**Problem:**

In this era the most valuable thing in this world is time, everyone wants to utilize their time more productively but what about big company’s HR they get millions of resumes/CV every year and they have to study all and then analyze them so that they can shortlist best one for their company.

So here is the solution for those HR’s, so that they can utilize their time more productively.

Here is a sample question -

You are given a dataset of 20 sample resumes with the personal details masked out. Your job is to come up with a hypothesis for a generalized technique to extract certain information from all of them. The information which is to be extracted is stated below. You can use an ensemble of machine learning algorithms and other techniques that you think can help in making the predictions better. Your focus should be to come up with an algorithm/technique which can generalize for most of the resumes. You can take the help of any AWS/GCP services in your approach.

What to extract:

1. Names of organizations they have worked in and the years they were working in them. Certain resumes might not have this information, so you can ignore them.

2. Names of institutions and colleges they have studied in and the associated years.

**Solution:**

For the given problem statement I have used Named Entity Recognition(NER) approach and trained my model against 200 resume which is in a pickle format.

Training a model -

I have used python’s spaCy module for training the NER model. I have used the Dropout technique to improve the learning results and it is set a dropout rate, a rate at which to randomly “drop” individual features and representations. This makes it harder for the model to memorize the training data. For example, a 0.25 dropout means that each feature or internal representation has a 1/4 likelihood of being dropped. I have trained my model for 10 epochs and keep the dropout rate as 0.2.

After training my model, I have loaded it to “nlp\_model” then I have applied my trained model to the resume which is at the 0 indexes of my training data.

After that using Fitz I have applied my trained model to the resumes in PDF format and extracted the required data.

**Application:**

We can deploy this model in big companies by making a DropBox and through DropBox, all the CVs will go to the model and the Information gets extracted and send to HR where it gets evaluated easily.

Although at this time the accuracy of this model is not so good, it can be improved by using larger training data and increasing the number of epochs.