**CAPSTONE PROJECT: DATA ASSESSMENT**

**PROJECT TOPIC: Building a Resume Parser in Python using NLP and deploying the model on a cloud-based platform using an application.**

**GROUP 1**

SIDDHARTHA PATRA – 0787887

USAMA NORAT – 0792225

AMIR DAHYA – 0791252

**INTRODUCTION ABOUT DATA**

The data for the Resume Parser project consists of 220 resumes that were downloaded from an online job’s platform. These resumes were then uploaded to Dataturks online annotation tool, where they were manually annotated with information about key entities of interest. This annotation process involved identifying specific information within the resumes, such as the candidate's name, education, and work experience, and distinguishing it from other information.

The Dataturks tool then automatically parsed the documents and generated JSON formatted training data. Each line of the training data contains the text corpus from the resumes, along with the annotations of the important entities. This data is the foundation of the Resume Parser, as it will be used to train the NLP model to accurately extract information from resumes and match it with job requirements.

The annotated training data provides a comprehensive resource for training the Resume Parser, and it can be found at the specified location. The training data represents a diverse and representative sample of resumes from the online jobs platform, ensuring that the Resume Parser will be able to handle a wide range of information and provide accurate results.

Overall, the annotated training data is a critical component of the Resume Parser project, as it will be used to train the NLP model to accurately extract and match information from resumes and job requirements. With this data, the Resume Parser will be able to provide accurate and relevant information to recruiters, making it a valuable tool for organizations and recruiters alike.

**DATA ASSESSMENT STEPS:**

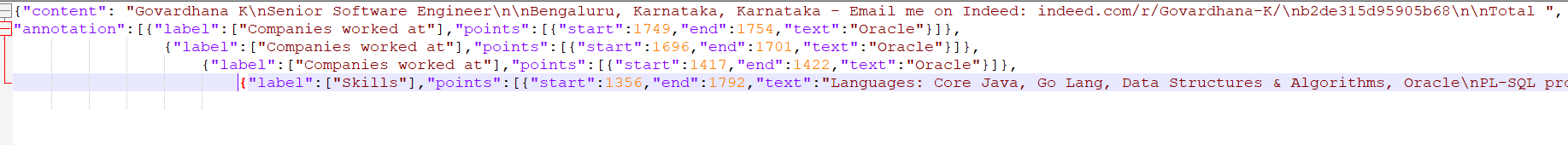
1. **Data Source(s) and Gathering:** For our project, we conducted a comprehensive search for the most suitable datasets to use. We analyzed different sources and critically evaluated the data collection methods and the quality of the data. Our analysis led us to the conclusion that the dataset of 220 annotated resumes collected from an online jobs platform was the most appropriate for our project needs. The resumes were parsed by the Dataturks online annotation tool, which then annotated the important entities such as the candidate's name, education, and work experience. The annotated data was generated in a JSON format and can be used to train machine learning models for named entity recognition (NER).

Additionally, the annotation process and the capabilities of the Dataturks tool played a significant role in our decision to select this dataset. The tool's automatic parsing of the resumes and manual annotation of important entities ensured that the data was accurate and reliable. Despite the limited size of the dataset, we believe it provides a solid foundation for our project work and can be augmented with additional data if required. The thorough analysis of the data collection methods and the quality of the data allowed us to make an informed decision and select the most appropriate dataset for our project.

1. **Dataset Features/Annotation:** The annotated dataset used in the Resume Parser project has a JSON format and contains important information about candidate resumes. The structure of each line in the dataset consists of two main elements: "content" and "annotation".

* "Content" represents the text corpus of the resume, while "annotation" is an array of objects that describes the annotations made on the text. Each object in the "annotation" array has two main elements: "label" and "points".
* "Label" is an array of strings that represents the categories or labels assigned to the annotated information. For example, the labels could be "Candidate Name", "Education", "Work Experience", etc.
* "Points" is an array of objects that represent the specific annotations made on the text corpus. Each object in the "points" array contains three elements: "start", "end", and "text". "Start" and "end" are the character positions in the text corpus where the annotated information begins and ends, respectively. "Text" is the actual text that was annotated.

In this way, the annotated dataset provides a structured representation of the information contained in the candidate resumes, making it easier to analyze and use for training the Resume Parser\_model.

Here is a snapshot for our data. 

1. **Dataset Terms of Use:** The annotated dataset generated for the Resume Parser project is a valuable resource for the development of machine learning models. The dataset has been created with the aim of enabling organizations and individuals to train models that can extract relevant information from resumes. To ensure that the dataset is used in an ethical and responsible manner, the following terms and conditions have been established for its use:

* The dataset is intended for use in the development of Resume Parsers and may not be used for commercial purposes.
* The dataset may not be re-distributed or made available to any third party without the explicit written consent of the creators.
* The creators of the dataset accept no liability for any damages arising from the use of the dataset, including but not limited to errors or inaccuracies in the data.
* Creators of the dataset reserve the right to revoke access to the dataset at any time for any reason.
* Any publications or presentations using the dataset must properly acknowledge the creators of the dataset and cite the source of the data.

1. **Dataset Reproducibility:** It is important for us to ensure that the results generated from the annotated dataset used in my Resume Parser project can be easily reproduced by others. To achieve this, we will carefully document the data preprocessing steps and clearly define the model architecture and hyperparameter tuning process. we will also use appropriate evaluation metrics and make the code used in the project publicly available. Moreover, we will provide thorough documentation of the methods used and the results obtained to ensure transparency and accountability in our project. These steps will help us to ensure that the annotated dataset used in my Resume Parser project can be used by others to develop machine learning models that can accurately extract relevant information from resumes.
2. **Dataset Storage & Handling:** The annotated dataset used in the Resume Parser project should be stored and handled with care to ensure its confidentiality and privacy. One option is to store the data securely on a cloud-based platform, Google Cloud Platform (GCP). This will ensure that the data is easily accessible and can be backed up to prevent loss in the event of a data breach or hardware failure.

The data should only be shared with authorized personnel and should be protected by appropriate security measures, such as encryption, password protection, and access control. It is also important to be mindful of privacy laws and regulations that may apply to the data and to take appropriate steps to comply with these regulations. By properly storing and handling the annotated dataset, the Resume Parser project can ensure that the data is used responsibly and ethically.

**REFERENCES**

1. <https://analyticsindiamag.com/dataturks/>
2. <https://github.com/DataTurks-Engg/Entity-Recognition-In-Resumes-SpaCy>
3. <https://towardsdatascience.com/a-review-of-named-entity-recognition-ner-using-automatic-summarization-of-resumes-5248a75de175>
4. <https://medium.com/@debanjanmahata85/natural-language-processing-with-spacy-36b90b9afa3d>