Infrrd Assignment
Task: Extract the following entities from the dataset
Entities present in the Dataset:

1. employerName
2. employerAddressStreet_name
3. employerAddressCity
4. employerAddressState
5. employerAddressZip
6. einEmployerIdentificationNumber
7. employeeName
8. ssnOfEmployee
9. box1WagesTipsAndOtherCompensations
10. box2FederalIncomeTaxWithheld
11. box3SocialSecurityWages
12. box4SocialSecurityTaxWithheld
13. box16StateWagesTips
14. box17StateIncomeTax
15. taxYear

Dataset:
The dataset contains images and tsv files.
The tsv files are token level extractions of their respective images with the columns of the file indicating the below:
For Train:
start_index, end_index, x_top_left, y_top_left, x_bottom_right, y_bottom_right, transcript, field
For Test:
start_index, end_index, x_top_left, y_top_left, x_bottom_right, y_bottom_right, transcript

- The start and the end index denote what the token's position would be, had the contents of the documents been represented in a single line.
- The subsequent 4 columns indicate the top left and bottom right coordinates of the token
  - Origin (0,0) of the document is the top-left corner of the image, y-coordinate increases as we go down the document.
- Transcript is the OCR extraction of the image
- Field is the ground truth or the label for the transcript

Evaluation

Make sure to add your model predictions in the last column of the tsv file for the val set (found in the val directory) and run eval.py (against val_w_ann), this will generate the metrics.tsv file - this contains a field-wise list of precision, recall and f1 score. Attach the metrics.tsv file and any analysis you've done when making your submission.

You're free to make any assumptions about the dataset and the training process. Think of it as an open-research problem and do as you please with the info/data you have at hand.
You're allowed to use any code that you find online, but please cite your

source/references/repos. You’ll be evaluated on the following items (in no strict order):
1. EDA (Exploratory Data Analysis)
2. Error Analysis
(You can use any State Of the Art models for this purpose)

Also delineate how your EDA affected any pre-processing steps you chose to take and how it influenced your choice of model/algo.
From your error analysis, if you’re able to reason out what’s boosting or hampering your model performance, please make sure to document that as well.