Entity Extraction from Financial Documents: EDA, Model Analysis, and Error Insights

Train

```
python train.py
```

This will create a model that fitted with train dataset (dataset/train/boxes_transcripts_labels) and it creates LabelEncoder for x and y in models folder.

Output

```
Score: 0.934306876525545
```

Predict

```
python predict.py
```

Output

```
100%|███████████████████████████| 207/207 [00:01<00:00, 117.17it/s]
```

This takes all the tsv files from `dataset/val/boxes_transcripts` and predicts the y value and combines that y to x in `field` column and save it as the same file name in the dir `dataset/predictions`

Eval

```
python eval2.py
```

Output

```
 100%|█████████████████████████████████████████████████████████████████████████████████████████| 207/207 [00:01<00:00, 148.38it/s]

Accuracy with `OTHER`: 1.0

100%|█████████████████████████████████████████████████████████████████████████████████████████| 207/207 [00:01<00:00, 137.62it/s]

Accuracy with `OTHER`: 1.0

Press enter to see sample data...
```

This will printout that accuracy in both with 'OTHER' and without 'OTHER'

And it shows `Press enter to see sample data...`

Output

```
 box17StateIncomeTax - box17StateIncomeTax
box17StateIncomeTax - box17StateIncomeTax
box17StateIncomeTax - box17StateIncomeTax
OTHER - OTHER
...
...
OTHER - OTHER
OTHER - OTHER
```