VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

Mini Project Report

On

Resume Parser in AI using NLP-NLTK

Problem statement

Under the Guidance of

Professor: Mrs. F. M. Inamdar

SUBMITTED BY

GR.NO	ROLL.NO	NAME
22010910	332002	Niraj Amrutkar
22010070	332010	Chirag Chawade
22010826	332011	Harsh Chawla
22010416	332025	Chetan Ingle

Resume Parser in AI using NLP-NLTK

Abstract:

A resume parser is a deep learning/AI framework that identifies complete information from resumes, analyses, store, organize, and enriches it through its taxonomies. Resume parsing software makes the hiring process quick and more productive.

In many major companies, recruiter doesn't have time to view or read full resume to select the person for a specific domain or skill. So, for that we designed a system where user upload resume docx file of person and required skills or education and in returns system returns the how much skills and education matches with required skills and education.

This entire system is based on The **Natural Language Toolkit** (NLTK) which is an open-source Python library for Natural Language Processing.

In order to achieve the desired goal, the entire process divided in 3 basic segments. The first segment consists function which extracts the text from given document file. And in second segment there is a function which extracts the skills and education from the extracted words from the document. And finally compares with given keywords of skills and education to founded keywords. And returns how much that candidate matches with recruiter's eligibility.

Keywords:

NLP: Natural Language Processing.

NLTK: Natural Language Toolkit

Introduction:

As a recruiter, have you found it hard to explain why the recruitment drive for a role is slow? Or finding it difficult to collate numbers on even surface level data? You are not alone. On average, every job posting receives almost 200 applications. A recruiter's job is already hard, trying to find the right candidates for their organization while balancing the pipeline of candidates coming in and sorting through documents apart from resumes.

Recruiters use resume parsers in order to streamline the resume and applicant screening process. Parsing technology allows recruiters to electronically gather, store, and organize large quantities of resumes. Once acquired, the resume data can be easily searched through and analysed.

Literature Survey:

Authors	Title of the Paper	Proposed Methodology	Positive Points	Discussion

Bichitra Mandal	Architecture of efficient word processing using Hadoop MapReduce for big Data Applications	Hadoop--Map Reduce, Hadoop Distributed File System	To--count the number of consecutive words and repeating lines	Time overriding method%
D.Celik et al.	Towards an Information Extraction system based on ontology to match resumes and jobs	Ontology based resume parser for finding Resume	Plain text resume into ontology form by Ontology Knowledge Base (OKB)	System calculates percentage completeness depend upon work experience, education etc.%
Sourav Madhesiya, Pranay Lonare, Tanuja Shelke, Swati Lokare, Vilas Khedekar	Resume Parsing and Processing Using Hadoop	Hadoop	Automatic extraction of desired information in vigorous classified sections helped in the selection process of effective document which may leads to achieve beneficial merits like automatic parsing, filtering and processing of respective document.	Apache Hadoop is a open source framework for storing, processing and analysing large amount of multi structured information in a distributed environment
OZGUR ULUSO Y et al.	Research Issues in Real-Time Database Systems	Transaction /query processing, data buffering, CPU and IO Scheduling	For satisfying Timing constraints in real time Database applications	Replacement of conventional database systems to eliminate disk access delays. %
Wen Hua et al.	Understand Short Texts by Harvesting and Analysing	Chain model, pair wise model, Monte Carlo method	For maintaining accuracy and efficiency in short texts to extract semantic Knowledge	For removal of ambiguity in short texts. %

	Semantic Knowledge			

Methods:

Previously Resume Parsing done using NLP, Entity Extraction Process and Big Data Tools and Using Hadoop. Here we are using NLTK which is Library of NLP.

The key reason I chose NLP for Hireability is that it can handle massive volumes of data in seconds or minutes that would take days or weeks to analyze manually. NLP technologies can instantly scale up or down to match demand, allowing us to have as much or as little processing capacity as per requirement.

Doc2txt

doc2text extracts higher quality text by fixing common scan errors Developing text corpora can be hard for everyone. Much of the text data we are interested in as scientists are locked away in pdfs that are poorly scanned. These scans can be off kilter, poor resolution, have a hand in them… and if you OCR these scans without fixing these errors, the OCR doesn't turn out so well. doc2text was created to help researchers fix these errors and extract the highest quality text from their pdfs as possible.

NLTK

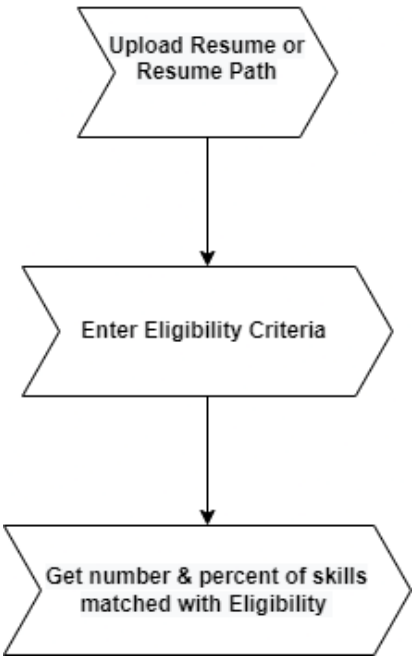
NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language.

Features used from NLTK Library

n-grams: N-grams analyses are often used to see which words often show up together. An n-gram is a contiguous sequence of n items from a given sample of text or speech. In the text analysis, it is often a good practice to filter out some stopwords, which are the most common words but do not have significant contextual meaning in a sentence (e.g., "a", "the", "and", "but", and so on). NLTK provides us a list of such stopwords. We can also add customized stopwords to the list.

System Architect:



Future:

Resume parsers are already standard in most mid- to large-sized companies and this trend will continue as the parsers become even more affordable.

A qualified candidate's resume can be ignored if it is not formatted the proper way or doesn't contain specific keywords or phrases. As Machine Learning and Natural Language Processing get better, so will the accuracy of resume parsers.

One of the areas resume parsing software is working on expanding into is performing contextual analysis on the information in the resume rather than purely extracting it. One employee at a parsing company said “a parser needs to classify data, enrich it with knowledge from other sources, normalize data so it can be used for analysis and allow for better searching.”

References:

- https://www.mecs-press.org/ijitcs/ijitcs-v10-n9/IJITCS-V10-N9-3.pdf
- http://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/9480/14101061,14101171_CSE.pdf?sequence=1&isAllowed=y
- http://www.ierjournal.org/pupload/vol2iss7/Resume%20Parsing%20And%20Processing%20Using%20Hadoop.pdf
- https://www.ijitee.org/wp-content/uploads/papers/v9i7/F4078049620.pdf
- https://en.wikipedia.org/wiki/R%C3%A9sum%C3%A9_parsing