**Scraping the Play Store**

1. Install the required libraries from the *requirements.txt*. Libraries include
   1. [Play-scraper](https://pypi.org/project/play-scraper/) - To scrape data from play store
   2. [Jupyter](https://jupyter.org/install) - Python Interpreter
   3. [Pandas](https://pandas.pydata.org) - Data manipulation
   4. [Tqdm](https://github.com/tqdm/tqdm) - Progress Bar library
2. Run the *play\_store\_scraper.py*.
   1. To change the countries, add the country codes of the countries of your choice to the *countries* list (Country codes can be found [here](https://github.com/danieliu/play-scraper/blob/master/play_scraper/constants.py#L87)). Presently set to **[‘in’, ‘us’]** for India and US
   2. To change the number of results you’d like to save, change the value in *number\_of\_results* variable. Presently set to **100**
   3. To change the categories you’d like to fetch, change the *categories* variable. Presently set to all categories from [here](https://github.com/danieliu/play-scraper/blob/master/play_scraper/lists.py#L12) ( Note that ‘ANDROID\_WEAR category is a list of all the apps that have an ANDROID\_WEAR version and hence we can exclude it from our consideration)
   4. This python file fetches the information of all the ‘TOP FREE’ apps from the play store and creates a JSON file in the *output/app\_info* folder with the file name *{country}\_{category}\_{number\_of\_results}.json*
   5. It also sums the number of downloads for each market (India and US) (in 1000s) and stores it in *output/installs\_per\_category\_for\_{number\_of\_results}.json*

**Selecting the Apps**

* To select the apps based on the distribution of downloads(installs), a double bar graph is plotted with Downloads v/s Categories (for IN and US)
* It’s noticed that there’s a distribution gap between Communication, Productivity, Social, Tools, Travel and Local, Video Players and other categories (closest is Photography and Tools with a gap of 1,475,500,000 installs (from output/installs\_per\_category\_for\_100.json)
* The top 10 categories (based on the number of installs) are similar for both the countries except for the order
* Hence the following categories are chosen
  + Communication
  + Productivity
  + Social
  + Tools
  + Travel and Local
  + Video Players

**Running the Code**

The steps to run the code are

1. cd into the /code/ folder
2. To install the required library and activate the virtual environment
   1. *pipenv install -r requirements.txt*
   2. *pipenv shell*
3. Run the play store scraper
   1. *python play\_store\_scraper.py*
4. Once the script finishes running, open the notebook
   1. *jupyter notebook 'Plotting Data.ipynb'*
5. Make sure you run all the cells in the notebook - [how to run all cells](https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/execute.html#executing-a-notebook)

PS: Sometimes, you may get an exception if there are many requests in a day (I’m not sure of the exact number, but it happened to me a couple of times. I waited for a couple of hours and tried again. I didn’t get any errors after the wait.)