Inspiring Excellence

CourseCode: CSE424

Pattern Recognition

Name: Ahsan Habib

ID: 22201027

Section: 02

Assignment 02

Date: 27 April, 2025

Project Title: Resume Screening For job-matching using Natural Language Processing and Machine Learning.

Submitting to,

Annajiat Alim Rasel Sir

DataSet: https://www.kaggle.com/datasets/noorsaeed/resume-datasets

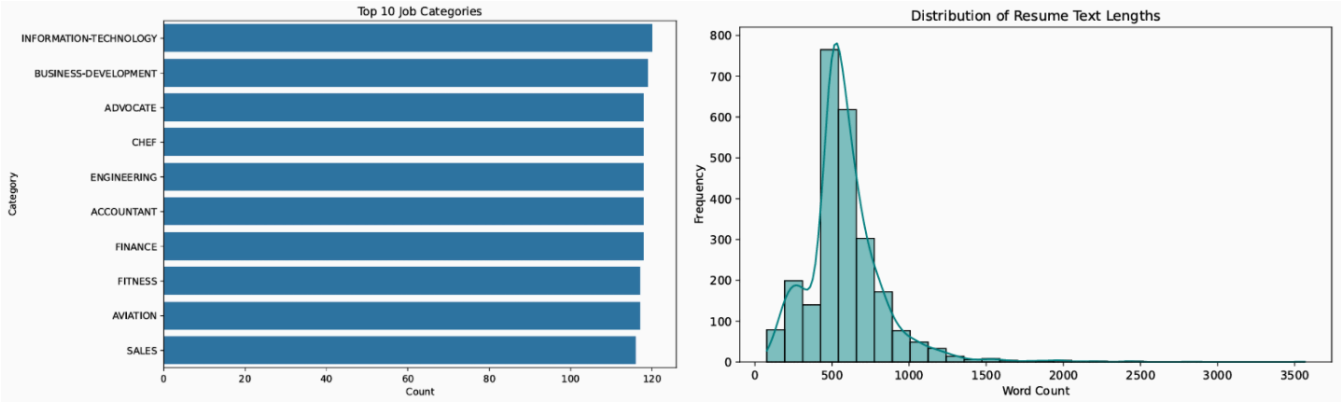
Dataset Report : By doing a basic analysis with OpenRefine on my dataset for my project (Resume Screening For job-matching using Machine Learning), we can see that there are mainly three columns in my dataset. First column is the ID of Applicants, which is being used to create anonymity for the candidates to reduce Bias. Second is job categories of different resumes. Third is the features about Job Descriptions. From the analysis of OpenRefine we can see there are a total of 2484 job resumes. Among them we can find 24 distinct job categories. Among them - { Information-Technology 120, Business-Development 120 , Advocate 118, Chef 118, Engineering 118, Accountant 118, Finance 118…..}and so on. For better understanding, I am attaching a EDA Report Summary, which I generated by using Matplotlib and Pandas.

EDA Summary

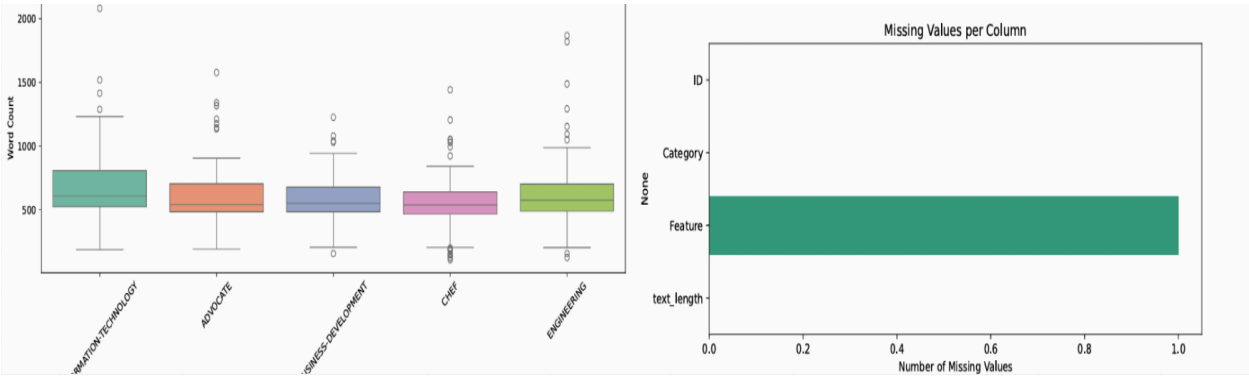
```
Dataset Summary:
----------------
Total Records: 2484
Missing Values:
ID             0
Category       0
Feature        1
text_length    0
Resume Text Length - Mean: 587.01, Median: 549.0, Mode: 482
Top Categories:
Category
INFORMATION-TECHNOLOGY    120
BUSINESS-DEVELOPMENT      120
ADVOCATE                  118
CHEF                      118
ENGINEERING               118
```

Fig 01: EDA Summary report

As the content of my dataset is more Long-text based, for which OpenRefine is not Enough, so to do more analysis I have used { Seaborn, Pandas, Matplotlib and collections } for further analysis. By using these awesome python Libraries we were able to create some graphs and analysis for better data Visualisation.



On the left side we can see Horizontal bar chart, which shows the Frequency of Job Categories, On the Right we can see have create a histogram with Kernel density Estimate (to smoothen Distribution Curve) using Seaborn, which reveals skewness, central tendency and outliers.



We have also added a box plot and a missing value bar plot, by using Seaborn, to help visualize the dataset better, The Box plot on the left side, helps to detect the median, outliers and quartiles of the length of Long-texts. The graph on the right shows us the frequency distribution of Missing values, which Helps us to Understand the missing Columns across all variables.

In conclusion, one must understand the data at a working level after combining OpenRefine for the initial cleaning stage and Python modules like Pandas, Matplotlib, Seaborn, and collections for the deeper analysis stage. The visualizations, including frequency bar charts, histograms with KDE, and boxplots with missing value graphs, have brought the main information about job category distribution, central tendency, or data completeness into the open while expressing the disadvantage of the said tools with long-text data. Such information is very much applicable in preparing the data for machine learning models for other resume screening processes, where ensuring the dataset is clean, balanced, and well-understood is done first before moving on to the modeling part.