社外秘

スクレイピングツールの仕様について

2023.08.24　作成

社外秘

仕様

処理内容

1. 検索キーワードからストアコードリスト取得（重複は削除）

   処理にはseleniumを使用。画面を直接操作しているので安定性が他より高め。

2. ストアコード一覧からそれぞれのストアで商品ページURLリストを取得（重複は削除）

   処理にはrequestsを使用。seleniumより処理は軽く高速だが、その反面アクセスが多くなりブロックを受けやすい。

   ※注文のマークがついているもののみ抽出。

3. 商品ページURL一覧からそれぞれの商品ページ記載の配送情報を取得する

   処理にはrequestsを使用。２.と同様にアクセスブロックされやすい。

※いずれの処理にも並走処理を実装しています。

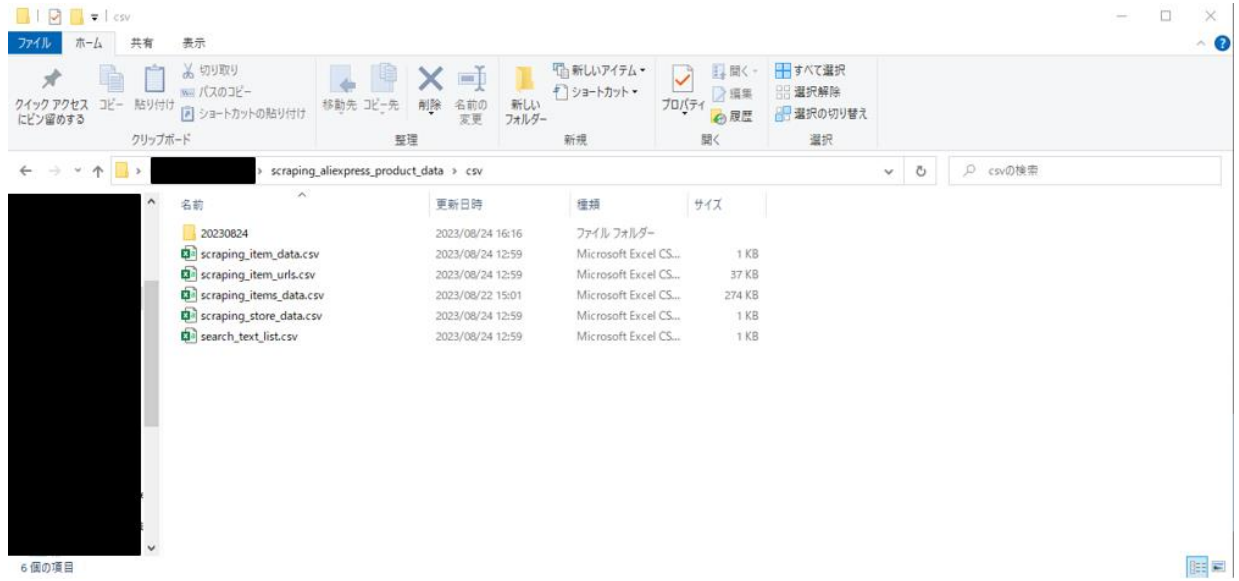
社外秘

入力ファイル／出力ファイル

インプット（入力）

1. 検索キーワードリスト（search_text_list.csv）
   もととなる検索キーワードを入力するCSVファイル。
2. ストアコードリスト（scraping_store_data.csv）
   商品ページURLを取得したいストアコードを入力するファイル。
3. 商品ページURLリスト（scraping_item_urls.csv）
   情報を取得したい商品ページURLを入力するファイル。

⇒デフォルトはcsvフォルダ内



アウトプット（出力）

1. ストアコードリスト（scraping_store_data.csv）
   キーワードそれぞれの検索結果で出てくるストアコード一覧ファイル。
2. 商品ページURLリスト（scraping_item_urls.csv）
   取得した商品ページURL一覧ファイル。
3. 商品ページ情報リスト（scraping_item_data.csv）
   取得した商品ページURL一覧ファイル。

⇒デフォルトはcsv¥yyyymmdd¥（当日日付）フォルダ内

※インプットとアウトプットでファイル名が共通になっているのでご注意ください

社外秘

設定ファイル

ファイルパスの変更

「CSV:」の箇所がCSVファイル名の設定箇所です。

・「path:」の下に書かれている「input:」 「output:」 を編集すれば、使用するフォルダを自由に変更できます。

・「input:」の下に書かれている「search:」「store:」「item_urls」を編集すれば参照するファイル名を変更できます。

・「output:」についても同様です。

並行処理数の変更

「TASK:」の箇所が並行処理の上限設定箇所です。

「max_workers:」に数字を入れることで並走数を制限できます。

画像では並走数は2に設定してあります。

※並走数は下げる場合にのみご変更ください（入力する数字は1～5以内）

CSVの設定

並走処理の設定

社外秘

アクセスブロック

認証画面

1. 商品ページにたどり着けない

   ページにアクセスしてもサイトにたどり着けず画像の画面が送られてくる。

ブロック発生時の挙動

1. Aliexpressからブロックを受けると表示される

   こうなると何回アクセスしてもこれが帰ってくるだけになる。

2. 例外処理は作成済み

   スライド操作をする処理は作成してあるので、

   この画面が表示された場合はseleniumで処理される。

3. ただしそれをやってもブロックは続く

   何回スライドしてもエラーで返ってくる場合がある。

   そうなるともう時間をあけて再度試すしかない。

社外秘

調査メモ

調査項目

1. AliexpressのAPIの利用

   販売者アカウントの作成が必要で、会社の証明書などできちんとした認証がある。

   ⇒アカウントを作成したとしても、APIで取得できるのは自分の販売サイト（ストア）内の情報だけ

2. Seleniumを利用しないで済ませられないか

   無理。ページのURLへアクセスした時点では情報がない。

   ページのレイアウトだけ最初に送られてきて後から中身の情報が入ってくる仕様になってる（JSで非同期的に呼んでる？）

3. IPブロック

   恐らくだが1つのIPアドレスから連続してアクセス（リクエスト）をするとそのIPアドレスへ情報を開示しないようにしてるよう。

4. 並走処理でブロックされやすくなってる

   並走処理だと並走させている分アクセスが集中してブロックされやすい。

   待機処理を加えても並走させているので大気のタイミングのずれより絶えずアクセスし続けているのと一緒になってる。

社外秘

対策方法

対策方法

1. 人間が操作させているように誤認させる
    1. アクセスする時間間隔を設定する

        処理速度は低下。安定性は向上。通常はこの手法を使うのが一般的。

2. 複数のコンピューターからアクセスされていると誤認させる
    1. プロキシサーバーを立てる

        このサーバーを経由すれば、アクセスを分けてくれるイメージ。

        無償のものもあるが安全性が保障されていないものが多い。