Extract Structed Data of Books from Amazon and Barnes & Noble

Lan Bai, Chaoqun Mei, Yuzhe Ma

March 22, 2018

Abstract In this project stage, we used the open-source tool Scrapy to extract structured data of books from Amazon and Barnes & Noble. For each book, the publisher, time, author, and title attributes were extracted as a tuple. We finally got two tables in CSV format, one has 3077 tuples for Amazon books, and the other has 3179 tuples for Barnes & Nobel books.

1. Data Sources

We selected Amazon books and Barnes & Noble books as the two Web data sources for this project. Since there are billions of books on Amazon books and Barnes & Noble books, we narrowed down to fiction, health, history, and business books on Amazon books, and history, health, business, and science books on Barnes & Noble books. For each book, the publisher, time, author, and title were extracted as a tuple.

2. Data Extraction

We extracted the structured data of books from Amazon and Barnes & Noble using Scrapy. Scrapy is a free and open source web crawling framework, written in Python. We first crawl pages that are lists of books and their URLs and extract book page URLs. Then publisher, time, author, and title of books were extracted from book pages by CSS. If there was a break during the extracting, we continued the extraction from a new start URL.

We construct rules to extract information by locating target information in source code of one page and comparing several pages.

Rules for Amazon books:

Title of books: inside <span> </span> tag whose id is “productTitle”.

Author of books: located by the hierarchy of three tags td, div and span whose class is “a-size-based”, “a-row” and “a-size medium” respectively.

Publisher: located by the hierarchy of three tags div (class = “content”), ul, and li, the prefix “Publisher” and the suffix “(”.

Publish time: inside <span> </span> tags whose class is “a-text-normal”. The format is “Month(letters) Day(digits), Year(digits)” and match by regex.

Rules for Barnes Noble books:

Title of books: indicated by the attribute “content” inside the element <meta> whose “property” entry has value “og:title”.

Author of books: indicated by the text inside element <a>, whose “itemprop” entry has value “author”.

Publisher: indicated by the text inside element <a>, whose “href” entry contains the string “Ntk=Publisher”.

Publish time: the text inside element <td>, which is the third child element of <tr>, which instead is a child element of <tbody>.

3. Type of Entity and Description of the Two Tables

The entity we extracted is book, we extracted the information of 3079 books from Amazon books, and 3179 from Barnes & Noble books. For each of the two Web data sources, we finally got a table in CSV format. There are 3078 tuples in the table for Amazon books, and 3179 tuples in the table for Barnes & Nobel books. Each tuple has four attributes, namely, publisher, time, author, and title.

4. Tool

We used the open-source tool Scrapy in this project stage. Scrapy is an application framework written in Python for crawling web sites and extracting structured data which can be used for a wide range of useful applications, like data mining, information processing or historical archival. Even though Scrapy was originally designed for web scraping, it can also be used to extract data using APIs (such as Amazon Associates Web Services) or as a general purpose web crawler.