The following project named – “Resume Classification using NLP and Machine Learning techniques ” is one of my first assignments that I had received while I was starting out to learn NLP and Machine Learning. (Disclaimer: I am still in learning mode). In this small project my main aim was to work with a JSON (JavaScript Object Notation) file format and convert it into a .csv file format for better interpretability. I have tried to access the JSON file from the path mentioned in the notebook which is a local path. This particular technique can be used on any file type whose path is known. Secondly I have tried to vectorize the .csv file by various NLP techniques and then have applied machine learning algorithms to find out which methodology works well and produces a better score and accuracy. The problem statement was to categorize the resumes into two groups’ viz. **‘Data Scientist’** & **‘Data Analyst’** based on their job titles and work experience.

## Notebook I –

## The Following notebook is aimed at loading and access the data directly from the files by mentioning the path names. The file format used while importing is of JSON. Loading the json files and retrieving the job titles and work experience.

## Removal of any redundancy for any non-alphanumeric values and storing the values into two separate dataframes. Finally converting the values stored in the dataframes into .csv file format as ‘job\_desc.csv’

* **Notebook II –**

1. This notebook shows the implementation of different machine learning packages and shows a way to vectorize and split the data present.

* Reading the given data from csv file
* Lemmatization and transformation of data
* Splitting and Vectorizing the data

1. Using LinearSVC a part of scikit-learn SVM package for classification
   * Building a LinearSVC model
   * Comparing the LinearSVC model with logistic regression o different metrics
   * Finding out the top positively and top negatively correlated features along with their importance in the model.
   * Lastly building a cloud word for visualization of data

**What is a cloud word?**



Many times you might have seen a cloud filled with lots of words in different sizes, which represent the frequency or the importance of each word. This is called [Tag Cloud](https://en.wikipedia.org/wiki/Tag_cloud) or WordCloud. In simple terms cloud word or word cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. They are widely used for analyzing data from social network websites.

For generating word cloud in Python, modules needed are – matplotlib, pandas and wordcloud. To install these packages, run the following commands:

* pip install matplotlib
* pip install pandas
* pip install wordcloud

The dataset used for generating word cloud is collected from UCI Machine Learning Repository.   
Dataset Link : <https://archive.ics.uci.edu/ml/machine-learning-databases/00380/>