Data Intake Report

Name: NLP: Resume Extraction
Report date: 05.06.2024
Internship Batch: LISUM32
Version: 1.0
Data intake by: Monisha Shree Senthil Nathan
Data intake reviewer: Data Glacier
Data storage location:

Tabular data details:

Total number of observations	200
Total number of files	1
Total number of features	2
Base format of the file	json
Size of the data	3.2+ KB

Proposed Approach:

- Dedup validation (identification) approach:
  - Utilize unique identifiers such as `content` to identify duplicate records within the dataset.
  - Use pandas functions like duplicated() and drop_duplicates() to identify and remove duplicate records based on the identified key fields.
  - Review the dataset before and after deduplication to ensure that duplicate records have been successfully identified
- Assumptions for Data Quality Analysis:
  - Resumes in the dataset adhere to a consistent format or structure.
  - The annotated entities in the dataset accurately represent the information present in the resumes.
  - Entity annotation is consistent across the dataset, with clear guidelines for labeling.