Project: Google Play Store

Google Play Store is a rich source of information about the various aspects of mobile apps. This includes information about customers, businesses, and technical information in the form of permission details, descriptions, version information and reviews. On July 24th, 2013, Google announced that they had crossed one million applications, and estimates now state about one and a half million apps and billions of downloads. This rich source of information has been largely untapped due to there being no public access to this data. This data is a potential source of information about customer sentiment and perception of apps. App developers can use such information to build better apps and respond to customer complaints faster. A few companies like '*AppAnnie*' and '*Distimo*' provide limited access to such data, and a few researchers have written APIs to scrape some selected amount of data to suit their needs. However, such a selective sampling of data using the custom written APIs can lead to an analytical bias.
The operation of every API is based on the "*docid*" parameter.

The "***docid*"** is the unique identifier for each application. This docid is available in the URL of every application page when you visit Google Play Store. For example, for the Facebook application the URL is "https://play.google.com/store/apps/details?id=com.facebook.katana" and the docid is "com.facebook.katana".
****

Scope of the Project

The project has two major folds:
Task-1: Extracting the Name, Docid and the app URL/Link from the Google Play Store.
Task-2: Extracting the Name, Docid and rank of the application for any of the website of your choice that rank the Google Play Store apps on either weekly basis or monthly basis. I need to approve the site of your choice before you start with the this task

****

Environment

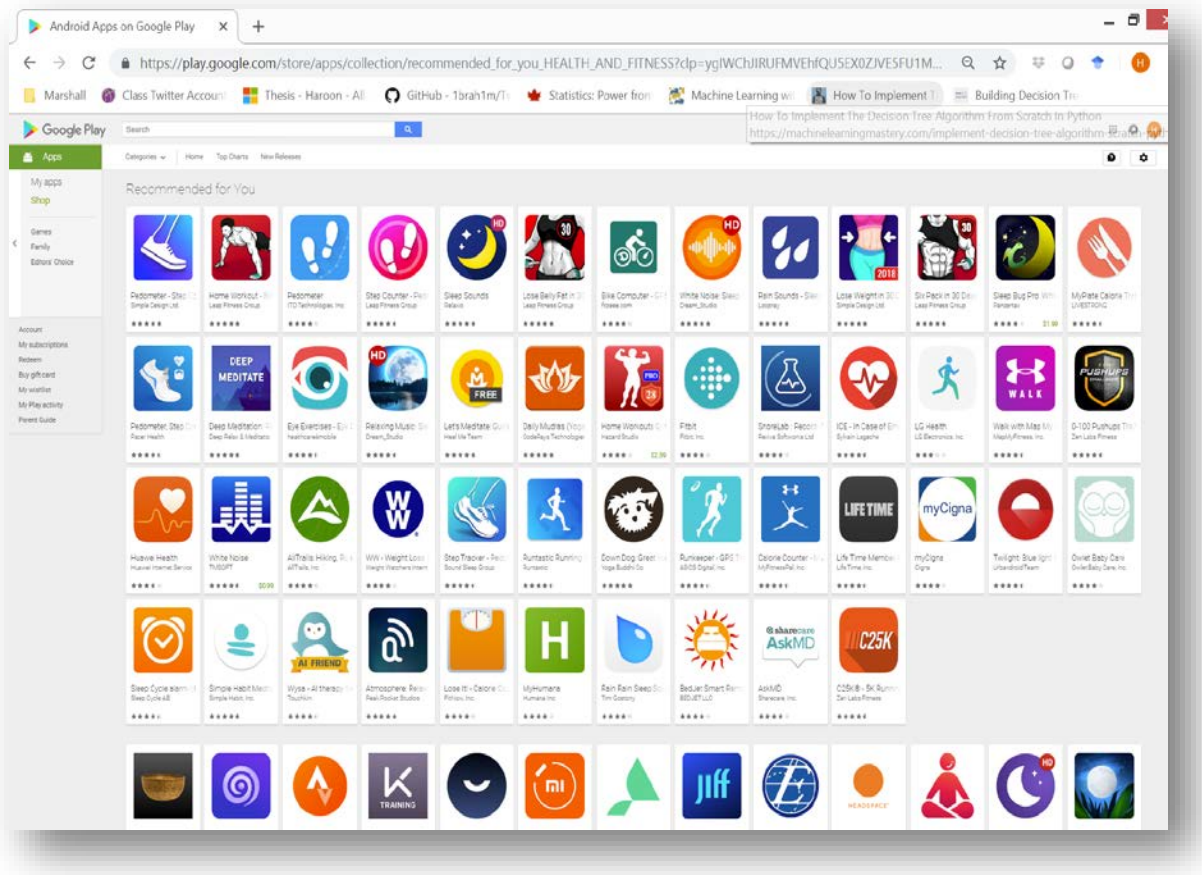
Language: Python
API: Any, provided it is free and had a decent limit on download quota. Before using an API, please consult with me.
Database: MySQL Version (8.0.12), community edition.

This is not the final draft of the project.

Task-1

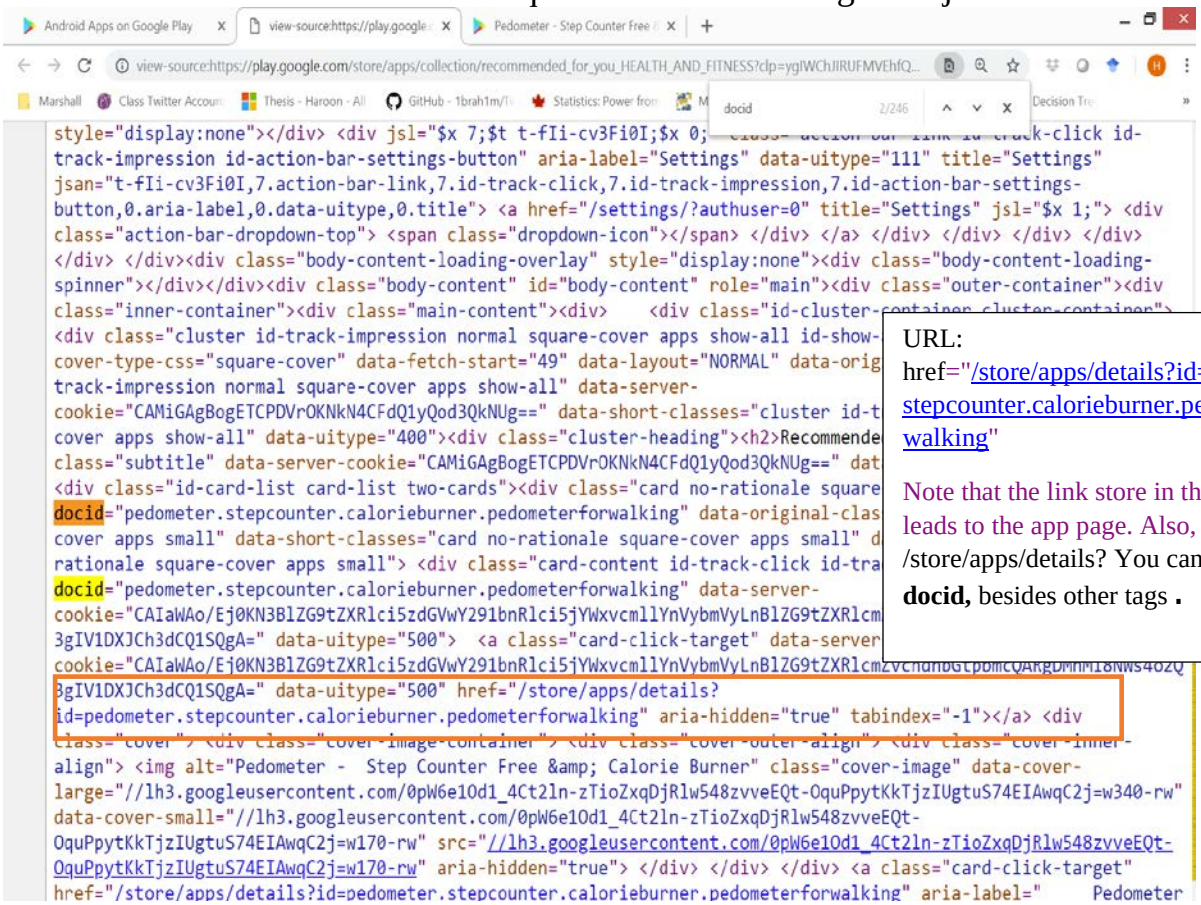
You need to produce a command line crawler, that will take at the minimum, a URL as an argument and fetch the App name, its corresponding Docid, its respective URL and save it into the DB.
For example, if the following URL is provide to the crawler:
https://play.google.com/store/apps/collection/recommended_for_you_HEALTH_AND_FITNESS?clp=ygIWChJIRUFMVEhfQU5EX0ZJVE5FU1MQAQ%3D%3D%3AS%3AANO1ljKVgyA&hl=en (Page shown below)

The crawler should retrieve all the Apps name, Doc ID and app URLs listed on the page (minimum of 30).



Hint: if you view the page source, i.e.,
view-source:https://play.google.com/store/apps/collection/recommended_for_you_HEALTH_AND_FITNESS?clp=ygIWChJIRUFMVEhfQU5EX0ZJVE5FU1MQAQ%3D%3D%3AS%3AANO1ljKVgyA&hl=en (Page shown below)

You should have access to all the required information to get the job done.



By simply knowing the appropriate tags, you can scrape the corresponding values for App name, URL and Docid.

****

Task-2

Pickup any website of your choice that maintains the Top-Chart(s) of google Play Store (Andriod) apps and scrape the App Name, its Docid, its rank and store them into the database.