**Document Parser**

**Name: Shashank Saumya**

This document has the steps explained in a very concise form. For long and full explanation, check the other document named "Long Explanation”.

**Exploratory Data Analysis:**

* Heavily inspired by the Exploratory Data Analysis done by Luis Fernando Torres in his [Wine Quality: EDA, Prediction and Deploy](https://medium.com/latinxinai/how-i-deployed-a-machine-learning-model-for-the-first-time-b82b9ea831e0) project.
* Found that the data is heavily skewed.
* Need for UnderSampling.
* After excluding “OTHERS” the data is not so skewed anymore.

**Feature Engineering:**

* Replaced features: coordinate of top\_left and bottom\_right replaced by the coordinate of their centers.
* Added a feature named index\_len which contains the difference of start\_index and end\_index.
* Used RFECV to decide the features to be used for training.
* All of the analysis is done in “Data Analysis full.ipynb” for full data and “Data Analysis filtered.ipynb” for filtered data (data without “OTHER” class).

**Model Pipeline:**

* Decided to go with two Random Forest Classifiers combined.
* 1st classifies if the label is “OTHER” or not.
* If not, then 2nd classifies the actual labels from the main dataset.
* Tried ClusteringCentroid under sampling. Took too much time
* Instead used RandomUnderSampling due to time constraints.
* Used Stratified K Fold Cross Validation to check model score.
* Average CV score around 96.57% for model 1 and around 94.96% for model 2.
* n\_estimators was also evaluated by training model on a range of estimator values. The accuracy achieved was very high (94-96%) and the variance in accuracy was quite low.
* All of the analysis is done in “Model Analysis.ipynb”

**Final Model:**

* n\_estimators used for both models: 200
* The model was trained on the entire training data.
* All the documents from validation were predicted using the model and the predictions were appended back and saved in a separate folder named “results”.