VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

Mini Project Report

On

Resume Parser in AI using NLP-NLTK

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. 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. Overworked recruiters are constantly searching for methods that will speed up procedures. However, it is of little use to believe that a system will expedite things only to discover that ideal candidates have been overlooked. 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:

Nowadays many hiring processes are done to get that skilled, qualified candidate and for that many students apply for a single job. And this is very hectic procedure for HR to view or read each and every resume fully. Also, new generation students have loss of patience to view their result as soon as possible. So, to avoid such type of hard work we designed this system which is do smart work and consume very less time comparatively old-fashioned resume checking.

As everyone know that hiring process is very slow and very lengthy for the recruiters and for that we have designed a small project, where a recruiter can give some eligibility criteria and the resume of some candidates and our system returns how much eligible that candidate is.

Free-form internet documents are transformed into organised sets of data via resume parsing technology. This is accomplished by carefully examining resumes and extracting relevant data. The information is immediately saved, and recruiters can use search tools to find specific abilities and expertise.

Every day, this procedure saves recruiters hours, allowing them to devote more time to engaging with and interviewing candidates. For both recruiters and applicants, resume parsers reduce a lot of manual tasks, accelerating the screening procedure. Unsuitable applicants are alerted via automatic responses, and suitable prospects are rapidly found.

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 determine the number of recurring lines and consecutive words	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	By using the Ontology Knowledge Base, transform a plain-text resume into an ontology (OKB)	Based on factors like work experience, education, etc., the system calculates the percentage of completion.
Sourav Madhesiya, Pranay Lonare, Tanuja Shelke, Swati Lokare, Vilas Khedekar	Resume Parsing and Processing Using Hadoop	Hadoop	Automatic extraction of desired information from active classified sections assisted in the selection of useful documents, which may have led to the achievement of advantageous merits like automatic document parsing, filtering, and processing.	An open source framework called Apache Hadoop is used to store, process, and analyse large amounts of multi-structured data in a

				distributed setting.
OZGUR ULUSOY et al.	Research Issues in Real-Time Database Systems	Transaction /query processing, data buffering, CPU and IO Scheduling	To meet timing restrictions in real-time database applications	replacing current database systems with new ones to get rid of disc access delays.
Wen Hua et al.	Understand Short Texts by Harvesting and Analysing Semantic Knowledge	Chain model, pair wise model, Monte Carlo method	To preserve precision and effectiveness when extracting semantic knowledge from brief texts	to eliminate uncertainty in brief texts.

Methods:

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

NLP technologies enable us to have as much or as little processing capability as necessary by instantaneously scaling up or down to match demand.

Doc2txt

We have used this function to check each and every word in given docx file which is inbuilt function NLP-NLTK library. This function checks words more accurately than OCR. And here in the resume parsing to get correct and deserved candidate we used NLP library. This makes our system more accurate and trustable for companies which got too much resume and candidates.

Features used from NLTK Library

n-gram extracted sequentially from a body of text make up an n-gram. The tools in this collection can be used to generate, show, summarise, and "babble" n-grams. Very efficient C code, which can even be written as its own standalone library, handles the "tokenization" and "babbling". Essentially, the babbler is a Markov chain. A vignette with thorough examples of "workflows" and details on the utilities provided by the package are also included in the package.

N-gram studies are frequently used to determine which words are frequently found together. A contiguous group of n items from a given sample

of text or speech make up an n-gram. It is frequently a useful practise in text analysis to eliminate some stopwords, which are the most prevalent words but lack important contextual meaning in a sentence (for example, "a," "the," "and," "but," and so on). We are given a list of these stopwords by NLTK.

Tokenization

Tokenization is the first step in text analysis. It makes it possible to identify the main ideas of the text. The fundamental units are tokens. Tokenization is advantageous since it breaks down a text into manageable pieces. Internally, spaCy decides if a "." is a period and splits it into tokens or whether it is an abbreviation, such as "B.Tech.," in which case it is not split. We may use word tokenization or sentence tokenization depending on the issue.

a. Sentence tokenization: breaking a paragraph up into a number of sentences using the sent tokenize () function.

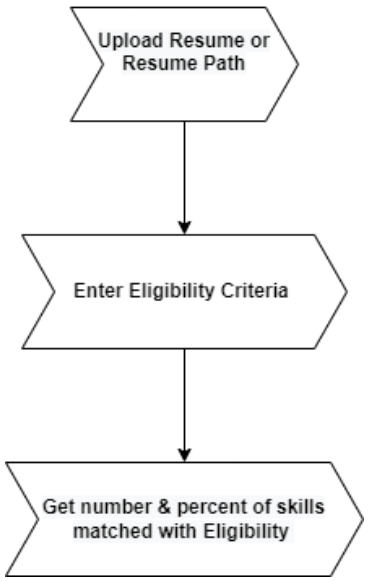
b. Word tokenization: Breaking a statement down into a list of terms using the word tokenize () method.

Removing Stop words

In NLP, cleaning the data is crucial to removing noise. The most often occurring repeating words in a text that provide no valuable information are called stop words. A collection of terms that are regarded as stop words in English is available in the NLTK library. These are just a few of them: [I, no, nor, me, mine, myself, some, such we, your'd, your, he, ours, ourselves, yours, yourselves, you, you're, you've, you'll, most, other].

A well-liked library for deleting stop words is the NLTK library, which gets rid of roughly 180 of them. We can create a unique set of stop phrases for particular challenges. We can quickly add any new word to a collection of terms by using the add approach.

System Architect:



Future:

Most mid- to large-sized businesses now use resume parsers, and as they become more accessible, this tendency will continue.
A resume for a good candidate may be overlooked if it is improperly formatted or lacks certain terms or phrases. The precision of resume parsers will increase as machine learning and natural language processing get more sophisticated. Performing contextual analysis on the information in the resume rather than just extracting it is one of the areas resume parsing software is striving to develop into. "A parser needs to classify data, enhance it with knowledge from other sources, normalise it so it can be utilised for analysis and allow for improved searches," one employee of a parsing business stated.

Here we have implemented resume parsing for the skill section but in future it can be implemented in each and every part of resume like hobbies, education, experience, etc.

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*