Scraping the Play Store

1. Install the required libraries from the *requirements.txt*. Libraries include
    a. Play-scraper - To scrape data from play store
    b. Jupyter - Python Interpreter
    c. Pandas - Data manipulation
    d. Tqdm - Progress Bar library

2. Run the *play_store_scraper.py*.
    a. To change the countries, add the country codes of the countries of your choice to the *countries* list (Country codes can be found here). Presently set to **['in', 'us']** for India and US
    b. To change the number of results you'd like to save, change the value in *number_of_results* variable. Presently set to **100**
    c. To change the categories you'd like to fetch, change the *categories* variable. Presently set to all categories from here (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)
    d. 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*
    e. 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
    a. *pipenv install -r requirements.txt*
    b. *pipenv shell*

3. Run the play store scraper
    a. *python play_store_scraper.py*

4. Once the script finishes running, open the notebook
    a. *jupyter notebook 'Plotting Data.ipynb'*

5. Make sure you run all the cells in the notebook - how to run all cells

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.)