NLP Internship Task

Jishnu S Nair
jishnunair1398@gmail.com
+91 8347508479

Task 1:

Classification of the documents.

Progress: Complete

1) At first the documents in the datasets were manually separated into two different folders.
Datasets:

Amendment Files:here
Agreement Files: here

2) The documents then undergoes preprocessing which lowers and removes white spaces and special characters from the text file

3) The document also undergoes tokenization

4) The final datasets is modified into a dataframe which consists of the text data from the file along with corresponding labels

label of 1 -> Employment Amendment file
label of 0 -> Employment Agreement file

5) As a part of feature engineering, the tf idf,count_vector score were calculated.

6) And on the basis of this scores, a multinomial Naive Bayes classifier is used.

Code:here

Task 2:

Progress: Partial

Named Entity Extraction

1) The initial thought process is use to a spacy model but the output where limited to the features like organisation name,person name,base salary,date and it is quite difficult to extract other features.

Code:here

.

2) So a custom model needs to be prepared for this, for which two approaches can be used.

a) Training a custom Spacy model by giving training inputs.
Code: offical_website

b) Using Rasa NLU(which basically uses a SVM model) extract features from the text.
a sample of Rasa NLU from previous work: here

Note: THe above two approaches which requires a decent amount of datasets for proper extraction.

Github link:here