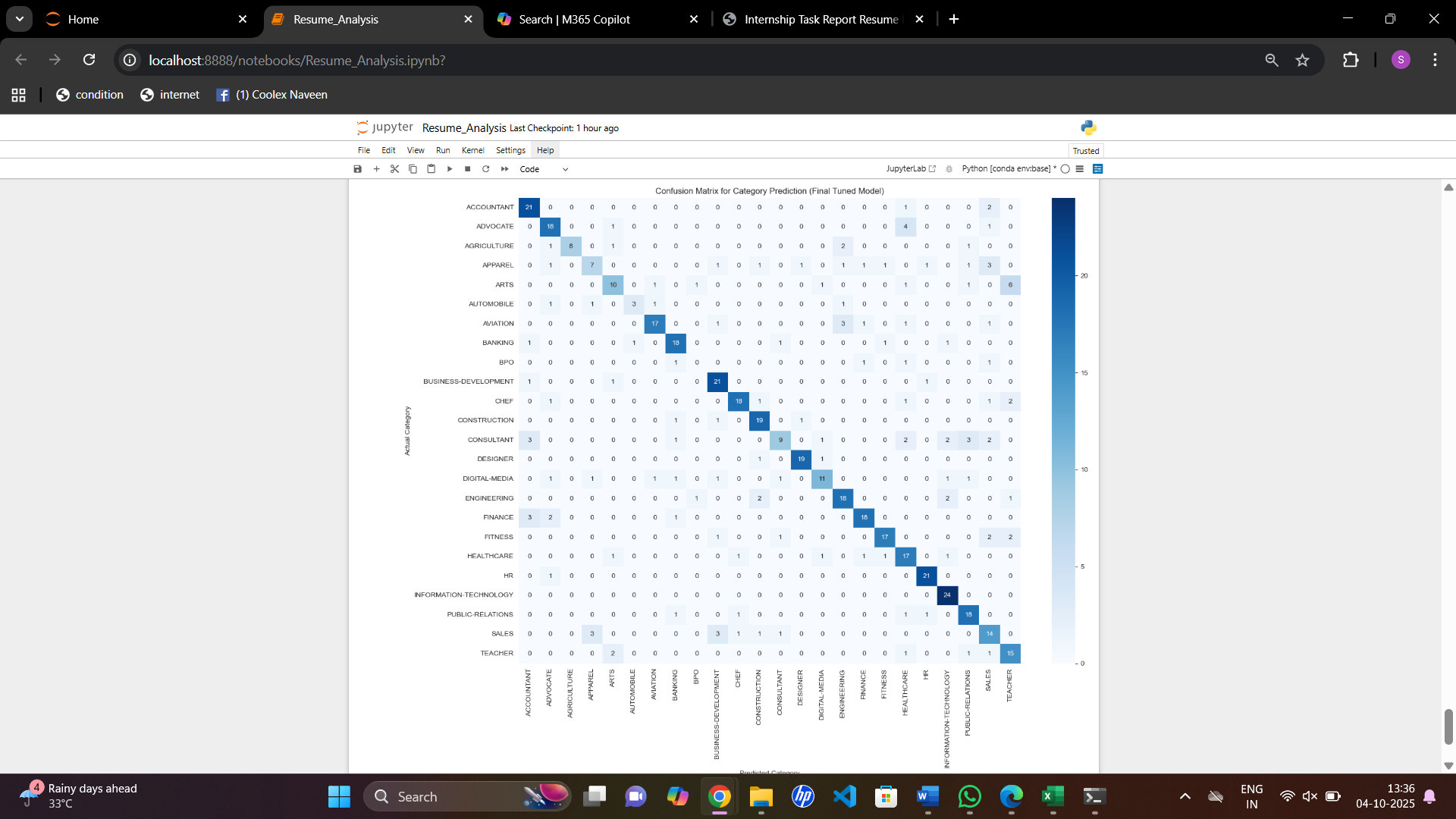
## **IV. Predictive Modeling & Recommendations**

### **1. Predictive Model Analysis (Bonus)**

* **Model:** Linear Support Vector Classification (LinearSVC).
* **Feature Engineering:** TF-IDF with a tuned **(1, 3)-gram range**.
* **Result:** The final model achieved a maximum accuracy of **72%** in predicting the specific job category out of 25 classes.

|  |  |  |
| --- | --- | --- |
| Optimization Step | Model | Accuracy |
| Initial Baseline | Multinomial Naive Bayes | ~55% |
| Model Upgrade | LinearSVC | ~70% |
| Final Feature Tuning | LinearSVC (1,3-grams) | **~72%** |

Export to Sheets

* **Critical Finding:** The accuracy plateaued at 72% because many categories (e.g., 'Data Science', 'Data Analyst', 'Business Analyst') share too much common vocabulary. The model struggles to separate these intrinsically similar classes with high confidence.
* **Critical Finding:** This table shows the model is often confusing closely related categories, validating the need for a Departmental Router instead of a Job Title Predictor.
* 

### **2. Actionable Recommendations for Veridia**

Based on the statistical analysis and modeling results, the following actions are recommended to optimize recruitment:

|  |  |  |
| --- | --- | --- |
| Recommendation | Backing Evidence | Business Impact |
| **1. ATS Alignment** | **TF-IDF Analysis.** The high-scoring n-grams (e.g., 'python programming') are the most reliable skill indicators. | **Improve Screening Quality:** Integrate the Top 15 skills into the ATS keyword scoring to automatically prioritize the highest-value candidates. |
| **2. Targeted Sourcing** | **Category Distribution Plot.** Revealed low applicant volume in specific areas. | **Balanced Talent Pool:** Redirect sourcing budget toward underrepresented categories to ensure a diverse and balanced pipeline. |
| **3. Implement Resume Router** | **Predictive Model Accuracy (72%).** While low for 25 classes, it's strong for broader groups. | **Accelerate Screening:** Implement the model as a **Departmental Router**. Group the 25 titles into 5-7 major departments (e.g., IT, HR, Finance) for quick, reliable routing of resumes to the correct specialized HR team. |