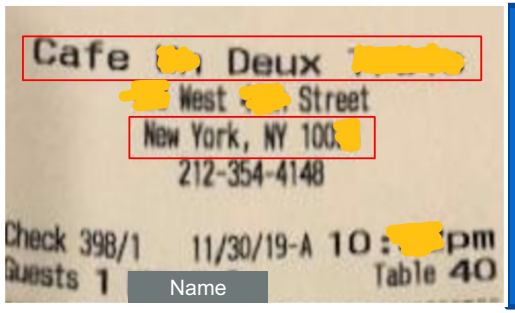
Amex Campus Challenge 2021
Data Science

Problem Statement

Given all the text that might appear in an invoice image, your goal is to classify all the text into pre-defined categories such as total amount/tax/tip, etc.

1 Sample Invoice



Invoice images will NOT be shared with participants

2 Bounding box & text info

```
[{"bbox": [...],
"description": “Café %$ Deux *@#^”
...},
{"bbox": [...],
"description": “New York, NY 100^&”
...}]
```

To be shared with participants /input for your model

3 Model

Model

Participants are expected to build a framework/ model

4 Label predictions

```
[{
“description” : “Café %$ Deux *@#^”
“label”: “w merchant name”
...},
{
“description” : “New York, NY 100^&”
“label”: “y merchant address inside”
...
}]
```

Expected results from participants

Business use case: Digitizing invoice, automating expenses, and for ensuring compliance and regulation.

Stages of Competition

Stage 1 Leaderboard Scoring

Data to be shared with participants for this round:

1. 900 training set
2. 100 validation set

Data expected from participants:

Upload predictions for the validation set in the output json format. Sample shared in next slide.

Results at the end of round:

Leaderboard Round 1 rankings and scores

Stage 2- Out of Time Scoring

Data to be shared with participants for this round:

100 test set (to be shared post conclusion of Round 1)

Data expected from participants:

Upload predictions for the test set in the output json format. Sample shared in next slide.

Results at the end of round:

Leaderboard Round 2 rankings and scores. List of teams selected for Stage 3.

Stage 3 - Virtual Interaction

Data expected from participants:

Share a deck or documentation which you'd be asked to present in front of a panel. It should broadly cover your understanding of the problem, insights and approach. Please find more details in following slides.

Results at the end of round:

List of participants qualifying for pre-placement interviews for final placement

Dataset Overview

Text extracted from receipts from Food and Beverage category will be shared with participants

- Dataset with 1,100 receipts is split as:
    - 900 receipts for training
    - 100 receipts for validation (Stage 1 Leaderboard)
    - 100 receipts for test (Stage 2 Submission)

- Each receipt will have two files associated with it (more details in the following slide)
    1. Input JSON containing meta data which includes
        a. Text Information
        b. All 4 coordinates of the bounding box for every text segment
            a. Clockwise - Top Left, Top Right, Bottom Right and Bottom Left.
            b. Top Left of image has coordinates (0,0)
    2. Output JSON containing the Label/Class information for every text segment that appears in input file

- There are 10 possible Labels
    - Please use the exact label mentioned in the list – eg. “total amount”
    - Comprehensive list of labels and their description are in the following slides

Dataset details

Each receipt will have two files associated with it. Samples below

1. Input JSON containing meta data which includes text info + text position
2. Output JSON containing the Label/Class information for every text segment that appears in input file

1. Extracted text
2. Text position
3. Text Label

Input Json


```
{"textAnnotations": [

{"boundingPoly": {"vertices": [{"x": 218.0, "y": 98.0}, {"x": 327.0, "y": 98.0}, {"x": 327.0, "y": 113.0}, {"x": 218.0, "y": 113.0}]}, "description": "ICHIKAWA"},

{"boundingPoly": {"vertices": [{"x": 220.0, "y": 112.0}, {"x": 316.0, "y": 112.0}, {"x": 316.0, "y": 121.0}, {"x": 220.0, "y": 121.0}]}, "description": "Japanese Restaurant"},
....

{"boundingPoly": {"vertices": [{"x": 63.0, "y": 289.0}, {"x": 299.0, "y": 285.0}, {"x": 299.0, "y": 310.0}, {"x": 63.0, "y": 314.0}]}, "description": "Refreshing Tuna Poke Rice Bowl"}]}
```

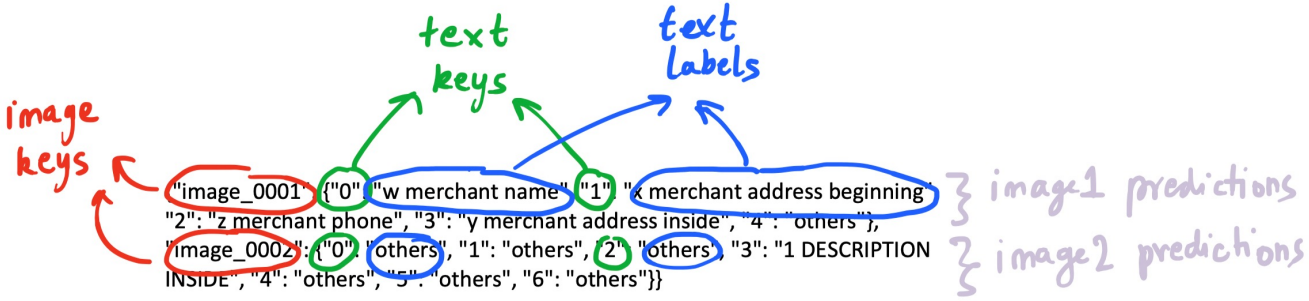

Output Json


```
[{"id": 1, "text": "ICHIKAWA", "label": "w merchant name", "box": [218, 113, 327, 98], "linking": []},

{"id": 2, "text": "Japanese Restaurant", "label": "w merchant name", "box": [220, 121, 316, 112], "linking": []},
...

{"id": 12, "text": "Refreshing Tuna Poke Rice Bowl", "label": "0 description beginning", "box": [63, 314, 299, 285], "linking":[]}]
```

Submission format

Please follow the following instructions for creating your submission file for leaderboard

1) File should be in json format
2) Name your file as your_team_name.json
3) The format of the file should be a nested dictionary
    1) The high level dictionary key should be the image names (for eg. for image 0001 the string should be 'image_0001')
    2) The nested dictionary key should be the index of the text you're predicting for (eg. "0" for the first text appearing in your input file for that image)
    3) The value of the nested dictionary should be the label (eg. "total amount" etc.)

Sample text file which has labels for 2 images.	"image_0001": {"0": "w merchant name", "1": "x merchant address beginning", "2": "z merchant phone", "3": "y merchant address inside", "4": "others"}, "image_0002": {"0": "others", "1": "others", "2": "others", "3": "1 DESCRIPTION INSIDE", "4": "others", "5": "others", "6": "others"}}
Details on how the above text file is structured.	

List of Labels

SNO	*Type*	*Label*	*Description of Label*
1	Item	0 description beginning	Description of Item Purchased
2	Item	1 description inside	Description of Item Purchased
3	Item	e quantity	Quantity of Item Purchased
4	Item	unit price	Unit Price of Item
5	Item	5 total price per item	Final Price of Item
6	Misc	sub total	Sub total in a receipt
7	Misc	2 tax total	Amount of tax paid
8	Misc	3 tip	Tips/Gratitude offered
9	Misc	total amount	Final amount paid
10	Misc	others	Any other text

Evaluation Criteria- Quantitative

We would consider label fields for every text segment appearing in input JSON & evaluate the F1 Score for each image. Final evaluation metric is the average F1 score for all invoices.

Final score is the average F1 scores across N test invoices as

$$\text{Evaluation metric} = \frac{\sum_i^N F1score(y_{true_i}, y_{pred_i})}{N}$$

where y_{true} is the array of true labels for i^{th}invoice and y_{pred} is your array of predicted label for i^{th} invoice

For more information about F1 score used here please refer to the following link:

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

In parameters average = 'macro' will be used

Evaluation Criteria- Qualitative

Once selected post Leaderboard round 2, you are expected to share a deck or documentation which you'd be asked to present in front of a panel.

It should broadly cover the following ideas :

1. Understanding of Problem:
    I. Understanding of the overall objective
    II. Data Exploration : Understanding trends, patterns & deriving insights
2. Methodology & Thought process: Ideas & Approaches Tried
    I. Data / feature engineering
    II. Choice of algorithms and models
    III. Validation
3. Assumptions

Guidelines – suggested packages and tools

- Any open-source packages
  - Machine/Deep Learning packages
    - TensorFlow
    - Torch
    - NLTK
    - Spacy
    - Scikit-Learn
  - Word Embeddings
    - Flair
    - Bert
    - Word2Vec
    - Glove
  - Graph Neural Network
    - Graph Convolutional Network
    - Graph Attention Network

- Recommended Paper
  - https://arxiv.org/pdf/1903.11279.pdf

Rules

1. A team can have 1 to 2 members
2. An individual employee cannot be a member of more than one team
3. In case of a tie between two teams, the lead would go to the team which submitted the results first
4. In case of a discrepancy, the decision by the organizing team would be final