Scraper

Was ist ein Scraper?

Ein Scraper ist ein Programm, das automatisch Daten von Webseiten extrahiert. Es gibt verschiedene Arten von Scrapern, die sich in ihrer Funktionsweise und ihren Anwendungsmöglichkeiten unterscheiden. Wir verwenden zwei Scraper-Bibliotheken: BeautifulSoup und Selenium. Diese unterscheiden sich in ihren Funktionen und Anwendungsbereichen.

BeautifulSoup `BeautifulSoup` ist eine Bibliothek, die es ermöglicht, Daten aus HTML- und XML-Dateien zu extrahieren. Sie ist jedoch nicht in der Lage, Formulare zu bearbeiten oder JavaScript auszuführen, weshalb sie nur für statische Webseiten geeignet ist. Das bedeutet, dass sie lediglich den HTML-Code der Webseite auslesen kann und nicht die dynamischen Inhalte, die durch JavaScript generiert werden (zum Beispiel nach dem Drücken eines Buttons).

Selenium `Selenium` ist ein Webdriver, der es ermöglicht, Webseiten zu steuern und mit ihnen zu interagieren. Ein Webdriver ist ein Programm, das die Steuerung eines Webbrowsers ermöglicht, also tatsächlich ein Browserfenster öffnet und dieses dann steuert. Selenium kann auch dynamische Webseiten auslesen, da es JavaScript ausführen kann. Warum wir Selenium nicht für alle Scraping-Aufgaben verwenden, ist, dass es langsamer ist als BeautifulSoup und auch mehr Rechenleistung benötigt (sowie wir noch herausfinden müssen, wie gut es funktioniert, das Ganze auf dem KI-Server mit GUI laufen zu lassen).

Was brauchen wir für Scraper für unser Projekt?

Für unser Projekt benötigen wir zwei Arten von Scrapern:

Historische Scraper Mit historischen Scrapern sammeln wir Artikel, die in der Vergangenheit auf den Webseiten veröffentlicht wurden. Diese lassen wir einmalig laufen, um das Archiv der Webseiten zu erstellen.

Aktuelle Scraper Die aktuellen Scrapern sind diejenigen, die wir regelmäßig laufen lassen, um kontinuierlich die neuesten Artikel von den Webseiten zu extrahieren.

Der `BaseScraper`

Für die Scraper haben wir eine Basisklasse `BaseScraper` erstellt, die die allgemeinen Funktionen und Methoden enthält, die für alle Scraper benötigt werden. Die Basisklasse enthält die folgenden Methoden:

- `base_metadata_dict()`: Gibt ein Dictionary zurück, das die Metadaten enthält, die wir für jeden Artikel speichern wollen (zum Beispiel Titel, Beschreibung, Datum, usw.).

- `_get_soup()`: Lädt die Webseite und gibt ein `BeautifulSoup` Objekt zurück.
- `_fetch_articles_from_feed()`: Extrahiert die Artikel-URLs aus dem Feed (oder auch anders, falls kein Feed verfügbar) der Webseite.
- `_get_metadata_and_content(url)`: Extrahiert die Metadaten und den Inhalt eines Artikels aus der URL.
- `scrape()`: Führt die einzelnen Schritte zum Scrapen der Webseite aus.

Dadurch müssen wir für jeden neuen Scraper nur noch die spezifischen Methoden implementieren, die für die jeweilige Webseite benötigt werden, und das Ganze ist ein wenig übersichtlicher.

WDR Scraper

Warum WDR? Der WDR Scraper extrahiert Daten von der Webseite des Westdeutschen Rundfunks (WDR). Der WDR ist ein öffentlich-rechtlicher Rundfunksender, der in Nordrhein-Westfalen ansässig ist. Der WDR bietet eine Vielzahl von Inhalten, darunter Nachrichten, Videos, Audios und mehr. Wir haben uns aus verschiedenen Gründen für den WDR als eine der Webseiten entschieden, von der wir Daten extrahieren wollen:

- Der WDR ist ein öffentlich-rechtlicher Sender, was eine gewisse Qualität der Daten sichert.
- Der WDR bietet eine Vielzahl von Inhalten.
- Die Nachrichten-Leicht Seite des WDR und die normalen WDR-Nachrichten sind unter einem Dach, was bedeutet, dass es zu jedem leichten Artikel auch einen normalen Artikel gibt.
- Die leichten Artikel verlinken immer direkt den normalen Artikel, weswegen wir hierfür keinen Matcher brauchen.
- Der WDR bietet Audios an, die von Menschen eingesprochen wurden.
- Der WDR lädt wöchentlich recht viele leichte Artikel hoch, was uns eine gute Datenbasis bietet (ca. 22 Artikel pro Woche).

Funktionsweise Der WDR Scraper extrahiert zunächst Daten von der Nachrichten-Leicht Seite des WDR und lädt die Audios herunter. An die Audios zu kommen war nicht ganz einfach, da der WDR die Audios nicht direkt verlinkt, sondern sie über eine JavaScript-Datei lädt. Daher muss hierfür leider nicht `BeautifulSoup`, sondern `Selenium` verwendet werden. Da die einfachen Artikel immer auf die normalen Artikel verlinken, benötigen wir hierfür keinen Matcher. Im Anschluss an den einfachen Artikel wird der normale Artikel gescraped.