Bill Challenge

Description: BILL subscribers are able to make purchases with our corporate card solution. When a purchase is made, users can fill in information about the purchase such as the date of the purchase, the amount, and vendor name. To validate the purchase, a receipt must be attached. Therefore reconciling all transactions is a time consuming process. To alleviate this time consuming task of matching uploaded receipts with past transactions we challenge you to design a Receipt Matching algorithm to save customers' time.

Data set: The dataset consists of two main sources. The first, is a data frame (**Users.csv)** of transactions entered by the users. The second source is the receipt data accompanying each transaction. The receipt data is composed of:

<u>Receipt Information</u>

- Receipt images (jpg) stored in **img** folder
- Optical Character Recognition (OCR) files (csv) of the receipts stored in **ocr** folder
    - Each row of a single file defines bounding box coordinates of a text string and its contents: x1,y1,x2,y2,x3,y3,x4,y4,TEXT
    - Coordinates start at top left (x1,y1) and go clockwise in rectangular fashion ending at (x4,y4)
    - Commas can exist in TEXT

Note, the majority of the receipts will have OCR data but a small fraction will **only** have the image data. The user information data (**Users.csv)** will have rows representing a single transaction and will have the following columns:

<u>User Entry Information</u>

- Id: unique row identifier
- Vendor Name: The vendor name entered by the user
- Amount: The amount entered by the user
- Date: The date of transaction entered by the user
- Receipt: The name of the receipt image associated with the transaction

Goals: Your goal is to create a model that assigns each receipt to its corresponding user entry in **Users.csv** by only using the information from the image and text and coordinates generated by OCR. The test dataset will be withheld until approximately 2pm on Saturday.

Please report your test accuracy and any other relevant metrics/visualizations for **test_transactions.csv**. When judging, we are interested not just in the raw performance of your algorithm in terms of test accuracy, but are also looking for interesting/novel problem-solving approaches, insightful exploratory data analysis and visualizations, possibilities for future work, and the potential challenges your model may face if deployed in a real-world setting.

Prizes:

- 1st place: ipad(s)
- 2nd place: ipod(s)
- 3rd place: Bill swag(s)