**SI507 Project Document**

Tianhao Liang

**Project code**

Link to github repo: <https://github.com/liangth35/SI507project>

**README:**

**To run the code:**

If you want to rerun the whole crawling process: delete gamelist.csv and gameData.json, then run all the cells in Project\_retrieve\_data.ipynb. It will take roughly an hour to crawl and scrap 6544 games from steam.

Then, run Project.py in command line, a link will be generated (http://127.0.0.1:5001), the flask app will be available in that address. On the homepage, you can type in the range of price, release date, rating, and number of reviews of games you would like to search for, you can also choose to search for discounted games only. After the ranges have been set, click on "Search" button and the results will show up. On the result page, you can choose to sort the games by price, release date, rating, and number of reviews in both descending and ascending order. You can also select an attribute and plot its distribution by clicking on "Plot" button. When in the plot page, clicking "Return" button will take you back to results page.

**Required packages:**

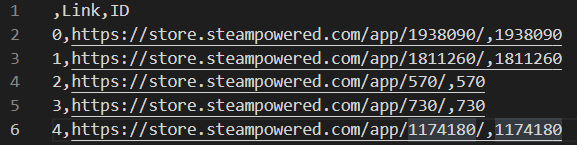
bs4, flask, requests, pandas, plotly

**Data sources**

First, I crawled the searching page on steam to obtain a list of games and the urls of their store pages. The url I used for crawling was [https://store.steampowered.com/search/?ignore\_preferences=1&category1=998&os=win&filter=globaltopsellers&page=%d](https://store.steampowered.com/search/?ignore_preferences=1&category1=998&os=win&filter=globaltopsellers&page=%25d), in which %d was replaced by the page numbers. There are about 25 games on each searching page. I scraped 404 pages in total and got 6654 valid game id–url pairs.

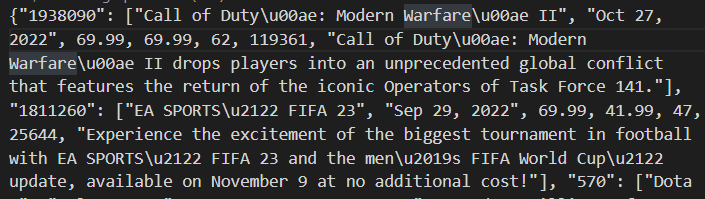
I obtained the html files of these pages using request.get, and I used beautifulsoup to extract the game ids and urls form the html files. I put the game ids and urls into a pandas dataframe. (The urls were used for crawling later and game ids were used as dictionary key in the next part.)

I implemented caching in this part by saving the dataframe as a csv file. When executed, if the program finds the csv file, it directly reads from the file instead of scraping the pages again. This is part of the cached csv file:



For the next step, I scraped the urls I obtained in the previous part, there were 6654 urls and I applied request.get to all of them and extracted the **name, release date, original price, discounted price, rating, review number, and description** of each game using beautifulsoup.select. Among those fields, release date, discounted price, reviews, and rating would be used as the for dimensions of the kd-tree, so users can sort the games by one of these fields. I saved the game data into a dictionary, with the game id being the key. The resulting dictionary has 6654 items.

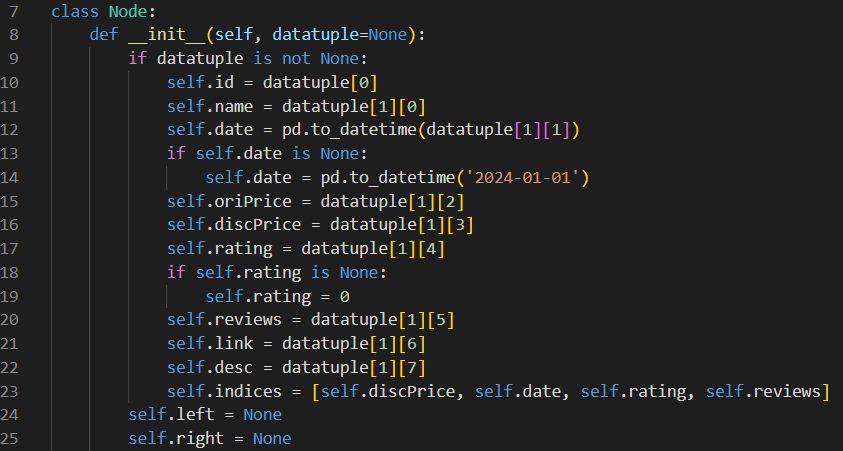
In this part, I also implemented caching. This time, I used a json file to cache the dictionary. This is part of my cached file:



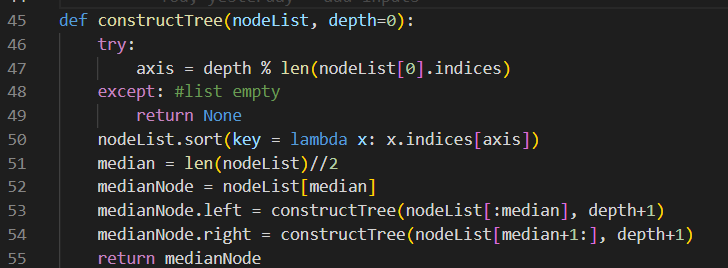
**Data Structure**

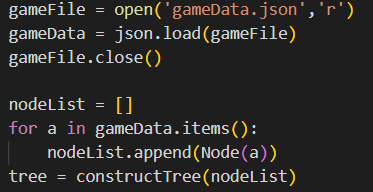
README:

The data structure I use in this project is a kd-tree. It is a variant of binary search tree, in which nodes are partitioned in 4 dimensions in turn as depth grows. I organize the game data into classes named Node. Other than the game data, each node has left and right child, which are set to None by default. I use 4 numeric attributes in game data: release date, original price, discounted price, and rating, as the 4 dimensions of this kd-tree.



When constructing the tree, the current dimension is calculated from the depth first. Then, the nodelist is sorted in that dimension and partitioned into two halves. These two halves of nodes are then fed to two recursive calls of the kd-Tree function and the returned results become the left and right child node. At last, the root node of the tree is returned.





Because I use classes to construct my tree, the resulting tree is not json serializable, and I cannot provide a json file containing my tree. Instead, I put both the tree-constructing code and the flask code in Project.py so that the tree can be directly passed to the flask app.

**Interaction and Presentation Options**