block6-anno-scraper

September 20, 2018

1 Block 6 Web Scraping Zeitungsportal Anno

URL: Übersicht der verfügbaren Zeitungen:

- http://anno.onb.ac.at/alph_list.htm

Das Skript kann verschiedene txt-Dateien aus dem österreichischen Zeitungsportal Anno herunter. Die txt-Dateien sind per OCR erfasst; Erkennungsfehler sind reichlich vorhanden.

1.1 Ermittlung der URL

- Zeitung auswählen
- Jahresübersicht auswählen
- Ausgabe auswählen
- Anzeige als txt auswählen
- am Ende der URL die Seitenzahl durch x ersetzen; auf diese Weise wird die gesamte Ausgabe ausgewählt
- in der URL findet sich ein Datumformat in der Form yyyymmdd; auf diese Weise kann in einer Schleife das Datum genutzt werden

1.2 Das Skript

```
In [ ]: import requests
        import time
        import pandas as pd
        from datetime import date
```

URL

```
In [ ]: url_root = "http://anno.onb.ac.at/cgi-content/annoshow?text=dar|" # URL der Zeitung ei
        url_root_2 = "|x"  # steht für alle Seiten einer Ausgabe
        name_zeitung = "die-arbeit"
```

Beginn und Ende der Schleife im Datumsformat

```
In [ ]: start_date = date(1886, 1, 1) # Start des Zeitraums
        end_date = date(1886, 1, 22) # Ende des Zeitraums
```

Mit Pandas ein date_range-Objekt erstellen

```
In [ ]: daterange = pd.date_range(start_date, end_date)
```

Über das date_range-Objekt iterieren

```
In [ ]: for date in daterange:

            date_id = date.strftime("%Y%m%d")

            response = requests.get(url_root + date_id + url_root_2)
            text = response.text

            print(date_id)

            if len(text)!= 0:

                # speichern der einzelnen Ausgabe
                with open(date_id + "-" + name_zeitung + ".txt", "w", encoding="utf-8") as file
                    file.write(text)

                # speichern eines gesamten Jahrgangs
                with open("1886-jahrgang-" + name_zeitung + ".txt", "a", encoding="utf-8") as f
                    file.write(text + "\n")

            time.sleep(2)
```