Resume Scanner

DATS 6312: Natural Language Processing
Fall 2023
Instructor: Amir Jafari

Project Proposal

Jadhav, Shubham and Ramagiri, Jahnavi
Department of Computer Science
School of Engineering
George Washington University

November 5, 2023

1 Project Summary

The problem selected for this project is to build a resume scanner tool. The primary motivation for this project is to address the challenges candidates face when applying for job roles, specifically the time-consuming process of shortlisting candidates based on their resumes. The resume scanner tool, aims to simplify and expedite the job application process, benefiting both candidates and recruiters. Through the use of advanced NLP techniques, it seeks to provide more accurate and efficient resume analysis and job role matching. This aims to streamline this process by leveraging NLP techniques to achieve the following objectives:

1. Resume - Job Description Similarity Score: Calculate the similarity score between a candidate's resume and a job description to assess how well the resume matches the job requirements.

2. Multi-class Resume Classification: Categorize resumes into predefined classes or job roles, making it easier for candidates to identify suitable job roles for their resume.

3. Resume Analysis: Provide comprehensive analysis and insights into a candidate's resume, including key skills, qualifications, and experience.

2 Dataset

Source Link: https://github.com/florex/resume_corpus/tree/master

The multi-labeled dataset of resumes labeled with occupations. The resume files have the extension ".txt" and the corresponding labels are in a file with the extension ".lab".

This dataset contains 3 files :

1. resumes_corpus.zip: This file contains a set of resumes files with the extension ".txt" with the corresponding list of labels in a file with the extension ".lab"

2. resumes_sample.zip : This file represents the data set of resumes in a single text file. Each line of the file contains information about a text resume. Each line has 3 fields separated by ":::". The first field is the reference id of the resume; the second field is the list of occupations separated by ";" ; and the third field is the text resume.

3. normalized_classes : This file contains the associations between the occupations as written by the experts and their corresponding normalized form.

3 Key concepts

1. Pretrained NLP (Transformers Based Networks)

2. Recurrent Network (Any RNN model LSTM, GRU)

4 Model Customization

The project will involve fine-tuning pretrained models for specific tasks such as document similarity and classification. While we will initially use off-the-shelf pretrained models, fine-tuning and customization will be essential to adapt the models to the unique requirements of resume analysis.

5 Packages

1. Pretrained PyTorch: As our primary deep learning framework for building and customizing NLP models.
2. Pandas: For data preprocessing, manipulation, and organization.
3. NumPy: For numerical operations and array handling.
4. Scikit-Learn: To assist in machine learning tasks, such as classification and model evaluation.
5. Seaborn: For data visualization and performance analysis.

6 NLP Tasks

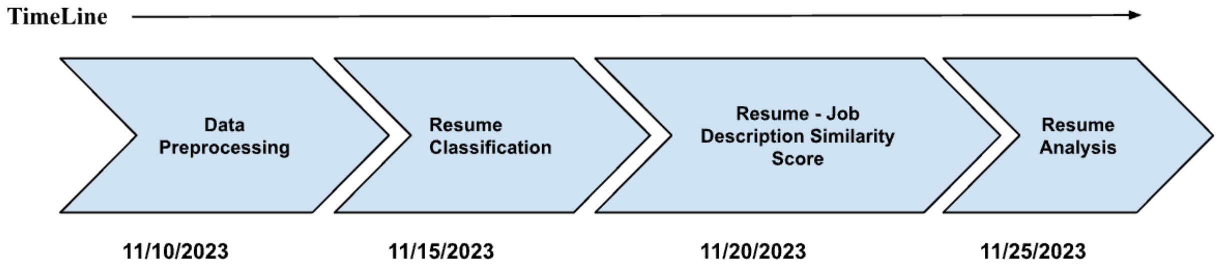
1. Document Similarity: To compute the similarity score between a candidate's resume and a job description.
2. Document Classification: To categorize resumes into predefined job roles or classes.
3. Resume Analysis: Extracting relevant information from resumes, such as key skills, qualifications, and experience.

7 Performance Metrics

We will judge the performance of our model using the following metrics:

1. Similarity
    - (a) Jaccard
    - (b) Cosine
2. Classification
    - (a) Confusion matrix
    - (b) F1 score
    - (c) Recall
    - (d) Precision

8 Timeline


TimeLine
Data Preprocessing
Resume Classification
Resume - Job Description Similarity Score
Resume Analysis
11/10/2023
11/15/2023
11/20/2023
11/25/2023