Graph2Table Challenge

Description

A company has a lot of client data in word, pdf, images etc. containing different types of graphs. They want to have this data stored in their database.

They approached you with a dataset of extracted plots of different types, and want you to:

1. Classify images of plots into different classes. (2pts)
2. Somehow able to read data in these images and convert it to row and column format, so that the data can be inserted into database. (1 pt)

1. **Classify images of plots into different classes:**
   You are provided with a dataset.zip, containing images of different types of plots.

   Tasks:

   a) Clean the dataset (remove any file which is corrupted, make sure images are in same format (.jpg or .png))

   b) Build a classifier training code (pytorch or keras or classical ML model anything is fine)

   c) Do the accuracy metrics analysis

   d) Save model

   e) Inference code testing your model on any image from the dataset

   Deliverables:

   1. Complete code or notebook
   2. Saved model file
   3. Documentation containing instructions to run your code and a requirements.txt (You can also make use of Markdown cells if you are using a notebook)

2. **Somehow able to read data in these images and convert it to row and column format, so that the data can be inserted into database:**
   To have a full-fledged solution for this part is really difficult task. For that reason, you are expected here to make a working solution for only the images mentioned in dataset_part2 (contains 2 bar graph plots). You can have different code that works for different images.

   Make use of existing OCR engines to extract the information in the given images.

   Tasks:

   a) Extract and print 'title of graph' and make bounding boxes around the texts detected in the images and save the resultant images with Bounding boxes.Save images as '{image_name}_bboxes.png'
   b) Make use of computer vision skills, get individual bars and their values (Hint : contour detection, edge detection etc). Write the individual results into '{image_name}.csv'

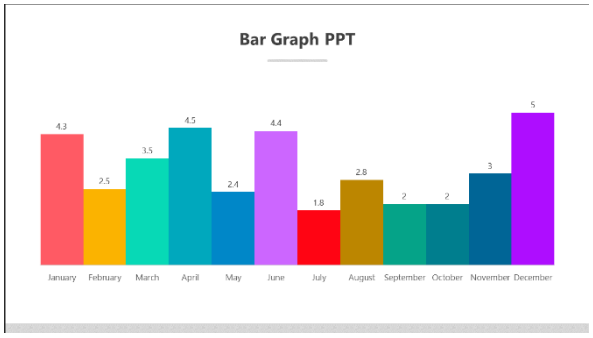
   Deliverables:

   4. Complete code or notebook
   5. Documentation containing instructions to run your code and a requirements.txt file (You can also make use of Markdown cells if you are using a notebook)

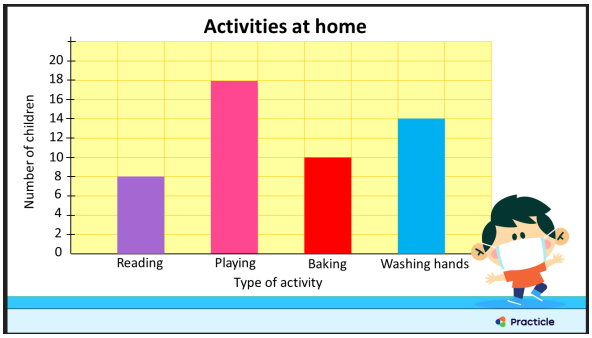
   Resources for part 2:

   a) The easiest way to get started is using EasyOCR (https://pypi.org/project/easyocr/) .
   b) Look at cv2 contour detection, edge detection, corner detection. (Feel free to ask and discuss approaches on the graph2table teams channel)

Expected results for part 2:



	A	B
1	January	4.3
2	February	2.5
3	March	3.5
4	April	4.5
5	May	2.4
6	June	4.4
7	July	1.8
8	August	2.8
9	September	2
10	October	2
11	November	3
12	December	5
13		



Type of activity	Number of Children
Reading	8
Playing	18
Baking	10
Washing hands	14