ТЗ на онлайн парсинг

Нужно по командной строке вида

Run Script – с параллельным запросом всех требуемых к поиску позиций( мультипроцессорная обработка).

Открыть файл и считать part#

1. Выполнить поиск на сайте

<https://www.findchips.com/>

и вывести результат в файл

MAX232EESE+T

Результат в виде

" nomenclature\_name": [] – запрошенная позиция

{

"name": [], - наименования, которые выдает сайт-дистрибьютор по данному запросу

"vendor": [], - наименования производителя

"seller": [], - наименование поставщика

"MoQ": [], - minimum order quantity

"PQ": [], - package quantity

"prices": [], - цена с количеством через «;»

"in\_stock": [], - количество штук в наличии

"on\_order\_quantity" : [], - количество уже заказанного товара

"lead\_time\_days": [], - время доставки, указанное на сайте, в днях

"delivery\_time\_days": [], - время доставки в днях

"date\_code": [], - дата производства

"short\_description": [], - краткое описание товара

"seller\_url": [], - ссылка на поставщика }

{

Следующий ответ

}

Вида Sponsored – пропускам

Пример для

MAX232EESE+T

<https://www.findchips.com/search/MAX232EESE%2BT?currency=USD>

" nomenclature\_name": [MAX232EESE+T]

{

"name": [MAX232EESE+T],

"vendor": [Analog Devices Inc],

"seller": [Arrow Electronics Authorized Distributor],

"MoQ": [1],

"PQ": [1],

"prices": [

3000, 4.3220;

1000,4.3850;

500, 4.4490

],

"in\_stock": [4798],

"lead\_time\_days": [42],

"date\_code": [2207],

"short\_description": [Dual Transmitter/Receiver RS-232 16-Pin SOIC N T/R],

"seller\_url": [https://analytics.supplyframe.com/trackingservlet/track/?r=gPVv3S5fz9WS0muSSENj7oXgYdMxSPAs-tUU3scemG1BDHNvR4jqJsRad8PBMMhJFxbCdDulNc25MbvIEJru-y1jNPJJR\_GKKlNjHo3XPIHoawrEyQM2x3zixDYA8nCQB5Rw5ftjxEBxwP3O9JSvWte4h5UxaoSAeW6X6c9qZi\_T2J0phyt4bx1wumuPQLvbX4ivbtZbW2EcQN42ZMcuYoPOxzKN\_dPw88id3hSFlJMqED4N4AWxEHAOnIlu9Iy3cEry2sbJNT802wP47L-fKRk2-6dVBmBsky9mTFuflKNivkkiXzihwGxEcUWhsvY-IPpQcwXeO2hj5Y07NxK0IXXyff5rXuajWUDsr6Orc5UX0Ig7W50xGKLuSNLgENYTH4yR4TVylcGawqReh0u7PEjCSV\_abTtTtn\_Db7jbhmpBhrog-PA2GRE0KtlM2AODRqkRcUWcJ65dRkic5tIiTExuMQdwMKo3USg0Bj9GZA-utT9co1qQEoAJcAEZ5JTIg\_VEbQTDRU-9PxBgVxRpFkFsopXZgJfxVTNW-Z7CMlaLg3Vvh7I1Rvn5NPCHn3NnEXYQUl4ow0VazNvSBJ9P5egR75Z\_4aJMLaPfuxrGqp93vzMSz-hXh6N9bYZFWUo-otXhcmNZZoPDtXrLGVc1lwaQ0R267ObaEe70oYoCoIo]

},

{

Следующий ответ

}

1. Отдельный исполняющий модуль с поиском на сайте

<https://www.sekorm.com>

пример запроса

https://www.sekorm.com/Web/Search/channel?tab=4&searchWord=MAX232EESE%2Bt&ugc=k

без цен ответы пропускам

ответ

" nomenclature\_name": [MAX232EESE+T]

{

"name": [MAX232EIM/Tr],

"vendor": [Huaguan Semiconductor],

"seller": [Shiqiang Advanced (Shenzhen) Technology Co., Ltd.],

"MoQ": [2500],

"PQ": [],

"prices": [0, 1.6501],

"in\_stock": [],

"lead\_time\_days": [],

"date\_code": [],

"short\_description": [Transceiver; driver/receiver: 2/2; data rate: 120Kbps; 5V power su],

"seller\_url": [https://www.sekorm.com/product/871455.html]

},

{

Следующий ответ

}

https://www.sekorm.com/Web/Search/channel?tab=4&searchWord=MAX232EESE%2Bt&ugc=k

Или

<https://www.sekorm.com/Web/Search/channel?tab=4&searchWord=bav99&ugc=k>

" nomenclature\_name": [bav99]

{

"name": [bav99],

"vendor": [[Gaote](https://www.sekorm.com/brand/%E9%AB%98%E7%89%B9" \t "_blank)],

"seller": [Shiqiang Advanced (Shenzhen) Technology Co., Ltd.],

"MoQ": [3,000],

"PQ": [],

"prices": [0, **0.0608**],

"in\_stock": [],

"lead\_time\_days": [],

"date\_code": [],

"short\_description": [Transceiver; driver/receiver: 2/2; data rate: 120Kbps; 5V power su],

"seller\_url": [https://www.sekorm.com/product/871455.html]

},

{

"name": [bav99],

"vendor": [[Anbang](https://www.sekorm.com/brand/%E5%AE%89%E9%82%A6" \t "_blank)],

"seller": [Shiqiang Advanced (Shenzhen) Technology Co., Ltd.],

"MoQ": [3,000],

"PQ": [],

"prices": [0, **0.0816**],

"in\_stock": [50],

"lead\_time\_days": [3],

"date\_code": [],

"short\_description": [225mW Surface Mount Switching Diode-70V],

"seller\_url": [https://www.sekorm.com/product/871455.html]

}

Комментарии к коду

Детально:

1. Мне очень не понравилось, что преднамеренно используются функции задержки программы на конкретное время.

По всей видимости, это сделано для того, чтобы дождаться прогрузки страницы (иначе незачем). Однако, данный подход сильно иррациональный, ввиду того, что мы никогда точно не можем знать, через сколько времени откроется страница, а пальцем в небо тыкать, особенно когда важна скорость - плохо, а если учесть тот факт, что страница попросту может не успеть прогрузиться за это время, то мы, и вовсе, не получим никакой информации, а лишь потратим время на ожидание. Для решения данной проблемы используются специальные методы для ожидания того, пока сайт не откроется - ни больше, ни меньше (<https://stackoverflow.com/questions/26566799/wait-until-page-is-loaded-with-selenium-webdriver-for-python)>. В текущей реализации время ожидания слишком мало и сайт вообще не успевает открыться, чтобы собрать с него данные -> на выходе имеем пустой json.

2. Так как данный способ парсинга (через Selenium) достаточно медленный, но, по всей видимости, единственный, а нам важна скорость, последовательная обработка сильно затормозит процесс.

Вообще говоря, все, что касается web'а - все делается асинхронно или параллельно, так как в данных подходах не приходится ждать, пока сайт прогрузится или придет ответ. А учитывая задержки, описанные выше... В общем, стоит пересмотреть процесс и, думаю, сделать обработку мультипроцессорной (<https://pythonhint.com/post/5428055958566135/python-selenium-multiprocessing#:~:text=Python%20selenium%20multiprocessing%20refers%20to,multiple%20web%20pages%20in%20parallel)>

3. Работа плохо оформлена.

Прописаны библиотеки, которые не используются. Все сделано в одной функции, не разбитой на логические части. Много хардкода, локаторы захардкожены и не вынесены в отдельный файл (что, если они поменяются - потом ходи в код и ищи их...), куча try except, много повторений кода и тд. А также нет даже файла с зависимостями, которые используются, чтобы запустить код (так, к примеру, оказалось, что использовалась далеко не последняя версия Selenium, так как код просто не запускался на последней и пришлось искать причину), я уже молчу про docker и тп...

4. Для того, чтобы нормально интегрировать код, по итогу нужна функция, к которой можно обратиться и подтянуть все необходимые данные. Сейчас код работает так, что в цикле проходимся по позициям и записываем каждую из них в итоговый файл. А тут как раз таки тема в том, чтобы сначала все данные получить, а потом уже из них сформировать итоговый файл. А также выходные данные должны быть строго в формате:

{'nomenclature\_name': [], 'name': [], 'other\_names': [], 'seller': [], 'manufacturer': [],

'MoQ': [], 'PQ': [], 'prices': [], 'in\_stock': [], 'expected': [], 'on\_order\_quantity': [],

'lead\_time\_weeks': [], 'delivery\_time\_days': [], 'date\_code': [], 'short\_description': [],

'seller\_url': [], 'eccn': [], 'obsolete': [], 'offered\_quantity': []}

И если где-то есть пропуски из-за недостатка данных, то пусть они будут. Главное, - чтобы были поля и структура данных сохранялась.

В общем, человек сделал... хреново либо из-за неопытности, либо просто на пофиг. Хотя, я больше склоняюсь к первому варианту. Не знаю, оплачено ли уже, но, возможно, стоит найти кого-то другого... Потому что это нужно переписывать полностью...