ABV-IIITM Gwalior, India

Resume Parsing with custom NER Training

Submitted By

Aakar Srivastava	2018IMT-002
Abhay Chaurasiya	2018IMT-005
C.Dheena	2018IMT-026
Himanshu Pandey	2018IMT-038
K Mallikarjun	2018IMT-046
Puneet	2018IMT-074

Under the supervision of
Dr. Debanjan Sadhya

Table of Content

1. Introduction
2. Problem Statement
3. Theory
4. Working of NER
5. Methodology
5. Applications
6. Conclusion and Future Scope

Introduction

- **Resume Parsing** is conversion of a free-form resume document into a structured set of information suitable for storage, reporting, and manipulation by software. Resume parsing helps recruiters to efficiently manage electronic resume documents and find suitable candidates quickly.

- Many HR Professionals use CV and Resume Parsing tools to automate the storage, import and analysis of data and text on CVs and resumes.

- Resume Parsing could be boon to HR. With the help of NLP, an accurate and faster system can be made which can save days for HR to scan each resume manually.

Problem Statement

- As the number of jobs are increasing every day, specially in IT so screening resumes has become the need of the hour.

- We want to build a tool that can quickly analyze the resume and find relevant fields we are interested in.

- With thousands of resumes being sent to companies each day, a resume parser saves a lot of time and effort.

Theory

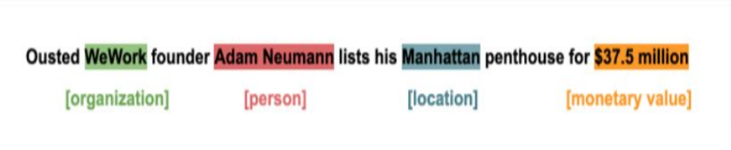
- **Named Entity Recognition (NER)**
  - Named entity recognition is a natural language processing technique that can automatically scan entire articles and pull out some fundamental entities in a text and classify them into predefined categories.

Entities may be:

•Organizations
•Quantities
•Monetary values
•Percentages and more.
•People’s names
•Company names
•Geographic locations (Both physical and political)
•Product names
•Dates and times
•Amounts of money
•Names of events

Theory

- In simple words, Named Entity Recognition is the process of detecting the named entities such as person names, location names, company names, etc from the text.
- It is also known as entity identification or entity extraction or entity chunking.



Ousted WeWork founder Adam Neumann lists his Manhattan penthouse for $37.5 million

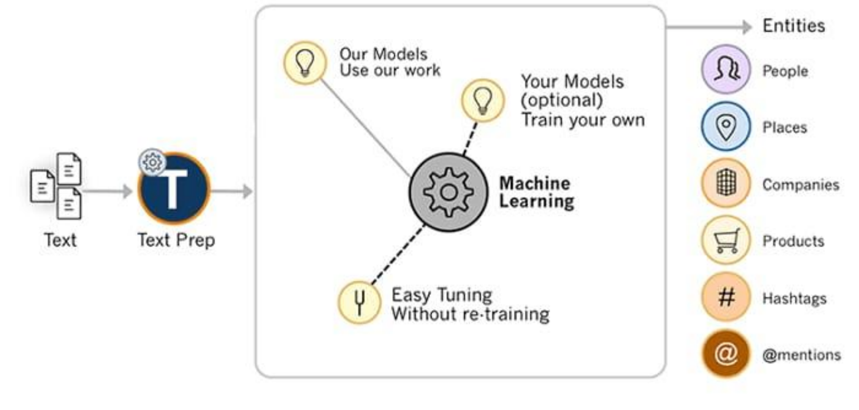
Working Of NER

- Named Entity Recognition NER works by locating and identifying the named entities present in unstructured text into the standard categories such as person names, locations, organizations, time expressions, quantities, monetary values, percentage, codes etc.

- spaCy NLP Library comes with an extremely fast statistical entity recognition system that assigns labels to contiguous spans of tokens.

Cont..

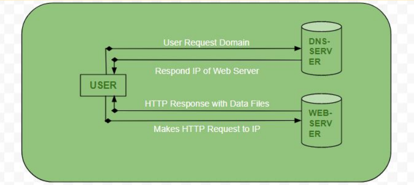
Basic Overview



Cont..

Client Server Model

- We are using client-server architecture for our system model.
- A model in which the server hosts, delivers and manages most of the resources and services to be consumed by client.
- A major motivation to use such architecture is
  - to modularized the client and server into separate services.
  - to have centralized system with all data in a single place.



Cont..

Client side

- All the user interaction is done through this side
    - The user is able to upload the resume from their Browser.
    - After the uploading, user/HR can get a detailed structured report of the candidate’s details with likeness prediction for a role based on the matching skills.
- We are using React.js Library for maintaining the Client side and sending requests to the backend.

Cont..

Server Side

- All the major internal working including various text processing is done in this side
    - The interaction with the client side is done by using API, it acts as the communication bridge between client and server.
    - The client side will send Resume as input to APIs, which server further responds to the client after running the algorithm and returns the results.
- We are using Flask framework for handling the APIs.

Cont..

Match Score

- We are allowing recruiters to add skills and get a percentage of match skills. This can help them filter out hundreds of Resumes with just one button.

- From a resume, we check which requirement skill are being satisfied and which are not. We then divide the number of satisfied skills to the total number of required skills to get the match score.

- Using this, a threshold can be set to automatically prune those resumes which have less skills which the company needs.

Methodology

DATASET -

Dataset is scarce in this area of research. We have chosen the DataTurks Dataset consisting of 200 resumes.

Each resume entity consists of a string containing the text form of resume and a dictionary of the entity labels and positions of that label.

Since new datasets could arise, we decided to store the dataset in a text file and convert the text to a **pickle file** for training easily. A script was written for the same by us.

Pickle file stores objects which can be used for other projects/files directly. This saves time

Methodology

- Training our custom NER model.
- Extracting data from PDF files for testing.
- Testing our NER Model using extracted data

Training Of NER Model

We use spaCy NLP Library to train a model with our NER data. spaCy’s tagger, parser, text categorizer and many other components are powered by statistical models. Every “decision” these components make – for example, which part-of-speech tag to assign, or whether a word is a named entity – is a prediction based on the model’s current weight values. The weight values are estimated based on examples the model has seen during training.

Training Of NER Model

Every iteration we shuffle the training data and feed it to the model which updates its weights. The training process with spaCy does take some time. We therefore limited our number of iterations to 100.

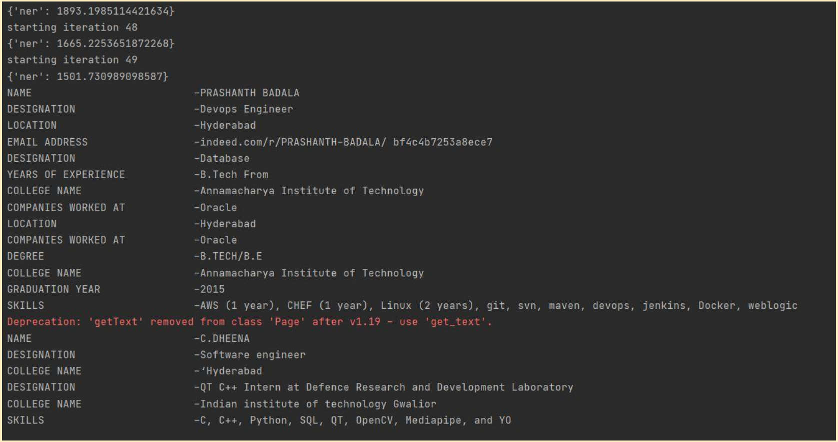
After 100 iterations, our model is ready for testing. We feed the test data and gather the outputs. To test a fresh test case, we convert the pdf file to raw text using Fitzz package.

Testing

- We have used our batchmates resumes and checked if those got parsed through our NLP model using the website we designed.
- We observed that all the skills were matching with the original resumes.
- We also matched the skills of the candidate with the required skills for the job role and if they get matched, then that candidates are eligible for that job role..

Result

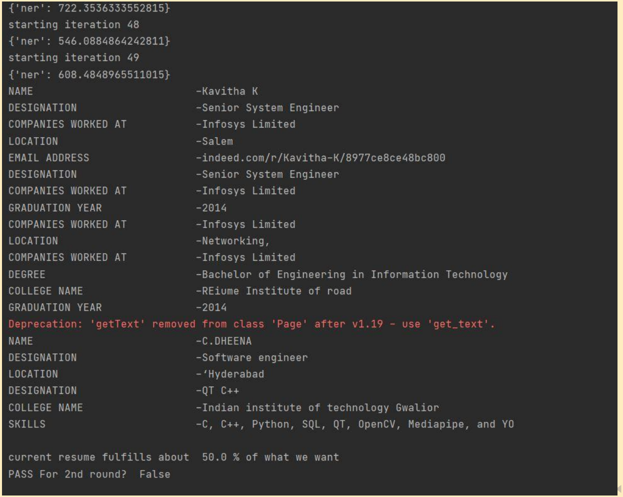
Testing for 100 iterations (50*2 separately)



```
{'ner': 1893.1985114421634}
starting iteration 48
{'ner': 1665.2253651872268}
starting iteration 49
{'ner': 1501.730989098587}
NAME                          -PRASHANTH BADALA
DESIGNATION                   -Devops Engineer
LOCATION                      -Hyderabad
EMAIL ADDRESS                 -indeed.com/r/PRASHANTH-BADALA/ bf4c4b7253a8ece7
DESIGNATION                   -Database
YEARS OF EXPERIENCE           -B.Tech From
COLLEGE NAME                  -Annamacharya Institute of Technology
COMPANIES WORKED AT           -Oracle
LOCATION                      -Hyderabad
COMPANIES WORKED AT           -Oracle
DEGREE                        -B.TECH/B.E
COLLEGE NAME                  -Annamacharya Institute of Technology
GRADUATION YEAR               -2015
SKILLS                        -AWS (1 year), CHEF (1 year), Linux (2 years), git, svn, maven, devops, jenkins, Docker, weblogic
Deprecation: 'getText' removed from class 'Page' after v1.19 - use 'get_text'.
NAME                          -C.DHEENA
DESIGNATION                   -Software engineer
COLLEGE NAME                  -‘Hyderabad
DESIGNATION                   -QT C++ Intern at Defence Research and Development Laboratory
COLLEGE NAME                  -Indian institute of technology Gwalior
SKILLS                        -C, C++, Python, SQL, QT, OpenCV, Mediapipe, and YO
```

Result

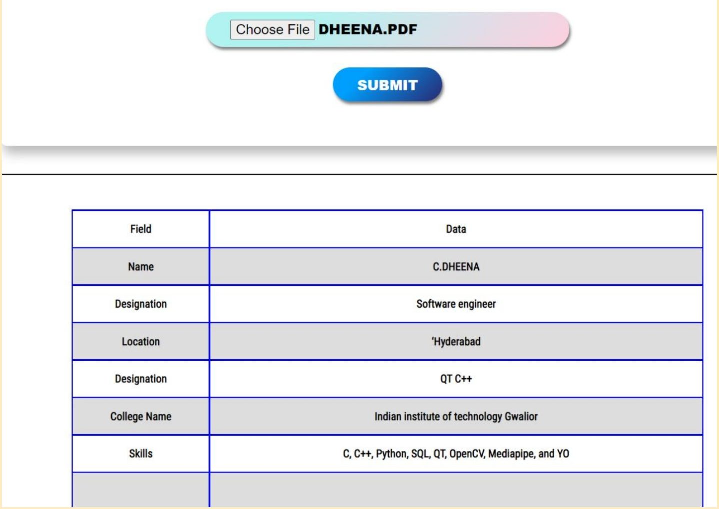
Testing for 100 iterations (50*2 separately)



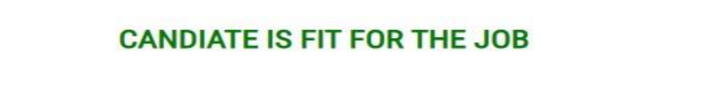
```
{'ner': 722.3536333552815}
starting iteration 48
{'ner': 546.0884864242811}
starting iteration 49
{'ner': 608.4848965511015}
NAME                          -Kavitha K
DESIGNATION                   -Senior System Engineer
COMPANIES WORKED AT           -Infosys Limited
LOCATION                      -Salem
EMAIL ADDRESS                 -indeed.com/r/Kavitha-K/8977ce8ce48bc800
DESIGNATION                   -Senior System Engineer
COMPANIES WORKED AT           -Infosys Limited
GRADUATION YEAR               -2014
COMPANIES WORKED AT           -Infosys Limited
LOCATION                      -Networking,
COMPANIES WORKED AT           -Infosys Limited
DEGREE                        -Bachelor of Engineering in Information Technology
COLLEGE NAME                  -REiume Institute of road
GRADUATION YEAR               -2014
Deprecation: 'getText' removed from class 'Page' after v1.19 - use 'get_text'.
NAME                          -C.DHEENA
DESIGNATION                   -Software engineer
LOCATION                      -‘Hyderabad
DESIGNATION                   -QT C++
COLLEGE NAME                  -Indian institute of technology Gwalior
SKILLS                        -C, C++, Python, SQL, QT, OpenCV, Mediapipe, and YO

current resume fulfills about  50.0 % of what we want
PASS For 2nd round?  False
```

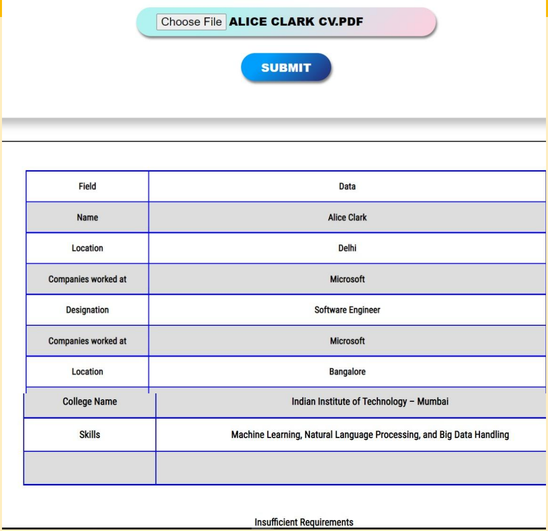
Result



Field	Data
Name	C.DHEENA
Designation	Software engineer
Location	'Hyderabad
Designation	QT C++
College Name	Indian institute of technology Gwalior
Skills	C, C++, Python, SQL, QT, OpenCV, Mediapipe, and YO



CANDIATE IS FIT FOR THE JOB



ALICE CLARK CV.PDF

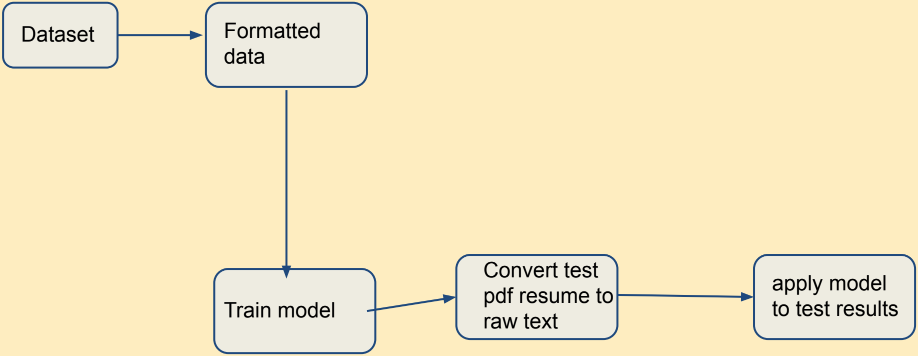
Field	Data
Name	Alice Clark
Location	Delhi
Companies worked at	Microsoft
Designation	Software Engineer
Companies worked at	Microsoft
Location	Bangalore
College Name	Indian Institute of Technology – Mumbai
Skills	Machine Learning, Natural Language Processing, and Big Data Handling

Insufficient Requirements

Candidate with lack of skills

Cont..

WorkFlow


Dataset
Formatted data
Train model
Convert test pdf resume to raw text
apply model to test results

Applications

Resume parsing can help ensure that you find the best candidate for an available position.

1) Resume parsing software can help you quickly sort resumes by searching for relevant keywords and qualifications automatically instead of manually searching each resume.

2) Because parsing software highlights a qualification you specify, your hiring team can easily identify the skills, qualifications and characteristics they want in an ideal candidate and maintain that standard throughout the hiring process.

Applications

3) Data from this parser could be used for future purposes.

Some purposes include - identifying behavioural characteristics, developing model that predicts a potential employee’s productivity rate, etc.

Conclusion

- In this project, we have used Name Entity recognition (NER) to create additional entities and then displayed them using custom colors. We have also visualized categories and skills distributions and allowed the user to add resumes directly which includes skills match percentage.

- It was a learning experience for us as we have never used spaCy in depth. We have also discovered various ways on how the project can be used to improve the hiring process in filtering out the perfect candidate for the job.

THANK YOU