Müller Crawler Hausarbeit

Dies ist das Software-Projekt **Müller Crawler Hausarbeit IT-Praxis/SoSe2024/Decission-Support-Systems**, das zum Ziel hat, Web-Daten zu crawlen, zu extrahieren, zu analysieren und die Ergebnisse entsprechend in einer SQL DB (SQlite) abzuspeichern.

Im Folgenden werden die benötigten Schritte zur Einrichtung der Software, die Verzeichnisstruktur und die Funktionalität der einzelnen Komponenten des Projekts beschrieben. Es wurde sich auf folgende Kategorie festgelegt **Düfte/Düfte-für-Ihn** URL: https://www.mueller.de/parfuemerie/duefte-fuer-ihn/duefte/

Team

Kristin Mederer, Niklas Rabus, Michael Mark

Einrichtung der Entwicklungsumgebung

Um das Projekt in einer isolierten Umgebung auszuführen, wird empfohlen, ein virtuelles Environment zu erstellen. Im Folgenden sind die Schritte für die Einrichtung unter Windows und Mac beschrieben.

Virtuelles Environment erstellen

Windows

1. Öffne die Eingabeaufforderung (CMD) oder PowerShell.
2. Navigiere zum Projektverzeichnis:

```
cd \Pfad\zum\Projekt\Mueller_Crawler_Hausarbeit_Mederer_Rabus_Mark
```

3. Erstelle ein virtuelles Environment:

```
python –m venv env
```

4. Aktiviere das virtuelle Environment:

```
.\env\Scripts\activate
```

Mac

1. Öffne das Terminal.
2. Navigiere zum Projektverzeichnis:

```
cd /Pfad/zum/Projekt/Mueller_Crawler_Hausarbeit_Mederer_Rabus_Mark
```

3. Erstelle ein virtuelles Environment:

```
python3 -m venv env
```

4. Aktiviere das virtuelle Environment:

```
source env/bin/activate
```

Installation der Abhängigkeiten

Nachdem das virtuelle Environment aktiviert ist, können die erforderlichen Abhängigkeiten installiert werden.

Windows

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

Mac

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

Verzeichnisstruktur

- **DB**
  - Enthält die Datenbank und Datenbank-Einstellungen.
  - Beinhaltet die Datei `models.py`, die die Datenstrukturen definiert.
- **Analyse**
  - Enthält den Unterordner **Data**, in dem die Analyse-Ergebnisse und die Ergebnisse der Queries abgespeichert werden.
- **Output**
  - In diesem Ordner werden die Crawling-Ergebnisse abgespeichert.
  - Enthält ebenfalls die Log-Dateien und die JSON-Daten, die während des Prozesses generiert werden.
- **Scrapers**
  - Beinhaltet alle notwendigen Dateien für den Crawling-Prozess:

- `Extractor`: Extrahiert relevante Daten aus den Webseiten.
- `Analyzer`: Analysiert die extrahierten Daten.
- `Product-Extractor`: Spezifisch für die Extraktion von Produktinformationen.
- `Review-Extractor`: Extrahiert Kundenbewertungen.
- `Web-Crawler`: Verbindet sich mit den Webseiten und sammelt Daten.
- `Link-Extractor`: Extrahiert Links von den zu crawlenden Seiten.
- Diverse Browser-Einstellungen für Chromium.

Hauptdateien im Projektverzeichnis

Im Hauptverzeichnis des Projekts, dem **Müller_Crawler_Hausarbeit_Mederer_Rabus_Mark_Ordner**, befinden sich folgende wichtige Dateien:

- **main.py**
  - Diese Datei startet den Crawler. Durch Ausführung dieser Datei beginnt der Crawling-Prozess.
- **query_all.py**
  - Fragt die Datenbank ab und erstellt CSV-Dateien mit allen Daten der gesamten Tabelle.
- **requirements.txt**
  - Enthält alle notwendigen Abhängigkeiten, die für das Projekt erforderlich sind.
- **run_analysis.py**
  - Nimmt die kompletten Tabellen und analysiert sie mittels `SpaCy`, einem NLP-Tool, um positive und negative Eigenschaften von Produkten zu extrahieren.

Ausführung des Projekts

1. **Crawler starten**: Führe die `main.py` aus, um den Crawler zu starten. Dieser durchläuft die vordefinierten Webseiten und sammelt die notwendigen Daten.
2. **Datenbankabfrage**: Nutze `query_all.py`, um die Datenbank zu durchsuchen und alle Daten in CSV-Form zu extrahieren.
3. **Datenanalyse**: Mit `run_analysis.py` werden die gesammelten Daten analysiert. Diese Analyse wird mittels `SpaCy` durchgeführt, um wertvolle Informationen über die Produkte zu erhalten.

Excel-Auswertung

Im Ordner Excel-Auswertung is noch ein Sheet enthalten welches die erstellten CSV Dateien verarbeitet und in einem Interaktiven Dashboard visualisiert.

Backup - Ordner

Die Projektdateien bleiben leer, damit man einen sauberen Start hat !! Aber... In diesem Ordner befinden sich unsere schon befüllte DB, JSONs und Analyse-Daten. mit dem Stand vom 09.08.2024.