DBscrayper

このリポジトリには、メルカリのスクレイピングとデータ処理を行うための一連のツールが含まれています。

概要

このプロジェクトは、メルカリから商品情報を収集し、それらをベクトルデータに変換するためのツール群で構成されています。それぞれのツールはGradioを利用したWeb UIを提供しており、ブラウザ上で直感的に操作できます。

各スクリプトの機能と使い方

1. `app.py` - AI搭載メルカリスクレイパー

機能:
メルカリから指定した条件で商品情報をスクレイピングし、CSVファイルとして保存します。
AI（DeepSeek）を利用してウェブサイトの構造変化に自動対応する「AI修復機能」を搭載しています。

主な特徴:

- キーワード、カテゴリ、価格帯、販売状況、並び順による絞り込み
- 取得件数の指定
- 複数ジョブの予約実行（キューイングシステム）
- PlaywrightとBeautifulSoupを組み合わせた高速かつ安定したスクレイピング
- CSSセレクタが古くなった場合にAIが自動で新しいセレクタを推定
- 商品画像のダウンロード機能

使い方:

```
python app.py
```

実行後、表示されるURL（例: `http://127.0.0.1:7860`）にブラウザでアクセスします。必要な情報を入力し、「キューに追加 (予約)」ボタンを押すと、バックグラウンドで処理が開始されます。

2. `converter.py` - データベクトル変換ツール

機能:
`app.py`で収集したCSVデータを元に、商品画像または商品名をCLIPモデルを用いて512次元のベクトルデータに変換します。

主な特徴:

- 画像とテキストの両方に対応
- 処理の重いベクトル化をバックグラウンドで実行するジョブキューシステム
- 複数のCSVファイルを順番に処理可能

使い方:

```
python converter.py
```

実行後、表示されるURLにアクセスします。「画像から生成」または「商品名から生成」を選択し、処理対象のCSVファイルと画像が保存されているフォルダのパスを指定して「キューに追加」ボタンを押します。

3. `concut.py` - CSV結合ツール

機能:
複数回に分けてスクレイピングした結果のCSVファイルを一つに結合します。

主な特徴:

- 複数のCSVファイルをドラッグ＆ドロップでアップロード
- 指定したカラム（例: `URL`）に基づいて重複データを自動的に削除

使い方:

```
python concut.py
```

実行後、表示されるURLにアクセスします。結合したいCSVファイルをアップロードし、「結合を実行」ボタンを押すと、処理済みのCSVファイルがダウンロードできます。

4. `diagnose.py` - パス診断ツール

機能:
`converter.py`で「画像が見つからない」というエラーが発生した場合に、原因を特定するための補助ツールです。

主な特徴:

- CSVファイルに記載された画像パスと、実際の画像フォルダのパスの整合性をチェック
- パスの間違いや、よくある設定ミス（サブフォルダの指定漏れなど）を検出し、ヒントを表示

使い方:

```
python diagnose.py
```

実行後、表示されるURLにアクセスします。診断したいCSVファイルと、`converter.py`で指定した画像フォルダのパスを入力して「診断開始」ボタンを押すと、ログが表示されます。

セットアップ

1. **必要なライブラリをインストールします。**

```
pip install -r requirements.txt
```

2. **Playwrightのブラウザをインストールします。**

```
playwright install chromium
```

3. **環境変数の設定（任意）**
   `app.py`でAI修復機能を使用する場合、`.env`ファイルを作成し、DeepSeekのAPIキーなどを設定してください。

```
DEEPSEEK_API_KEY="YOUR_API_KEY"
```

注意事項