**CSCE 5300 Big Data and Data Science.**

**Project Proposal.**

**Natural language Processing for Resume Screening.**

**Team Members:**

1.Litesh Perumalla.

2.Chinmai Kaveti

3. Pavan kalyan Natukula.

4. Arun Govindgiri.

**Overview**

The goal of this project is to create a resume screening system which works with Natural Language Processing (NLP) and a Machine Learning Model. This project will extract relevant information from resumes, categorize candidates based on criteria, and rank their appropriateness.

For data processing, we will use Hadoop and Cassandra to store and manage data. We use Natural Language processing libraries for extracting information. For data analysis and machine learning, we will leverage Spark and ML lib and Sage Maker will be used for visualization and model training.

**Objectives:**

**Project 1:** Process resume data using Hadoop and store structured information in Cassandra.

**Project 2:** Analyze processed data using Spark, apply ML models for ranking candidates, and visualize the results in Sage Maker.

By using the analysis, we try to match the resume and job description to get a perfect candidate for the company.

**Tools and Technologies:**

**Project 1:**

Data Processing: Hadoop, Cassandra.

Data Analysis: Spark.

**Project 2:**

Visualization & Model Training: Sage Maker, Spark Mlib.

Natural Language Processing: NLP libraries such as NLTK and BERT for text extraction and classification.

**Dataset Selection:**

We will use publicly available datasets for resumes and job postings:

Resumes Dataset: [Hugging Face Resume Dataset] – structured (CSV).

(<https://huggingface.co/datasets/brackozi/Resume>)

Job Postings Dataset: [Bright Data Job Postings Dataset] – structured (CSV).

(<https://brightdata.com/cp/datasets/browse/gd_lpfll7v5hcqtkxl6l?tab=sample&camp=plg>)

The important attributes of the job posting dataset are:

url: The url for the job posting

job posting id – The unique attribute for each job.

Job title: The title of the job.

Company name: The company of where the job posting is available.

Job location: The location where the job is available.

Job Summary: The description of the job

Job seniority level: The level of the job in the company.

Job employment type: The type of employment the company offers.

Job industries: The type of industry the company has business.

Job base pay range: The minimum pay for the job.

Company url: The company portal of the job.

Country code: The country code where the job is present.

The attributes for Resume dataset are:

Category: The role of the job that company offers.

Resume: Parsed data of all the information in a resume.

**Methodology:**

**Project 1:**

We will extract text data from resumes and job postings. Then, we will process and clean this data using Hadoop. Finally, we will store the structured information in Cassandra so it can be queried efficiently.

**Project 2:**

We will Utilize Spark’s for largescale analysis. Then, we apply NLP techniques to extract information and Implement ML models to rank candidates. Later, we visualize using Sage Maker to perform analysis by the following insights provided from it.

**Deliverables:**

**Project 1:**

We will Process and structure resume and job listings data stored in Cassandra.

**Project 2:**

We will implement NLP for extracting resume data, perform text processing and use ML model for candidate ranking. After that, we'll use Sage Maker to analyze the data and create visualizations to present the results.

**Timeline (12 Weeks):**

**Weeks 1-7:** We plan to use data processing and store the data using Hadoop and Cassandra.

**Weeks 8-12:** We will implement data analysis, NLP, machine learning, and visualizations in Sage Maker.