**Resume Analyzer Model- Project Report**

* Project Objective: To build a model which would take input as resume and determines the specific role which the resume fits for.
* Scope of the project:
  + Reading the input file
  + Tokenization using nltk library
  + Pre-processing
  + Finding similarity score between resume and each of the job description.
* Tools/Technology used:
  + Natural language processing
  + Numpy & pandas
  + Scikit-learn library
* Approach:
* The resume will be matched against three job descriptions which are “Data Analyst”, “Data Scientist” and “Data Engineer”.
* Natural Language processing tools are used for further analysis.
* Following steps are done on this corpus:
  + Tokenization — convert sentences to words
  + Removing stop words — frequent words such as ”the”, ”is”, etc. that do not have specific meaning
  + Lemmatization — Another approach to remove inflection by determining the part of speech and utilizing detailed database of the language.
  + Bag of Words Modelling concept is implemented, which is a way of extracting features from text for using in modelling.
  + A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:
  + A vocabulary of known words.
  + A measure of the presence of known words (we are taking count of words here as I assumed repetition of words can help in find the similarity)
  + All the above steps are implemented on the uploaded resume file(pdf).
  + Once we get vectorized form of the resume, cosine similarity is found out between each of the job description and resume. The pair which gives highest similarity score can be selected as the best job role as per the attached resume.
  + Cosine similarity is a metric used to determine how similar two entities are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. As cosine similarity doesn’t get effected by the size of the vectors, I selected it rather than going with Euclidean Distance.
  + the cosine similarity of two documents will range from 0 to 1. Hence, higher the value we get for resume-job description pair, the more suitable the job description is for the resume.
* Future Scope of the project:
  + We can include more job description so that resume gets compared with various job descriptions and best JD can be selected.
  + Further pre processing can be done to remove unnecessary words from the corpus which can increase the accuracy of the model.
  + Option to read different types of input files and also images. We can use pytesseract library for accurate OCR.
  + Model deployment.