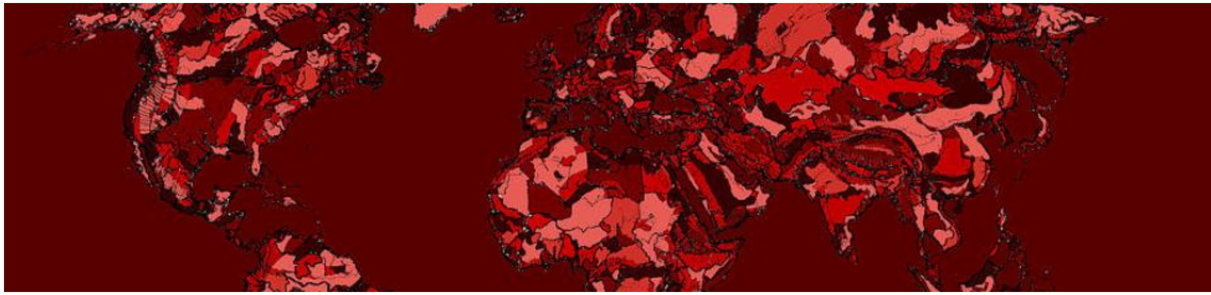
An automated data pipeline for scraping book information from an online bookseller and storing it to SQL/ NoSQL databases

Term project of COM506 – Data Management

Jeonghun Song (20212210010), September 2022

Swiss Institute of
Artificial Intelligence



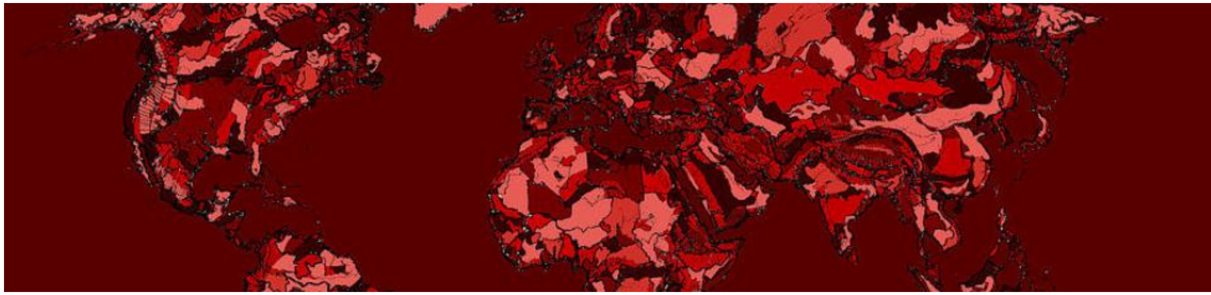
Contents

Section

Introduction

Swiss Institute of
Artificial Intelligence

RERUM
COGNOSCERE CAUSAS



Background

- Big online booksellers viewed as a database

■ Big online booksellers sell most of the existing books, including technical books of certain area of expertise.

■ Also, big online booksellers provide not only product details of a book but also evaluations on the book based on ratings and reviews of many people.

■ Thus, big online booksellers is not only an online website for purchasing books but also a large database of information of books.

Problem statement
- Lack of service offering for book data acquisition

- People may want to gather data of book information in a certain area of expertise for the following purposes:
  - Deciding books with higher priority in a specific field with multiple searchwords
  - Investigation on authors or publishers who are eminent in a specific field
  - Estimation of popularity trend in a specific field by investigation on time-based publishing records

- However, most of online booksellers do not provide a convenient way to get bulk data of books. In other words, there is no service offering to enable data acquisition via sending queries or open APIs.

- For the purpose, a customized data pipeline for web-scraping book information and storing the scrapped data to common SQL/ NoSQL databases is required.

Development Objective

- Building a data pipeline for scraping book data from a big online bookseller and storing the data to databases

- ■ **The objective of this development is to build an automated data pipeline which provides the following functions:**
  - – **Collects information of books (including reviews) from a target online bookseller for given searchwords**
  - – **Processes the data and stores to SQL/ NoSQL DBs**
  - – **Sends queries for generating desired subsets of the stored book data.**

- ■ **The target online bookseller is BookDepository.com, which allows scraping HTML documents of its webpages[1].**

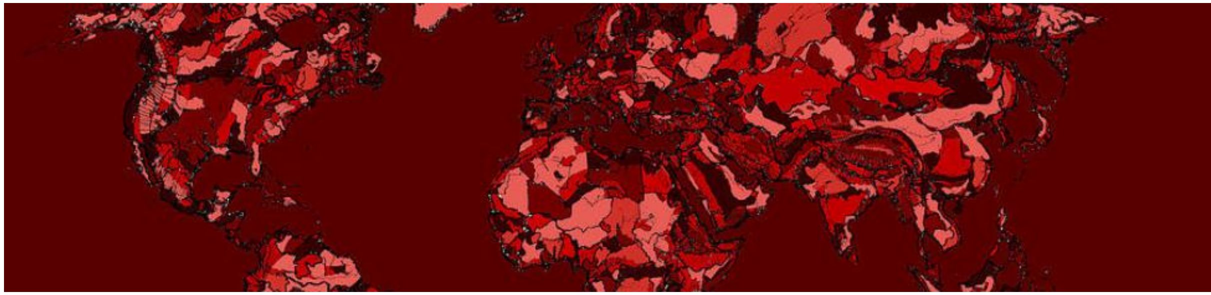
- ■ **The target attributes are book title, author, price, publisher, description written by the author, publication date, page number, ISBN-13 (book identifier), rating, and reviews.**

- ■ **SQLite and Elasticsearch have been selected as the SQL/ NoSQL DBs for data storage.**

[1] Amazon books does not allow scraping HTML documents of its webpage.

Methodology

Swiss Institute of
Artificial Intelligence



Configuration

- Collect – Process – Store by a data pipeline, and Query by separated codes

- Using python 3, a data pipeline which automates collect – process – store stages has been built, to collect information of books for a given searchword and store the results to SQLite and Elasticsearch DBs. For sending queries, separated python codes are provided.

Configuration of the data pipeline

Aggregated & Automated

	Collect	Process	Store	Query
Functions	• Get HTMLs from BookDepository.com and Goodreads.com • Parse the HTMLs • Create a initial table of book information	• Cleanse the initial table (dropping duplicates and unavailable books) • Create additional tables (author, publisher) • Create a list of dictionaries containing description and reviews	• Store the three tables to SQLite DB • Store the list of dictionaries to Elasticsearch DB (A docker container running Elasticsearch is required)	• Send a query to SQLite DB for getting a table of book information, and save it as a CSV file • Send a query to Elasticsearch DB for getting lists of book reviews, and save it as a JSON file
Packages	urllib[1]		sqlite3	
	BeautifulSoup		elasticsearch	
	pandas			

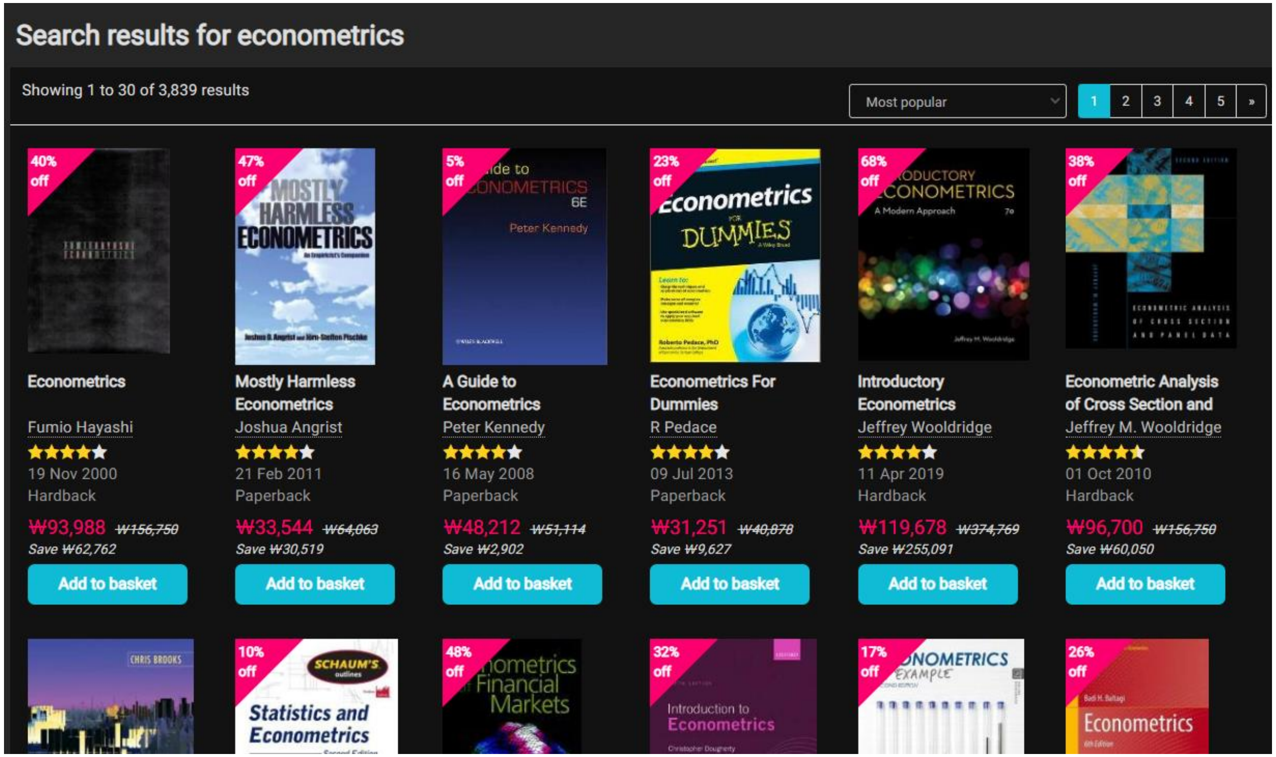
[1] Using requests package, wrong html document of product details page is obtained occasionally. Therefore, urllib package has been used.

Collect stage: scraping the bookseller website

- Search results page of BookDepository.com

■ For a given searchword and a given number of books, HTML documents of each page of search results are scraped and parsed to get the link of the product details page of the books.

Search results page of BookDepository.com



Search results for econometrics

Showing 1 to 30 of 3,839 results

Econometrics
Fumio Hayashi
19 Nov 2000
Hardback
₩93,988 ~~₩156,750~~
Save ₩62,762

Mostly Harmless Econometrics
Joshua Angrist
21 Feb 2011
Paperback
₩33,544 ~~₩64,063~~
Save ₩30,519

A Guide to Econometrics
Peter Kennedy
16 May 2008
Paperback
₩48,212 ~~₩51,114~~
Save ₩2,902

Econometrics For Dummies
R Pedace
09 Jul 2013
Paperback
₩31,251 ~~₩40,878~~
Save ₩9,627

Introductory Econometrics
Jeffrey Wooldridge
11 Apr 2019
Hardback
₩119,678 ~~₩374,769~~
Save ₩255,091

Econometric Analysis of Cross Section and
Jeffrey M. Wooldridge
01 Oct 2010
Hardback
₩96,700 ~~₩156,750~~
Save ₩60,050

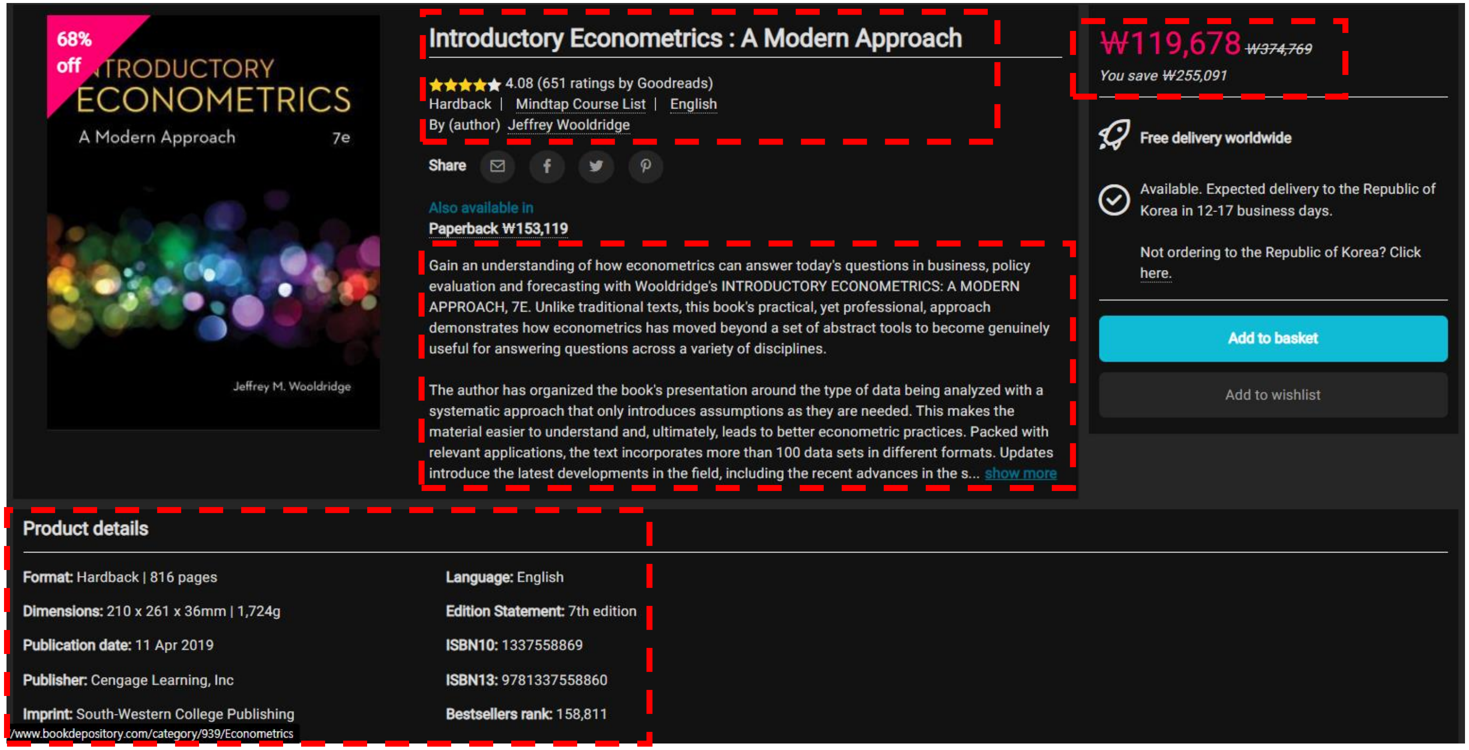
Source: Bookdepository.com

Collect stage: scraping the bookseller website

- Product details page of BookDepository.com

- For each book, the HTML document is scraped and parsed to scrap title, author, description, price, rating, and product details such as publication date, publisher, ISBN-13.

Product details page of BookDepository.com



Introductory Econometrics : A Modern Approach

★★★★★ 4.08 (651 ratings by Goodreads)
Hardback | Mindtap Course List | English
By (author) Jeffrey Wooldridge

₩119,678 ~~₩374,769~~
You save ₩255,091

Free delivery worldwide

Available. Expected delivery to the Republic of Korea in 12-17 business days.

Not ordering to the Republic of Korea? Click here.

Also available in
Paperback ₩153,119

Gain an understanding of how econometrics can answer today's questions in business, policy evaluation and forecasting with Wooldridge's INTRODUCTORY ECONOMETRICS: A MODERN APPROACH, 7E. Unlike traditional texts, this book's practical, yet professional, approach demonstrates how econometrics has moved beyond a set of abstract tools to become genuinely useful for answering questions across a variety of disciplines.

The author has organized the book's presentation around the type of data being analyzed with a systematic approach that only introduces assumptions as they are needed. This makes the material easier to understand and, ultimately, leads to better econometric practices. Packed with relevant applications, the text incorporates more than 100 data sets in different formats. Updates introduce the latest developments in the field, including the recent advances in the s... show more

Product details

Format: Hardback | 816 pages
Dimensions: 210 x 261 x 36mm | 1,724g
Publication date: 11 Apr 2019
Publisher: Cengage Learning, Inc
Imprint: South-Western College Publishing
Language: English
Edition Statement: 7th edition
ISBN10: 1337558869
ISBN13: 9781337558860
Bestsellers rank: 158,811

/www.bookdepository.com/category/939/Econometrics

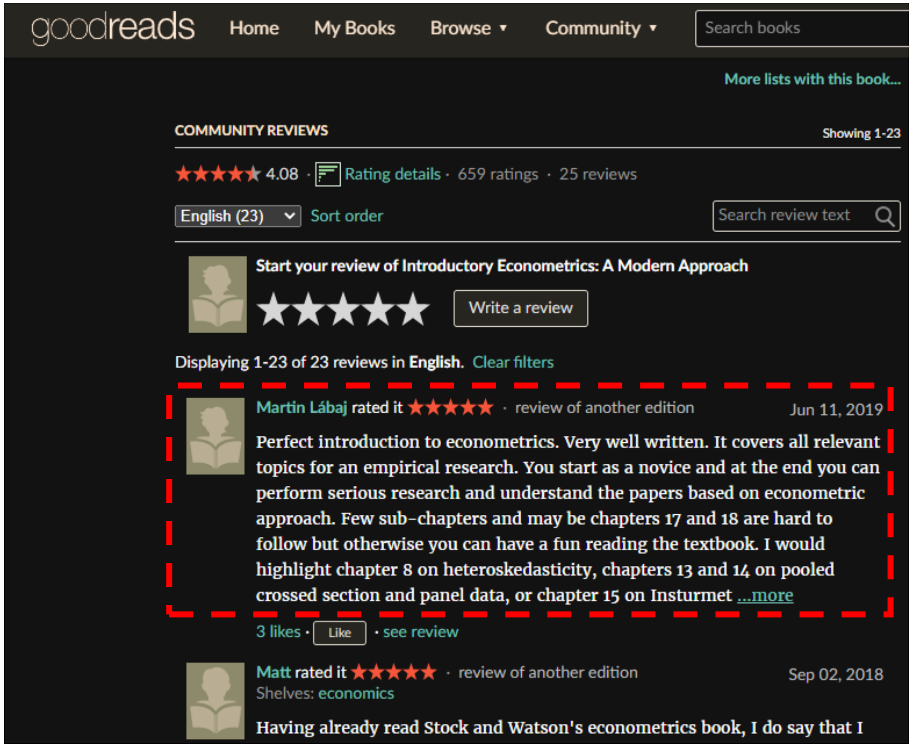
Source: Bookdepository.com

Collect stage: scraping the external website for book reviews

- Review page of GoodReads.com

- For each book, the HTML document of the page of the external website for book reviews is scraped and parsed to gather reviews on the book.

Review page of a book in GoodReads.com



Start your review of Introductory Econometrics: A Modern Approach

Displaying 1-23 of 23 reviews in **English**.

Martin Lábaj rated it ★★★★★ · review of another edition — Jun 11, 2019

Perfect introduction to econometrics. Very well written. It covers all relevant topics for an empirical research. You start as a novice and at the end you can perform serious research and understand the papers based on econometric approach. Few sub-chapters and may be chapters 17 and 18 are hard to follow but otherwise you can have a fun reading the textbook. I would highlight chapter 8 on heteroskedasticity, chapters 13 and 14 on pooled crossed section and panel data, or chapter 15 on Insturmet ...more

Matt rated it ★★★★★ · review of another edition — Sep 02, 2018

Having already read Stock and Watson's econometrics book, I do say that I

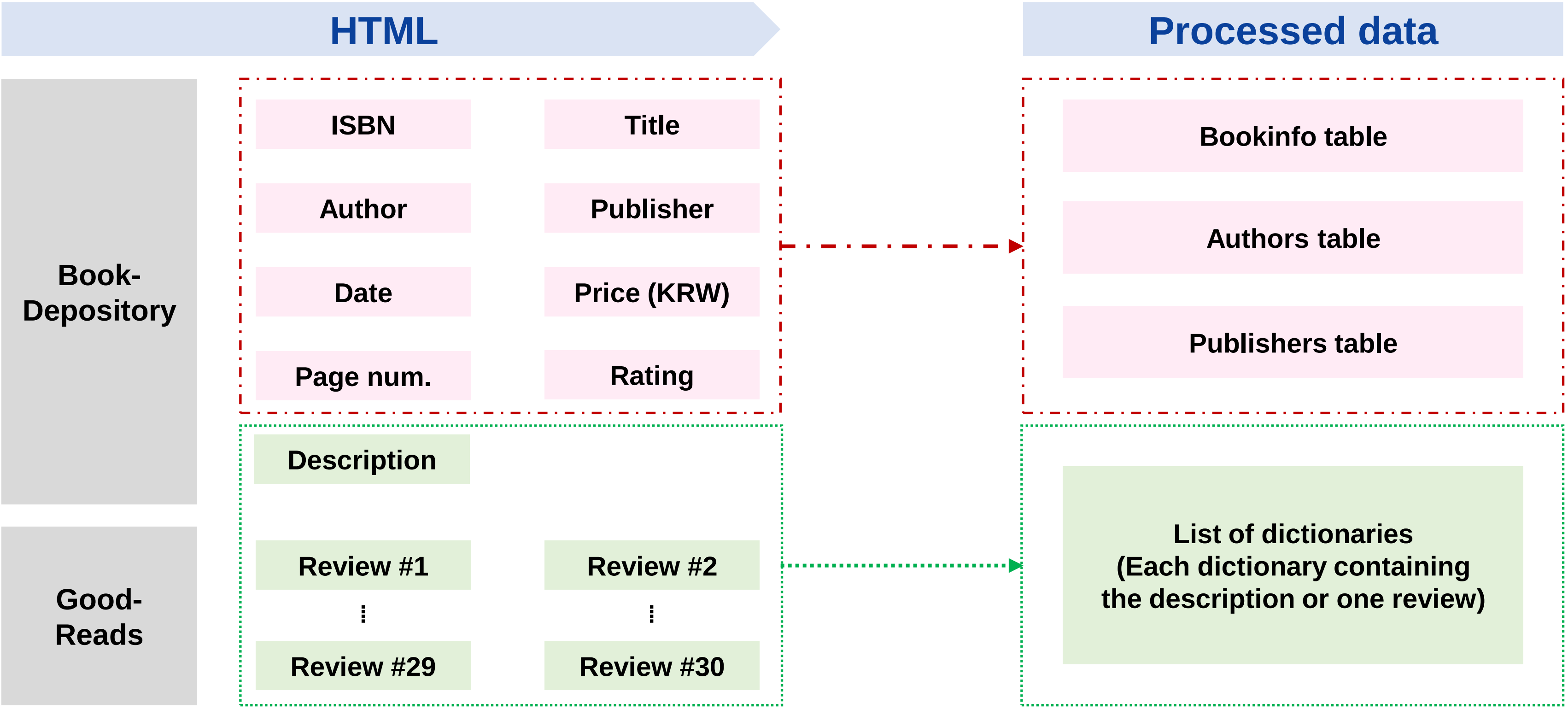
Source: Goodreads.com

Process stage: Data transformation and cleansing

- Conversion into tables and list of dictionaries, and dropping useless parts

■ The information from the parsed HTML documents are converted to i) three tables containing product details, and ii) one list of dictionaries containing description and reviews. Duplicates (ex> hardcover and paperback) and unavailable books are dropped.

Data processing in the pipeline

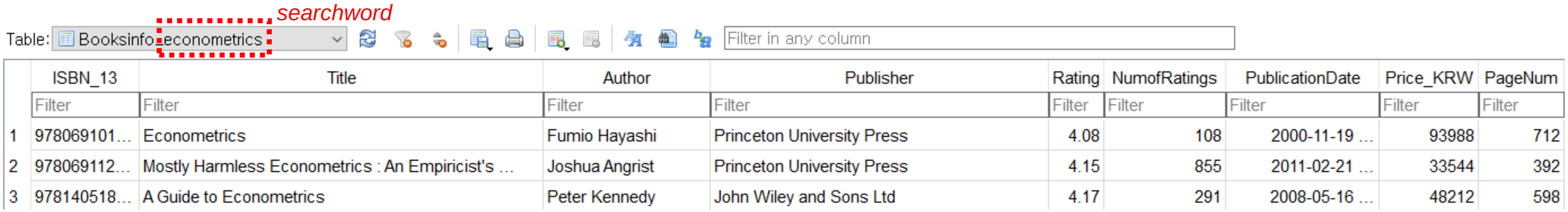


Store stage: Storing tables to SQLite DB

- Three tables with foreign key constraints for each searchword

- Authors and Publishers tables are additionally created to arrange authors and publishers corresponding to the searchword, and show their bestseller and the number of ratings of the bestseller.

Sample of Booksinfo table (searchword: econometrics)

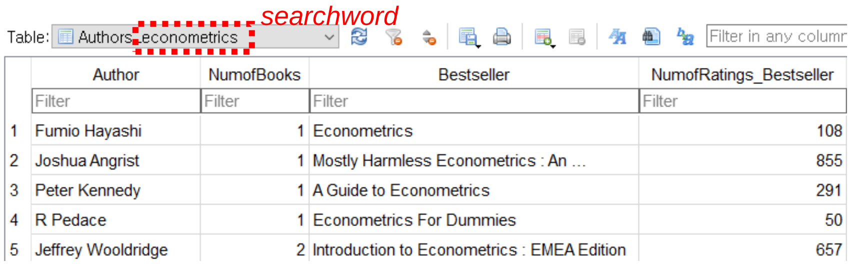


searchword

Table: Booksinfo_econometrics

	ISBN_13	Title	Author	Publisher	Rating	NumofRatings	PublicationDate	Price_KRW	PageNum
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	978069101...	Econometrics	Fumio Hayashi	Princeton University Press	4.08	108	2000-11-19 ...	93988	712
2	978069112...	Mostly Harmless Econometrics : An Empiricist's ...	Joshua Angrist	Princeton University Press	4.15	855	2011-02-21 ...	33544	392
3	978140518...	A Guide to Econometrics	Peter Kennedy	John Wiley and Sons Ltd	4.17	291	2008-05-16 ...	48212	598

Sample of Authors table

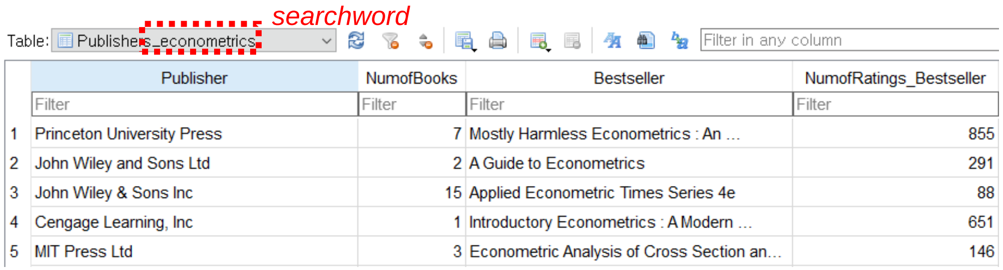


searchword

Table: Authors_econometrics

	Author	NumofBooks	Bestseller	NumofRatings_Bestseller
	Filter	Filter	Filter	Filter
1	Fumio Hayashi	1	Econometrics	108
2	Joshua Angrist	1	Mostly Harmless Econometrics : An ...	855
3	Peter Kennedy	1	A Guide to Econometrics	291
4	R Pedace	1	Econometrics For Dummies	50
5	Jeffrey Wooldridge	2	Introduction to Econometrics : EMEA Edition	657

Sample of Publishers table



searchword

Table: Publishers_econometrics

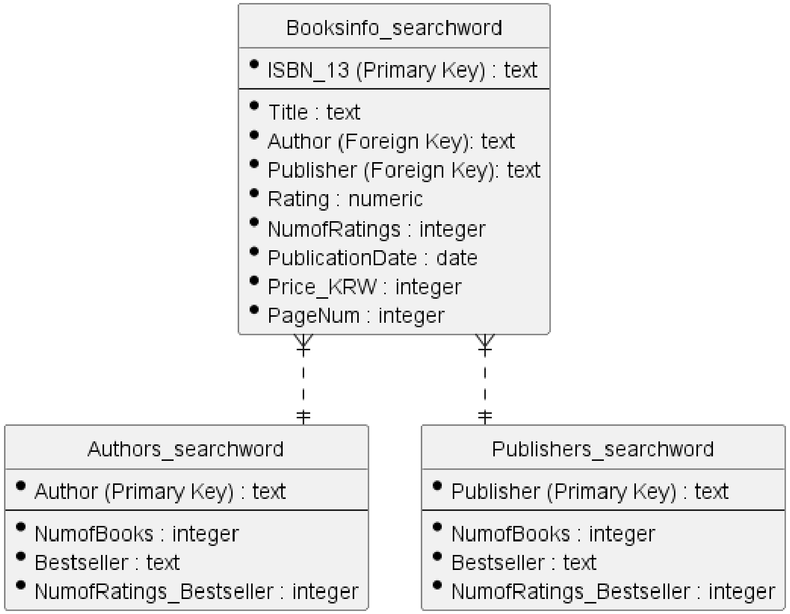
	Publisher	NumofBooks	Bestseller	NumofRatings_Bestseller
	Filter	Filter	Filter	Filter
1	Princeton University Press	7	Mostly Harmless Econometrics : An ...	855
2	John Wiley and Sons Ltd	2	A Guide to Econometrics	291
3	John Wiley & Sons Inc	15	Applied Econometric Times Series 4e	88
4	Cengage Learning, Inc	1	Introductory Econometrics : A Modern ...	651
5	MIT Press Ltd	3	Econometric Analysis of Cross Section an...	146

Store stage: Storing tables to SQLite DB

- Three tables with foreign key constraints for each searchword

- **The primary key of Booksinfo table is ISBN-13 (book identifier). The primary keys of Authors and Publishers tables are author and publisher, respectively, because these attributes are unique in each of these tables, though not unique in Booksinfo table. Author and publishers are foreign keys in Booksinfo table.**

Entity-Relationship Diagram of the tables



Booksinfo_searchword
- ISBN_13 (Primary Key) : text
- Title : text
- Author (Foreign Key): text
- Publisher (Foreign Key): text
- Rating : numeric
- NumofRatings : integer
- PublicationDate : date
- Price_KRW : integer
- PageNum : integer

Authors_searchword
- Author (Primary Key) : text
- NumofBooks : integer
- Bestseller : text
- NumofRatings_Bestseller : integer

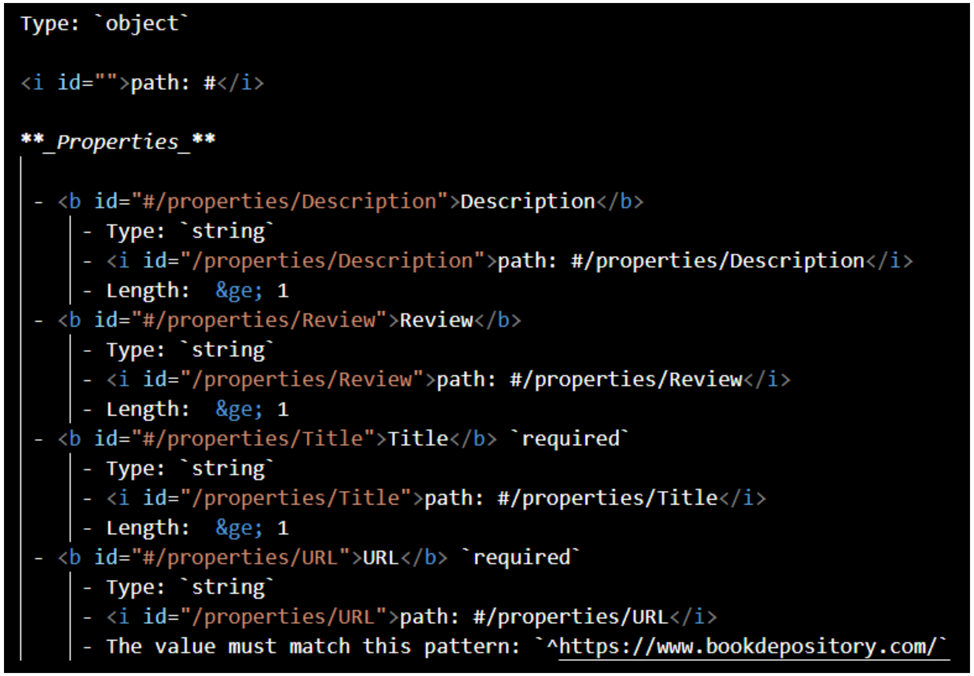
Publishers_searchword
- Publisher (Primary Key) : text
- NumofBooks : integer
- Bestseller : text
- NumofRatings_Bestseller : integer

Store stage: Storing reviews to Elasticsearch DB

- One index for each searchword

- ■ **For each searchword, an index is created to store the description and reviews on the books (up to 30 for each) to Elasticsearch DB[1].**

JSON Schema



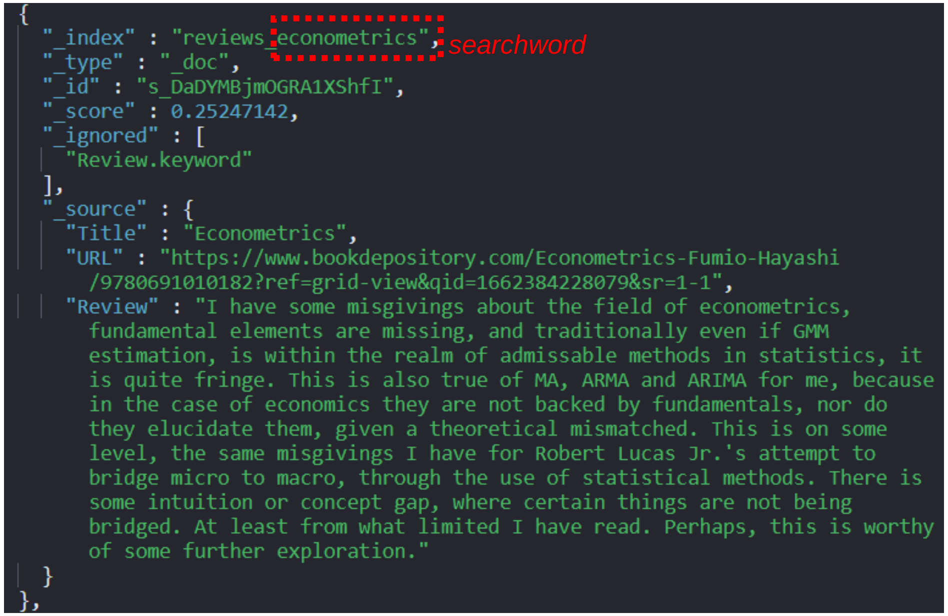
```
Type: `object`

<i id="">path: #</i>

**_Properties_**

 - <b id="#/properties/Description">Description</b>
    - Type: `string`
    - <i id="/properties/Description">path: #/properties/Description</i>
    - Length:  &ge; 1
 - <b id="#/properties/Review">Review</b>
    - Type: `string`
    - <i id="/properties/Review">path: #/properties/Review</i>
    - Length:  &ge; 1
 - <b id="#/properties/Title">Title</b> `required`
    - Type: `string`
    - <i id="/properties/Title">path: #/properties/Title</i>
    - Length:  &ge; 1
 - <b id="#/properties/URL">URL</b> `required`
    - Type: `string`
    - <i id="/properties/URL">path: #/properties/URL</i>
    - The value must match this pattern: `^https://www.bookdepository.com/`
```

Sample of index (searchword: econometrics)



```
{
  "_index" : "reviews_econometrics",
  "_type" : "_doc",
  "_id" : "s_DaDYMBjmOGRA1XShfI",
  "_score" : 0.25247142,
  "_ignored" : [
    "Review.keyword"
  ],
  "_source" : {
    "Title" : "Econometrics",
    "URL" : "https://www.bookdepository.com/Econometrics-Fumio-Hayashi
      /9780691010182?ref=grid-view&qid=1662384228079&sr=1-1",
    "Review" : "I have some misgivings about the field of econometrics,
      fundamental elements are missing, and traditionally even if GMM
      estimation, is within the realm of admissable methods in statistics, it
      is quite fringe. This is also true of MA, ARMA and ARIMA for me, because
      in the case of economics they are not backed by fundamentals, nor do
      they elucidate them, given a theoretical mismatched. This is on some
      level, the same misgivings I have for Robert Lucas Jr.'s attempt to
      bridge micro to macro, through the use of statistical methods. There is
      some intuition or concept gap, where certain things are not being
      bridged. At least from what limited I have read. Perhaps, this is worthy
      of some further exploration."
  }
},
```

searchword

[1] *A docker container containing Elasticsearch must be in running.*

Query stage: Sending queries to the DBs

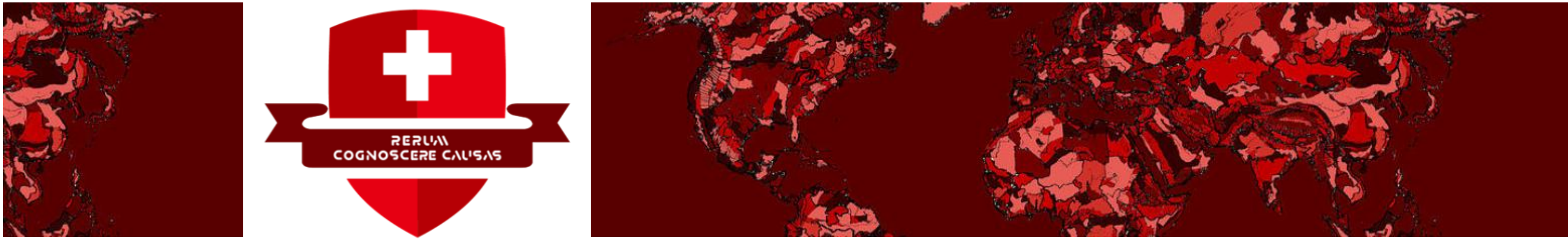
- Codes separated from the data pipeline

- After the information of books for a given searchword is stored in the DBs, user can send queries to the DBs in separated python codes.
- One python code is for SQLite DB, and the other python code is for Elasticsearch DB[1].
- The results of the queries are saved as CSV files (SQLite) and JSON files (Elasticsearch).

1) A docker container containing Elasticsearch must be in running.

Results

Swiss Institute of
Artificial Intelligence



Data pipeline and query operation
- Confirmed to work well

- The data pipeline works well.

- Among the three stages of the pipeline, collect stage takes the most of the time.

Elapsed time : scraping 100 books for searchword “econometrics”

Stage	Collect	Process	Store	
			SQLite	Elasticsearch
Time [sec.]	524.2	0.3	0.7	0.4

- For storing data to Elasticsearch DB, bulk upload (bulk function of helpers module) has been used instead of using PUT for every book, to achieve quicker upload of the set of many reviews.

- The separated codes for sending queries and saving the results also works well.