Datenstruktur (von Felix)

Auf Anraten von Professor Bauman wird für die Speicherung keine SQL-Datenbank benutzt sondern wie in der Abbildung dargestellt eine Ordnerstruktur.

.. .. *Abbildung: Struktur des zur Speicherung genutzten Dateisystems am Beispiel von Deutschlandfunk(hard) und Nachrichtenleicht(easy) (reduziert auf jeweils einen Artikel)*

Für jede Nachrichtenquelle findet sich im `data` Verzeichnis ein Unterordner. Da das Matchen lediglich innerhalb derselben Quelle (also z.B. nur Deutschlandfunk zu Nachrichtenleicht, nicht DLF zu MDR) stattfinden soll, findet sich die Datei mit den jeweiligen Matches (`matches_<Nachrichtenquelle>.csv`) auf dieser Ebene (z.B. `deutschlandfunk`). Jedes der Unterverzeichnisse ist wiederrum aufgeteilt in die Ordner `easy` und `hard`, wobei easy die Nachrichten in leichter Sprache enthält und hard die Nachrichten in Standardsprache. Hier findet sich für jeden gespeicherten Artikel ein eigener Ordner mit der Benennungsstruktur `<Jahr>-<Monat>-<Tag>_<Titel>`. Für eine effiziente Suche der Artikel nach ihren jeweiligen Links ist jeweils ein so genannter `lookup-<Nachrichtenquelle>.csv` implementiert. In diesem wir für jeden gespeicherten Artikel jeweils der Dateipfad und der Link im CSV-Format abgespeichert. Im Ordner zum jeweiligen Artikel findet sich jeweils `content.txt`, der Haupttext des Artikels, `metadata.json`, der Verschiedene Metadaten wie URL, Autor und Datum in einem über alle Nachrichtenquellen standarisierten JSON- Format enthält, sowie `audio.mp3`, falls der Artikel als vorgelesene Version als Audio verfügbar ist.

Die Speicherung im eigenen Format bietet viel Flexibilität und Unabhängigkeit von Versionen eines Datenbankmanagementsystems. Allerdings stellt sich die Herausforderung eines komfortablen, einheitlichen und effizienten Zugriffs auf die Daten. Hierfür wurde im Projekt die Klasse `DataHandler` definiert. Diese bietet ein Interface für den **Zugriff** auf die Daten durch Funktionen wie `head`, welcher die ersten n Artikel als Pandas DataFrame zurück gibt. Des Weiteren soll eine einheitliche **Speicherung** durch vordefinierte Speicherfunktionen sichergestellt werden. Auch ermöglicht der DataHandler eine **Suche** im Verzeichnis nach Metadaten. Um keine Artikel doppelt zu Scrapen gibt es außerdem die Funktion `is_already_saved`, welche effizient über den Lookuptable (ohne das gesamte Unterverzeichnis zu durchsuchen) zurückgibt, ob die URL bereits gescraped und gesaved wurde. Das DataHandler Objekt muss mit der jeweiligen Nachrichtenquelle initialisiert werden (aktuell `"dlf"`, oder `"mdr"`) und kann dann für das jeweilige Unterverzeichnis genutzt werden. Die Initialisierung mit der Nachrichtenquelle soll unter anderem einer Vermischung der Daten vorbeugen. Den meisten Funktionen muss übergeben werden, ob im `"hard"` (`"h"`), oder `"easy"` (`"e"`) Unterverzeichnis gelesen oder geschrieben werden soll.

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.

Deutschlandradio

Deutschlandradio ist ein integraler Bestandteil des öffentlich-rechtlichen Rundfunks und verantwortlich für die Produktion verschiedener Nachrichtenangebote, darunter Deutschlandfunk und Nachrichtenleicht, die als bedeutende Datenquellen dienen. Eigenständige Redaktionen sind dafür zuständig, die jeweiligen

Inhalte zu konzipieren und zu veröffentlichen. Deutschlandfunk

Die Gründe für die Berücksichtigung von Deutschlandradio sind vielfältig:

```
Es ist Teil des öffentlich-rechtlichen Rundfunks.
Seine Reichweite erstreckt sich über ganz Deutschland.
Es bietet eine breite inhaltliche Abdeckung aktueller Themen, darunter Politik, Wirtschaft,
Es besteht eine redaktionelle Nähe zwischen Deutschlandfunk und Nachrichtenleicht.
Die Inhalte werden von Menschen erstellt, sowohl in Form von Artikeln als auch von Audios.
```

Nachrichtenleicht Auf der Internetseite von Nachrichtenleicht werden jeden Freitagnachmittag etwa 5-6 Artikel in leicht verständlicher Sprache veröffentlicht. Zusätzlich werden die Artikel häufig als Audio angeboten, wobei sie von menschlichen Sprechern eingesprochen werden.

Weitere Nachrichtenangebote

Zur Auswahl standen auch die Nachrichtenangebote der APA (Austria Presse Agentur), des NDR und des SR. Die APA bietet Nachrichten in leichter Sprache an, die von capito.ai generiert werden, einem vollautomatisierten KI-Tool zur Übersetzung von Texten aus der Standardsprache in leicht verständliche Sprache. Auf ihrer Website sind die Originalartikel sowie Übersetzungen in die Sprachniveaus B1 und A2 verfügbar. Wir haben uns gegen die Verwendung der APA als Datenquelle entschieden, da keine Audioversionen der Artikel vorhanden sind und die Artikel in leichter Sprache ausschließlich von KI generiert werden, was zu einem möglichen Bias in den Daten führen könnte. Darüber hinaus erreichen die Artikel nicht die qualitative Standards der öffentlich-rechtlichen Sender. Sowohl der NDR als auch der SR sind Mitglieder des öffentlich-rechtlichen Rundfunkverbunds ARD und bieten ebenfalls Nachrichten in leichter Sprache mit Audio an. Allerdings unterscheiden sich die Texte in leichter Sprache formell stark von denen des Deutschlandradios und des MDRs. Zudem gibt es keinen separaten Nachrichtenfeed, was das Scrapen der Artikel erschwert. Das Angebot des SR in leichter Sprache konzentriert sich hauptsächlich auf regionale Nachrichten aus dem Saarland. Aus diesen Gründen haben wir vorläufig beschlossen, diese Nachrichtenangebote nicht zu berücksichtigen.

Scraping

Dank der redaktionellen Nähe zwischen Deutschlandfunk und Nachrichtenleicht sind die Internetseiten größtenteils strukturell identisch aufgebaut. Dadurch konnte ein DeutschlandradioScraper basierend auf dem BaseScraper entwickelt werden, um Redundanzen zu vermeiden. Die Unterschiede liegen hauptsächlich in den Metadaten und der Verfügbarkeit von Audio bei Nachrichtenleicht-Artikeln. Ausgehend vom DeutschlandradioScraper konnten entsprechende Scraper für Deutschlandfunk und Nachrichtenleicht abgeleitet werden. Für den Nachrichtenleicht-Feed wurde eine API-Schnittstelle gefunden, die das Scrapen erleichtert.