# INFORMATION EXTRACTION FROM RESUME

**OVERVIEW:**

In this work, I have done a resume parser based on regex techniques. It can find phone number, skill set, educational qualification and experiences, by using given PDF resume.

**WORKFLOWS CONSIST OF 4 STEPS:**

* Introduction.
* Implementation
* Other Technologies
* Conclusion



**INTRODUCTION:**

The object of this POC is to tackle the required information from given dataset. The problem statement says that we have to extract Name of the organization, year of experience, Education/Institution, year of studies from the given dataset, my curiosity has led me to extract few more information such as phone numbers and skills.

**IMPLEMENTATION:**

Technologies and Platform used :( Google Colab, Python, pdfminer, regex techniques)

To make use of the functions in a module, you'll need to import the module with an import statement. In a Python file, this will be declared at the top of the code.

First step is to read all the files from particular folder for extracting required information from resumes. We need to split the lines in a resume to identify/ extract required information for that pdfminer (**PDFMiner** is a tool for extracting information from PDF documents. It includes a PDF converter that can transform PDF files into other text formats. It has an extensible PDF parser that can be used for other purposes than text analysis) tool is used so that it will be easy for pattern matching.

If we have resumes in document format we have to convert that into text format so that it will be easy to fetch the information. For conversion, you will have to import docx2txt library.

We need to remove all punctuations from resumes. Regex (**regex** or regexp for short, is a sequence of letters and symbols that defines a logical pattern. Strings of text can then be compared to the pattern in order to identify strings that match the logical pattern defined by the **regex**) has been used to extract phone numbers from resumes.

For finding education qualification, it is again based on keyword matching. Keywords should be mentioned in education.txt.

For skill set finding, we have a set of all skills in valid\_skill.txt file. The program compares each word in the resume and looks whether that word is there are not in valid\_skill.txt.

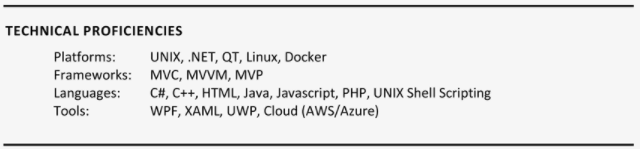
Experience are usually mentioned with dates in the format of Jan 2019 to Nov 2019 and some other format. The program looks for such formats and extract experience.

Combine all the extracted information for visualization purpose and store all these into data frame. And we will have to convert the dataframe into excel format for better visualization and understanding.

**OTHER TECHNOLOGIES:**

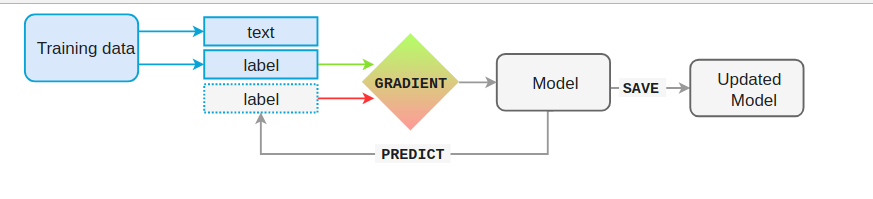
* **Named Entity Recognition: Approach 1**

It is used when you want a specific set of strings from the extracted regions. For example, consider the component of a resume below

****

Say we want only programming languages he’s good at. These type of problems can be solved using NER.

NERis an algorithm where it takes a string of text as an input (either a paragraph or sentence) and identifies relevant nouns (people, places, and organizations) and other specific words. For example, in a given resume if you want to extract, only Name and Phone Number using NER would make our job much easier.  It can be achieved by deep learning. This is because of a technique called word embedding’s, which is capable of understanding the semantic and syntactic relationship between words. There’s some pre-processing involved for most of the programs that involve data, even this Resume Parsing includes one.

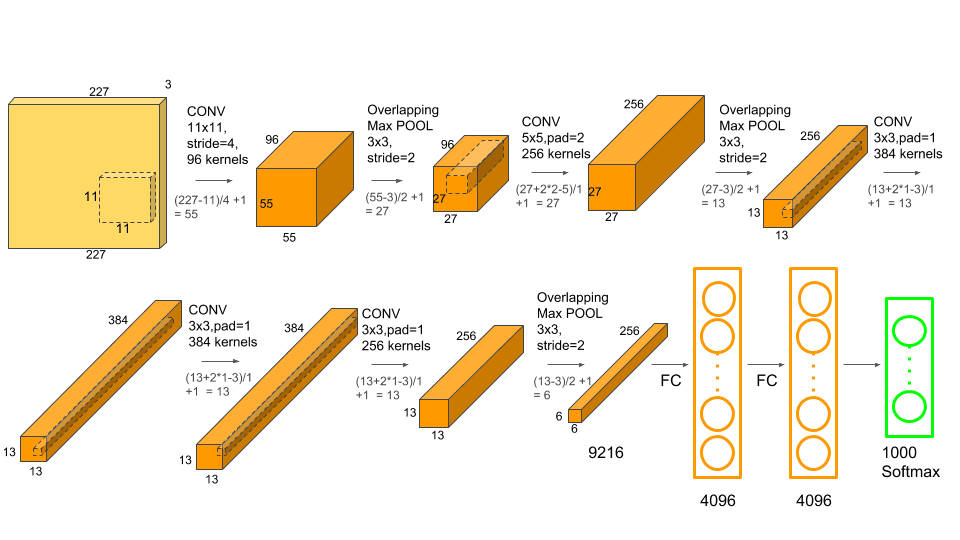


In most of the cases, resumes are saved as PDFs or DOCX, hence to make it easy,

* **Object Detection and OCR: Approach 2**

Resumes are basically live templates consisting of different sections. Meaning most of the sections in the Resume are similar but they are organized in different formats. The various components of Resumes are [Career Objective, Educational Background, Work Experience, Leadership, Publications, etc.]. To extract these components we consider these as the objects and detect them through an object detection algorithm. There are several challenges that need to be addressed such as table extraction (sometimes components like education background are added in tables), font-variation, template variation, etc.

We’ll now delve into how object detection is used to extract components from resumes. To achieve this Convolution Neural Networks (CNNs) are commonly used. There are several applications based on CNNs that achieved a state of the art performance for Image Classifications and Segmentation problems. Below is an image of a simple CNN,

****

For resume parsing using Object detection, page segmentation is generally the first step. The main goal of page segmentation is to segment a resume into text and non-text areas. Later, we extract different component objects, such as tables, sections from the non-text parts. Unlike traditional rule-based methods where a lot of parameters are involved, the main goal of learning-based (CNN in this case) methods is to split document pages into lines at first, then to classify each line and combine the classification results by different rules.

* **Using Nanonets: Approach 3**

Using the Nanonets API we can automatically extract all the necessary information from the Resumes required for job searching and matching. Just upload a Resume and get all the extracted fields returned in the format of resume choosing.

**CONCLUSION**

Pdfminer is used to extract all the required information from resumes and I have extracted all the required outcome into separate excel sheet for visualization. Since dataset is not huge I have come up with the solution of pdfminer. We can also use other technologies such as NER algorithm, Object Detection and OCR and Nanonets on huge datasets. If we follow these technologies it will give better performance as compared to pdfminer