# NLP Resume Extraction - Project Document

## Team Member Details

- **Group Name:** Lone Data Enthusiast
- **Team Members:**
    1. Name: Ritu Kukreja
       Email: rk977@drexel.edu
       Country: United States
       College/Company: Drexel University
       Specialization: Data Science, NLP, Data Analyst

## Problem Description

- **Problem:** Resumes contain surfeit information that is not relevant for the HR/authority, and they have to manually process the resumes to shortlist the promising candidates for them. And, thus making the shortlisting task a herculean task for HR. By making use of the NER(Named Entity Recognition) model of NLP this problem can be solved by finding and classifying the entities that are present in each resume into predefined classes such as person name, college name, academics information, relevant experiences, skill set, etc.

- **Importance:** The problem of manually processing resumes is of significant importance due to its widespread impact on various applications within the recruitment and HR domain. The traditional method of sifting through resumes is time-consuming and often results in inefficiencies and errors. The use of Named Entity Recognition (NER) models in natural language processing offers a transformative solution to this challenge. Automating the extraction and classification of entities within resumes, such as personal details, educational qualifications, experiences, and skills, holds the potential to revolutionize the HR industry. It can lead to faster, more accurate candidate shortlisting, improved recruitment efficiency, reduced workload for HR professionals, and enhanced scalability to handle large volumes of applications.

- **Project Objectives:** The primary objectives of this project are to develop and implement an NER model tailored to resume analysis. This model will enable the efficient extraction and classification of entities within resumes, categorizing them into predefined classes like personal information, educational history, experiences, and skill sets. The project aims to automate the entire resume processing workflow, creating a user-friendly interface for HR professionals. Additionally, the project will focus on achieving scalability to handle a high volume of resumes effectively. Performance evaluation will be a key component to ensure the accuracy and reliability of the NER model. Ultimately, the project's goal is to provide a valuable tool that streamlines and enhances the resume processing workflow for HR authorities.

## Business Understanding

- **Business Relevance:** This project aligns closely with the needs and objectives of businesses and industries, particularly within the HR and recruitment domains. HR departments in various organizations face the daunting task of manually processing large volumes of resumes during the recruitment process. The traditional approach is time-consuming and prone to errors, making it challenging to efficiently shortlist promising candidates. By automating the extraction and classification of entities within resumes using Named Entity Recognition (NER) models, this project addresses a critical business need. It streamlines the hiring process, improves the quality of candidate selection, and significantly reduces the manual workload for HR professionals.

- **Applications:** The applications of resume extraction are diverse and extend beyond HR and recruitment. Any industry or organization that deals with a substantial influx of unstructured text data can benefit from this technology. Beyond shortlisting candidates, it can be applied to automate the extraction of valuable insights from text data, such as academic research, content curation, and information retrieval. Resume extraction has the potential to enhance efficiency and accuracy in processes that involve parsing and analyzing textual information.

# Project Lifecycle and Deadlines

- **Project Lifecycle:** The project can be divided into several key phases. The initial phase involves data collection, where a diverse dataset of resumes will be gathered. The next phase focuses on data preprocessing, including text cleaning, formatting, and structuring the resumes. Subsequently, the model development phase entails creating an NER model for resume analysis. Post-model development, the project will enter an evaluation phase to ensure the model's accuracy and effectiveness. Finally, an implementation phase will integrate the model into a user-friendly interface for HR professionals.

- **Deadlines:** The project timeline will span approximately six months. Data collection is estimated to take one weeks. Data preprocessing is expected to be completed within two days. Model development, evaluation, and fine-tuning will collectively require 2 more days. The final phase, which involves implementation, interface development, and user testing, is estimated to take a day. This timeline is subject to adjustment based on project complexity and the availability of resources.

# Data Intake Report

- **Dataset:** Provide the source of the dataset you are using for this project. In your case, it's the dataset from this GitHub link.

- **Data Preprocessing:** In the initial stage of the project, data preprocessing played a pivotal role in preparing the raw resume dataset for analysis. The dataset, consisting of resumes in various formats, underwent a series of preprocessing steps to ensure uniformity and quality. These steps included the removal of irrelevant or sensitive information to comply with privacy and confidentiality regulations. Additionally, text cleaning techniques, such as removing special characters, standardizing text case, and handling missing data, were applied to enhance the quality and consistency of the textual content within the resumes. Formatting adjustments were made to ensure a consistent structure across all resumes, facilitating the subsequent Named Entity Recognition (NER) model's accuracy. Moreover, data was transformed into a structured format for effective model training, making it ready for the

NER model development phase. These preprocessing measures aimed to create a clean and standardized dataset, which is a critical prerequisite for the success of the resume extraction project.

Lone Data Enthusiast
10/19/2023