Resume Parsing Using NLP - Proposal

Abstract

In this project, we propose a natural language processing model that can extract useful information from unstructured 'Resume documents'. Specifically known as Resume parsing, this technique extracts useful information from resumes for further processing such as resume ranking and selection. During the recruitment process companies get thousands of resumes, using traditional manual methods and by providing unique resume templates to applications. The model aims to identify relevant document sections and corresponding specific information at a lower hierarchy level. To achieve this, we will use various tools such as docx2txt and pdfminer to extract text from docx, doc and pdf formats. Additionally, we will employ SpaCy for tokenization and apply its named entity recognition (NER) and part-of-speech (POS) tagging capabilities to identify relevant terms and entities in the text.

Keywords*: Natural Language Processing, Resume Parsing, Named Entity Recognition, Part-of-speech(POS) tagging, Text Classification*

Introduction

Resumes are a great example of unstructured data, as each resume has its unique formatting style and data blocks, and there are many forms of data formatting. During the recruitment process companies get thousands of resumessing by using traditional manual methods and by providing unique templates to application. To ensure that recruiters can quickly identify the most relevant information, it is essential to develop a tool that can extract useful information from these unstructured resumes[1].For the project we researched on few related work and came to a conclusion to use tools like docx2txt and pdfminer to extract pdf, docx and doc file format as most resume documents are in either of these formats. Additionally we will use SpaCy and NLTK for tokenization and apply its named entity recognition(NER) and part-of-speech(POS) tagging capabilities to identify relevant terms and entities in the text.

In this project, we propose to use SpaCy for entity recognition on resumes and experiment with various NLP tools. The model aims to identify relevant document sections and corresponding specific information at a lower level.

The problem we aim to solve is the extraction of useful information from unstructured resumes. Specifically, we aim to develop a model that can identify relevant document sections and corresponding specific information at a lower hierarchy level.

Related Work

This section gives a brief overview of previous research done in the field of natural language processing and Resume Parsing.

Resume Information Extraction with A Novel Text Block Segmentation Algorithm [2]

In this paper the authors have discussed resume parsing using neural networks-based classifiers and distributed embeddings. The proposed pipeline leverages position-wise line information and integrated meanings of each text block to segment a resume into predefined text blocks and perform named entity recognition within labeled text blocks. The study confirms the effectiveness of the proposed method through comparative evaluations of sequence labeling classifiers and comparison with three publicized resume parsers.

Resume Information Extraction with Cascaded Hybrid Model.[3]

In this Paper authors have proposed a cascaded information framework to perform resume information extraction. Instead of searching the entire resume, as is done with the flat model, a resume will firstly be segmented into consecutive blocks attached with labels indicating the information types, followed by the identification of detailed information within a specific block.

Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing[4]

In this paper authors talks about the ongoing approaches and relevant to those approaches proposes a system for resume parsing using deep learning models such as Convolutional Neural Network (CNN), Bi-LSTM (Bidirectional Long Short-Term Memory) and Convolutional Random fields(CRF). CNN model for classification of different segments in a resume. Bi-LSTM for sequence labeling in order to tag different entities. Pretrained GloVe model is used for word embedding.

Information Extraction From Free-Form CV Documents in Multiple Languages[5]

In this paper authors propose two natural language processing models for extracting useful information from multilingual unstructured CV documents. The approach employs the transformer architecture and its multilingual implementation of the encoder part in the form of the BERT language model. The models were trained and tested on a large dataset of manually annotated CV documents, which achieved high scores on standard accuracy measures.

Problem Statement

The problem we aim to solve is the extraction of useful information from unstructured resumes. Specifically, we aim to develop a model that can identify relevant document sections and corresponding specific information at a lower hierarchy level.

Resumes do not have a fixed file format, and hence they can be in any file format, such as .pdf, .doc, or .docx. So, our challenge is to read the resumes and convert them into plain text.

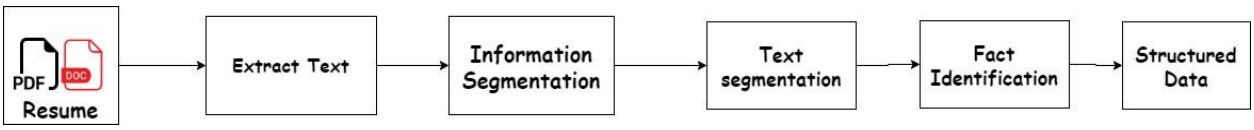
Objectives

The objectives of this proposed system are :

- Successfully develop a model capable of identifying relevant document sections and corresponding specific information at a lower hierarchy level..
- To develop the model and to gain a better understanding of the various concepts of natural language processing.

Workflow

Pipeline for the project is presented as the following figure.



Steps Involved in the project.

1. Read documents and Extract Text: This step involves reading pdf, docx and doc file type and gives extracted plain text. For this project we will be using pdfminer for pdf document and doc2txt for reading doc and docx file type as most Resumes are either in doc format or pdf format.
2. Information segmentation: This step involves extracting information such as personal information, Education, work information and skills. For this we will be using NLTK and spaCy.
3. Fact Identification: This step involves Named Entity Recognition which is the task of identifying key information and categorizing the information in text. For this, we propose to use spaCy.
4. Structured Data: Finally, the generated output will be in the form of structured data JSON or csv.

Task Division

The task is equally divided between the two members and both members will work interchangeably without any specific task allocation. This approach allows for flexibility and ensures that all responsibilities are shared equally.

References

[1] "Writing Your Own Resume Parser," *Omkar Pathak*. https://www.omkarpathak.in/2018/12/18/writing-your-own-resume-parser/ (accessed Mar. 02, 2023).

[2] S. Zu, X. Wang, and S. Darren, "Resume Information Extraction with A Novel Text Block Segmentation Algorithm," *Linguistics*, vol. 8, pp. 29–48, Oct. 2019, doi: 10.5121/ijnlc.2019.8503.

[3] K. Yu, G. Guan, and M. Zhou, "Resume Information Extraction with Cascaded Hybrid Model.," presented at the ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Jan. 2005. doi: 10.3115/1219840.1219902.

[4] C. H. Ayishathahira, C. Sreejith, and C. Raseek, "Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing," in *2018 International CET Conference on Control, Communication, and Computing (IC4)*, Jul. 2018, pp. 388–393. doi: 10.1109/CETIC4.2018.8530883.

[5] D. Vukadin, A. S. Kurdija, G. Delač, and M. Šilić, "Information Extraction From Free-Form CV Documents in Multiple Languages," *IEEE Access*, vol. 9, pp. 84559–84575, 2021, doi: 10.1109/ACCESS.2021.3087913.