DBS Partners internship assignment

**Objective** = To build a model that classifies whether documents are employment or amendments agreement and extract some features from employment agreement.

**Approach** = import all text document file in python and we use **NLP** for classification of these text files, after importing we do the following steps:

**Preprocessing:** In this method, we cleaned the data by removing an extra character or often use a most frequent word like the, an etc. after removing these word we use regex to remove special characters or alpha numeric as this creates **noise** in data.

**Vectorization:** after cleaning our text file we use vectorization technique to make sentences from a text file and then these sentences to the word. We use lemitization technique to make a word in the same form and remove other form.

**Assign a label to document to test the document**: we assign a label to documents to classify the documents, in this case, I use the label according to **‘AMENDMENT TO EMPLOYMENT AGREEMENT’** if this string found in files then assign label 0 otherwise assign 1.

**Train-test-split function**: after assigning a label to documents we use train test split methods to split our data into three sets which are train for training the model, test for testing the model accuracy. Now we do all the above methods to fit the data into training data and testing data.

So that all data in the correct form and does not create noise.

**Classifier**: after splitting the data we need to build a classifier which predicts our accuracy so to do this we have various built-in classifier. We use different classifier and measure accuracy.

One major problem comes when we predict the model is to remove a most frequent word or do **Tf-idf** and we also use ngram to build a model because it defines a relation in word which is helpful to predict better.

After applying various model based on this and various ngrams and **Tf-idf** values, we conclude that **LogisticRegression(C=100)** is best out of all data. And have an accuracy of **.981**.

**Extract Features**: after applying a model in documents we classify our document now we build a model to extract features. We take an employment document, and we use a library called **spacy**

In which we put the document in the scapy model, and this gives a result as a tuple in which word and their part of speech happen, and we use this speech method to classify our employ or employee and other features.

**Incomplete Answer** : I do not implement on code about extracting features.