Project NLP: Resume Extraction - Week 8 Deliverables

NLP Interns at Data Glacier

BatchCode: LISUM20

Group Name: TeamNLP

Name	Email	Country	College/Company	Specialization
Nilsu Bozan	bozannilsu@gmail.com	Turkey	Binghamton University/Istanbul Technical University	NLP
Nishchay Vaid	Nishchay89@gmail.com	USA	Rutgers University	NLP
Anish Mitra	anishmitra9666@gmail.com	USA	Montana State University	NLP
Sukriti Macker	sm11017@nyu.edu	USA	New York University	NLP

Data Understanding:

.json file provides an understanding of the individual's professional background, skills, work experience, and educational qualifications. It can be useful for assessing the candidate's suitability for specific job roles, evaluating their expertise in relevant technologies, and understanding their career progression.

What type of data you have got for analysis?

Total Rows : 200

Total Fields in Json File : 2 (Content and Annotation)

In a nutshell we have the following information:-

Personal Information: (Dataframe 1)

1. Name (string): The name of the individual.
2. Designation (string): The designation or job title of the individual.
3. Location (string): The location of the individual
4. Email Address (string): The email address of the individual

Work Experience: (Dataframe 2)

1. Company (string): The name of the company where the individual worked.
2. Designation (string): The job title or designation held by the individual.
3. Location (string): The location of the company.
4. Duration (string): The duration of employment or work period at the company.

Education: (Dataframe 3)

1. Degree (string): The degree obtained by the individual.
2. College Name (string): The name of the college or institution attended by the individual.
3. Graduation Year (string): The year of graduation.

Skills : (Dataframe 4) A (string) list of skills possessed by the individual, including programming languages, tools, technologies, and databases.

What are the problems in the data (number of NA values, outliers , skewed etc)

The dataset exhibits certain challenges that need to be addressed for effective analysis and processing. Specifically, the number of columns and rows in each label is inconsistent, leading to a non-uniform structure. For instance, some entries in the skills dataframe contain only one skill, while others have up to five. This inconsistency results in numerous NA values, making it difficult to accurately determine the number of skills through resume parsing. Ideally, a consistent number of bullet points for each section would have facilitated a more straightforward extraction process.

It is important to note that there are no outliers or skewed data in this qualitative NLP data project. The primary objective of this project is to streamline the transfer of information from 200 raw resumes into a structured database. By leveraging NLP techniques, the project aims to identify and classify relevant entities such as person names, college names, academic information, relevant experiences, skill sets, and more.

To ensure the reliability and quality of the extracted information, steps will be taken to address the issues related to non-uniformity and NA values. This may involve applying data normalization techniques to establish a consistent structure across the dataset, filling in missing values using appropriate imputation methods, and implementing validation checks to validate the accuracy of the extracted entities.

By addressing these challenges, the project aims to streamline the process of resume analysis and enable HR professionals to efficiently shortlist promising candidates based on relevant qualifications and experiences.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

To overcome problems like NA values and outliers in the dataset, you can apply various approaches to ensure the data is handled effectively and accurately. Here are some commonly used techniques:

1. Handling NA Values:

- Data Imputation: NA values can be replaced with estimated or calculated values. This approach helps in maintaining the integrity of the dataset and avoids the removal of valuable information. Imputation

techniques can include mean, median, mode imputation, or more advanced methods like regression imputation or multiple imputation.
- NA as a Separate Category: Another option is to treat NA values as a distinct category or class. For example, in the case of missing skill values, you can assign a specific label like "Unknown" or "Not Specified" to indicate the absence of a skill.

2. Handling Outliers:

- Outlier Detection: Outliers are observations that significantly deviate from the normal behavior of the data. Various statistical techniques and algorithms can be employed to identify outliers. Common approaches include the use of statistical measures like z-scores or interquartile range (IQR), visualization techniques like box plots or scatter plots, or machine learning algorithms like isolation forest or k-nearest neighbors (KNN) for anomaly detection.
- Treatment of Outliers: Once outliers are identified, you have several options for handling them. You can remove the outliers if they are due to data entry errors or anomalies that don't represent the true distribution. Alternatively, you can transform the outliers using techniques like winsorization, where extreme values are replaced with values closer to the rest of the data. Another approach is to keep the outliers as a separate category if they are meaningful and significant to the analysis.

3. Balancing Classes: In the case of non-uniformity in the number of columns and rows for each label, you can balance the classes to ensure equal representation and improve the overall performance of the model. Techniques like undersampling (removing samples from the majority class), oversampling (replicating or generating new samples for the minority class), or a combination of both (SMOTE - Synthetic Minority Over-sampling Technique) can be employed to achieve class balance.

4. Other NLP Techniques: Text cleaning and preprocessing techniques like removing stop words, handling misspelled words, or using lemmatization/stemming can enhance the quality and consistency of the data.