**Documentation of Liquor**

**(Amazon Japan)**

**System Details and Database:**

|  |  |
| --- | --- |
| System IP | 172.27.132.55 |
| System Password | Act@101 |
| Database IP | 172.27.132.55 (localhost) |
| Database Name | amazon\_japan |
| Page Save Path | D:\PAGESAVE\AMAZON\_JP |

**VPN/Proxy Required:**

* Zyte(Crawlera) Proxy require and also require VPN as well if not getting response please change the proxy or VPN.

**Script/Code Details:**

|  |  |
| --- | --- |
| **Project Path** | D:\DATA\HINAL\LIQUOR\_SITES\GERMANY\amazon\_jp |
| **Project Back-up path** | D:\BACKUP\Liquor\_backup\amazon\_jp.zip |

### Points to be take care before running the script:-

### Change month date:

### Go to the db\_config.py file and change the month\_date variable value according to “DayMonthYear” format.

### Date will be fixed “08”

### For example: month=11, date: 08, year=2024 then month\_date = ‘08112024’

### Page Save file drive available in PC (for example: D) and check the file path as mentioned in db\_config.py file for page save & check EXPORT.py for excel output file.

### Check in spirit\_links\_japan table of amazon\_japan database where status column values are “Pending” and all type column values of all links are “old”.

### If status column values are not “Pending” then execute below query in SQL and update status

### “UPDATE spirit\_links set `status`=”Pending””

### 

### If type column values are not “old” then execute below query in SQL and update type status

### “UPDATE spirit\_links set `type`=”old””

### 

### Step for run the script for data extraction:-

### **Run the **link\_extractor.py** file** for extract links into the table as db\_links\_table name provide in db\_config.py file.

### You will also check during the count of all insert link using mention query as below.

### After running the ****link\_extractor.py****, you have require to run data\_extractor.py for extract data.

### The data\_extractor.py get links from spirit links table and request for getting the data if running from spider make sure to check start and end value at the end of spider. “execute('scrapy crawl data\_extractor -a start=1 -a end=10000000000000000'.split())”

### You have also run it’s in parts if require. For running in parts use run.bat file available in spider folder.

### In this project new link add and also check old link as well for data extraction.

### For checking how many new links added please refer SQL queries mention as below.

### While running the script please check pages are proper save or not

### During running the script check the status:-

### In SQL execute query:

### for check the count of new links added in spirit\_links table.

### “SELECT COUNT(\*) FROM ` spirit\_links` where type=”new””

### for checking the status of extract data is “Pending”, “Done” or “Not Found”.

### “SELECT "Total", COUNT(\*) FROM `spirit\_links` UNION

### SELECT "Done", COUNT(\*) FROM `spirit\_links` WHERE STATUS="Done" UNION

### SELECT "Pending", COUNT(\*) FROM `spirit\_links` WHERE STATUS="Pending" UNION

### SELECT "Not Found", COUNT(\*) FROM `spirit\_links` WHERE STATUS="Not Found"”

### How running further process:-

### Run the first file for link extraction and then after require to run data extraction file.

### Just run the script I have already set the file generation code while process is done and when spider close the file generation script run and excel file available on given path.

### Make sure for points for file generation

### I have try to convert the file into English language but don’t depend on this translation make sure and verify from your side. (It’s not proper working for large data)

### If any cases file not generate after completion the process, run EXPORT.py file manual and if facing any issue then contact to developer.

### You have require to change manual translate from native language to English. (Using Google Sheet)

**Store Information:**

* **Primary Developer:** Hinal Patel
* **Secondary Developer:** Siraj Dal

**Note:** Additional point for developers