CSE 576: Phase 2: Automated Data Creation

Siddharth Nilesh Joshi

ASU Id: 1217923356

We have exploited the FigureQA dataset and DVQA Dataset. I have worked with **the FigureQA Dataset**:

figureqa-train1-v1.tar.gz(2.12 GB) . The Python script to generate the dataset is named **SiddharthDataset_script.py** and the dataset dumped into a JSON file by name **Siddharth_Dataset.json** .

Column details

- **Entry**: It is a reference to maintain a log of generated dataset.
- **dataset_id**: It refer to the dataset considered as the base. Here, it was FigureQA.
- **Image_index:** The image index if referred to under Image_index attribute.
- **Graph_type:** Elicits the type of figure (i.e. line, dot_line, pie, hbar_categorical and vbar_categorical)
- **Passage:** It contains the text passage of the particular dataset to suffice the multi-modality of the dataset.
- **Question:** It is the text question to be asked; created from the provided annotations.
- **Answer:** The answer based on the dual modality of image and text-passage is provided in this column.

Contribution

- I have worked on the '**Change**' type of manipulation of the FigureQA dataset. Drawing references from the FigureQA dataset, individual datasets were synthetically created. Based on the type of graph, passage, question and answer were created. For each of these datasets, fields namely 'Entry', 'dataset_id', 'Image_index' and 'Graph_type' were majorly common; however, the passage, questions and answers were significantly distinct. The templates were built to generate the datasets as per graph types.
- To 'change' a particular entity in the graph, a list of available color choices was created with every iteration, and choosing randomly from the same, that particular entity was then altered and further used to generate the answer.
- Finally, the synthetic dataset is converted to a list, and dumped into an outfile named by **Siddharth_Dataset.json .**

Instruction to run Code

- The code uses basic python libraries such as ***json*** and ***random.*** The interpreter used was Python 3.7. The *annotations.json* file (obtained from the FigureQA dataset) should be in the same directory as the dataset generation script.
- Download and save the *annotations.json (*aforementioned link*) and SiddharthDataset_script.py* files in the same directory.
- Open the terminal and access the directory in which the above files are stored; and execute the command *python SiddharthDataset_script.py*
- This should create file *Siddharth_Dataset.json* .