CS 839 Stage 2 Report

Crawling and extracting structured data from Web pages

Xiuyuan He, Chenlai Shi, Mingren Shen

Description of Web Data Sources

We basically extracted data science books information, i.e. using “data science” as keywords in search, from library databases of UW-Madison and UIUC.

- UW-Madison

https://search.library.wisc.edu/search/system?q=Data+Science

- UIUC

https://vufind.carli.illinois.edu/vf-uiu/Search/Home?lookfor=Data+Science+&type=all&start_over=1&submit=Find&search=new

Data Extraction Process

The process to extract structured data from the website is described in the following,

(1) Input the URL of of the searching result pages of both libraries(the two URL described above) and get the source code of this webpage.
(2) From the source code of research results, extract all the URL for details pages of each book and store this URL into a dictionary where the key is the URL and value is the book title.
(3) Do step 2 for next search result pages until we get 10000 different URLs
(4) For each URL of the book detail page,
    (a) Obtain the source code of the book pages
    (b) Get the data field form the source code and store them in a pandas dataframe
        (i) If the item name already contain in the dataframe, update data of that column
        (ii) If not, create a new column to store that data, dataframe insert NaN for missing value automatically and output a warning.

(5) Then finally we post-process the result CSV to make them have the same schema and format. We also deal with Null values and mismatches in this steps.

Code for Step (1) to (4) is contained in webcrawl_UWM.py and webcrawl_UIUC.py and code for Step (5) is contained in Project2_processing.ipynb.

We also uploaded the log files of the scripts running in Github named log_UWM.txt and log_UIUC.txt.

Table Description

The entity we want to extract is book information like ISBNs, Title, authors. etc. The tuples of 7 entities extracted include

1. Title: book title, string
2. Author: book authors, string
3. Publication:information of publishers, string
4. Format: book types, category, e.g. journal, magazines
5. ISBN: Search Results International Standard Book Number,integer
6. Series: book series information, string
7. Physical Details:Physical description in book cataloging, string

- Table A-UIUC: 6963 tuples
- Table B-UW-Madison: 5730 tuples

Open-source Tools

- BeautifulSoup

Python library for pulling data out of HTML and XML files.

- Pandas

Open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

- Requests

Requests is an elegant and simple HTTP library for Python, built for human beings.

- Jupyter Notebook

Jupyter Notebook was used to deal with Null values and mismatches.