AI Project Proposal

Team Members:

Harshitha Girish - 011809582
Mahesh Kumar Srinivas - 011845961
Sharath Kumar Karnati - 011852253
Sreehari Guruprasad – 011809587

Project Statement:

The main idea of this project is to develop an application for resume/CV parsing. The system will automatically extract key information and simplify the recruiting process from an HR professional point of view. By utilizing Natural Language processing (NLP) and Machine Learning, we will be able to extract crucial details like contact information, education background, work experience and skills.

Algorithm selected:

Natural Language Processing (NLP)
Named Entity Recognition (NER)

Named Entity Recognition (NER) is a technique in natural language processing that identifies and classifies key information in text into predefined categories, such as names, locations, and companies. NER automates the extraction of essential data from large texts. We are planning to use Spacy to train the NER model

Methodology:

1. Pdf Data Collection.
2. Converting pdf data into text data using *pymupdf.*
3. Splitting data into training and testing datasets for cross validation.
4. NLP Named Entity Recognition Model training using spacy.
5. Evaluation and Testing.

Selected Data Set:

We are using a dataset that has 2400 resumes from Kaggle. The data set consists of:

1. ID – id and names of the resume pdfs.
2. Resume String – All the text present in the resumes in string format
3. Resume HTML – Same data is in the html format
4. Category – Category of the resume.

Link:
https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset

Team Roles:

- **Harshitha -** Responsible for acquiring, cleaning, and pre-processing resume dataset from Kaggle.
- **Sharath, Sreehari** – Responsible for implementing NER, and training the machine learning model.
- **Mahesh:** Responsible for testing, analyzing results
- **All** – Documentation, reporting and presentation.

Even though the above delegation work indicates the primary distribution of work, all the team members will be contributing equally to each task at hand. We are collaborating as a cohesive team, sharing project responsibilities equally, and working together to achieve optimal outcomes. Each team member actively contributes their ideas and participates wholeheartedly in the project discussions, ensuring a comprehensive and inclusive approach.

Timeline:

Weeks 1-2: Our initial focus will be on acquiring relevant data from Kaggle, a reputable platform for datasets, and conducting preprocessing tasks to ensure the dataset is clean and ready for analysis.

Weeks 2-3: During this phase, we will delve into implementing Natural Language Processing (NLP) functionalities, crucial for extracting meaningful insights from the resume dataset. Additionally, we will dedicate time to training our machine learning model using processed data.

Weeks 3-4: With the model trained, we will shift our attention to conducting comprehensive evaluation experiments. This phase is pivotal as it allows us to gauge the performance of our model and make any necessary adjustments. Subsequently, we will meticulously analyze the experiment results to derive valuable insights.

Week 5: As we approach the final week, our primary focus will be on documentation and presentation preparations. We will document our methodologies, findings, and outcomes while we craft a presentation to effectively communicate our project's objectives, methodology, and results.