The Data Gathering Process

Timothy Schwieg

2018 年 5 月 27 日

Counter-Strike

- What do the items do?
- How are they obtained?
- Do people seriously pay money for this stuff?

Loot Boxes

- What are loot boxes?
- Is this the only way to obtain items?
- How does the steam community market function?

Steam Community market

- What data is given?
- Price and Quantity history over time
- Current buy and sell orders unfulfilled
- Not everything is sold on this market, but recently much more has been.
- Valve takes a percentage of sales, and is thoroughly incentivized to ensure this.

Volume of items

- There are ~ 11000 items sold on the market
- While many are just multiple variants of the same item (quality or slight color changes). This still is a lot of data to mine.
- To achieve this, need to use the steam market api.

Steam Community Market API

- First, we must get a list of all the available items on the market.
- For each item in the market, there is a price history as well as buy and sell orders available.
- This data is returned via json format, and must be converted to .csv.

How this is accomplished

- Using the python packages Beautiful Soup, requests and json.
- The pages that contain the items sold at market are returned in html, and must be parsed for the internal api name of each item.
- Once the internal name has been recovered, a request is sent for the market data for this item, it is returned in json format, and converted to a .csv table.

Building a file structure

- For Loot boxes, we are primarily interested in weapons which can be found in the boxes.
- The names of these weapons follow a particular pattern. Gun | Skin (Condition)
- This is the ideal place to use regular expressions to parse.
- `regX = re.compile( r'(?P<gun>[a-zA-Z0-9 \-]+)\|(?P<skin>[a-zA-Z0-9 \-\'\!` 🄵`]+)(?P<quality>\([a-zA-Z0-9 \-]+\)).csv', re.UNICODE)`

Exceptions

- However there are a few exceptions that I handled by hard coding them.
- First among them is a skin titled: M4A4 | 龍王 (Dragon King)
- Besides the obvious issue with Unicode, there is parenthesis in the title that are not the quality.
- The other problem is knives with no skin, which must be hard coded, but there are only four cases to handle.
- Once these are known, the file structure to be designed will be: gun / skin / condition

Merging Loot Boxes and their contents

- All the contents of the box are hand coded into files, and then their rarity is recorded into a script which finds available rarities for each item, and then gives each individual weapon its probability of being found.
- The last script takes the data on the price of each item in the market at the time the loot box was sold, and records it in a .csv file.

The End Result

- The end result is csv in the format where each row is a price and quantity sold, or a buy order placed. For each of these rows: the first three columns are the price, quantity, an indicator whether or not it is a buy order or a sold item, and the remaining columns contain price and probability of each item in the box.