AI Approaches for Automatic Resume Prescreening

Data Extraction:

Input Formats:

- PDF
- Doc
- Docx
- Image

Approaches to extract data:

PDF

- Row Based
- Column Based

Row Based:

Open Source libraries:

There are multiple open source libraries to extract data from pdf document.

- OCR
- PDF Plumber
- PyMuPDF
- PDF Miner
- PyPDF2

Column Based:

If we extract the data using the open source libraries for these kinds of document data is merging since all available libraries read document row wise, so we need to do different kind of approach for this as mentioned below.

Approach:

- Object Detection
- OCR

Object Detection:

First we will convert any kind of resume to image, then we will do object detection using these open source libraries.

- aspose.words (doc to image)
- pdf2image (pdf to image)

Object detection is helps in getting the bounding boxes of the parts of resume whether it is of any format.

OCR:

From the bounding boxes from object detection, we will use OCR technique to extract the data from these chunks (bounding boxes).

Docx:

To extract data from Docx we can use open source libraries mentioned below.

- docx2txt

Images:

There are multiple approaches to extract data from documents in image format. The approaches best suitable to our need.

- OCR
- Object Detection

Note:

As we discussed for first go we will go with row based resumes by using open source libraries to extract data from word and pdf formats.

Information Extraction:

Once we extract the data from resumes from previous step, then we need to extract the required information from these which helps us to provide end result.

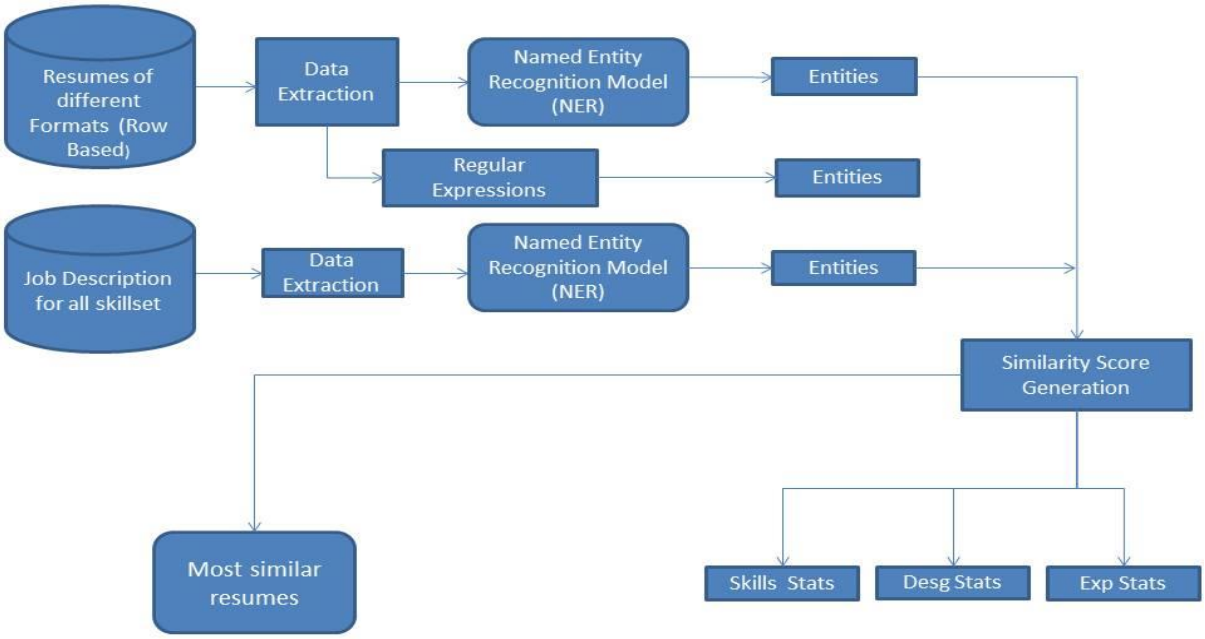
To do the information Extraction we will use one approach mentioned below.

Named Entity Recognition (NER):

Named Entity Recognition is technique in Natural Language Processing which helps in extraction of information like Name, Education, Designation, etc.,.

Similarity Computation Engine:

- Once we extract the required information from Resumes and Job Description we will build a Similarity Computation Engine to get the matched resumes for that particular JD along with the percentage of similarity.
- We will display the statistics for individual fields which we required.
- This helps to recruiter to finalize the resumes for that particular JD.



Dependency:

- Need resumes of different skillsets and proper Job Description for all the Skillsets provided.
- JD should contain minimum and preferred qualification to provide weightage accordingly.
- Once we extract the information from resume based on the requirement we need to use some rules to get the final requirement mentioned below.
    - Companies (Tier 1, 2, 3)
    - Colleges (Tier 1, 2, 3)
- For these need information from respective department.

Challenges:

- Extraction of data from column based resumes.
- Getting Tier1, Tier2, Tier3 Companies and college details are a challenge as they are not available in open source and they also frequently gets updated.
- Getting company and respective duration is bit challenge.