hypothizer

Assignment for Machine Learning Internship

Hypothizer Technologies Pvt. Ltd.

Address: L-Incubator, IIM Lucknow Noida Campus, B1, Institutional Area, Sector 62, Noida – 201307

Email ID: info@hypothizer.com

Problem Statement: Data Capture From Invoices

Background

Hypothizer's flagship product is a tool to capture intelligence-ready data from documents. It eliminates the need for manual data entry, as it automatically captures relevant data from scanned paper documents and populates the database accordingly.

Data capture from unstructured documents is an interesting problem being researched from decades. The problem statement is to capture the fields data from unstructured invoice documents.

Problem Description

The task is to develop a Machine Learning model to capture the data of fields, Invoice Date, Invoice Number, Buyer GST Number, Seller GST Number, and Total Amount.

Dataset is provided in a pandas dataframe saved in a pickle file. It has columns; groups, type, data, label, coords.

Description of all the columns:

1. Groups: It contains the group of words which can also be called as tokens. These are the lines which needs to be classified.
2. Type:
    a. It contains a tuple with first element being the processed version of words contained in token present in groups column.
    b. Second to before last two elements are the type of the words present in the token. It is found in this range [1:-2]
    c. Last two elements are the page height & width.
    d. Ex: (ncube solut, text, text, 3507, 2481) here ncube solut is the processed word, text, text are the types of the processed word, 3507 & 2481 are height & width.
3. Data: It contains image name from which the data has been extracted.
4. Label: It contains labels for different classes.
    a. 0: It stands for other class
    b. 1: It stands for invoice date
    c. 2: It stands for invoice number
    d. 8: It stands for Buyer GST Number
    e. 14: It stands for Seller GST Number
    f. 18: It stands for Total Amount in the invoice
5. Coords: It contains the placement of the tokens in the image. It would be in a list form & Order being xmin, ymin, xmax, ymax, midpointX, midpointY.

Example of pandas dataframe row:

	groups	type	data	label	coords
0	NCUBE Solutions	(ncube solut, text,text, 3507, 2481)	14805	0	[195.0, 246.0, 530.0, 278.0, 362.5, 262.0]

Resources

We don't have any restrictions on libraries you may want to use though we recommend following libraries:

1. OpenCV
2. Scikit-learn
3. Skimage
4. Scipy
5. Pytorch
6. Tensorflow

Dataset

Input dataset can be downloaded from here:
https://drive.google.com/file/d/1cXIXLv3etD594T8ryYEsHrV-u5FLvipt/view?usp=sharing

Dataset images are available at this link:
https://drive.google.com/open?id=11Wuo89Ai260LA42t1s_3cZK0y9qxzpFh

Dataset is a pandas dataframe pickle file, which can be easily loaded using the below command:
import pickle
import pandas
dbfile = open(path of the pickle file, 'rb')
df=pickle.load(dbfile)

Deliverables

1. Source code in IPython Notebook with proper comments.
2. Model file in pickle format.
3. Any other dependency to generate result on test data.

Term

Please confirm your availability with exact dates.

Stipend

Please mention your expected stipend.

For further information, clarification and submission of assignment, please email at raghu@hypothizer.com.