

[과제 4] Web Data Scraping 학습

순서대로 과제를 수행한 후 아래 4개의 파일을 제출하세요.

- R script 파일
- books.csv 파일
- multi_books.csv 파일
- 문제의 정답이 적힌 pdf 파일

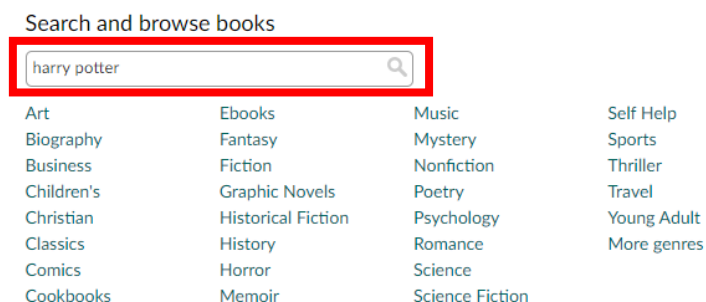
[Notice] 본 과제는 Good Reads 사이트에서 서적의 제목, 저자, 평점, 평점의 수, 리뷰 개수, 요약문을 Scraping 하는 것을 목적으로 합니다. 문제를 풀며 R script를 함께 작성하는 것을 권장합니다. 본 과제에 제시된 Step은 가이드라인으로, 다른 방법으로 같은 결과를 도출할 수 있다면 다른 방법을 사용하여도 무관합니다. 본 과제의 Task는 다소 성능이 좋은 PC를 사용하더라도 3시간 이상이 소요됩니다. 이를 고려하여 과제를 수행해 주세요.

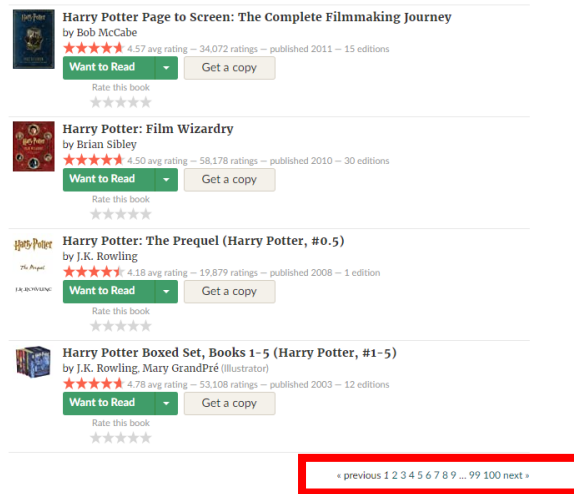
[Step 1] Good Reads 웹 사이트(<https://www.goodreads.com/>)에 접속하세요.

#Good Reads란?

다양한 서적과 서적에 대한 평점, 리뷰를 확인할 수 있는 사이트

[Step 2] 페이지 중앙의 Search and browse books에 “harry potter”를 검색하세요. 이후 맨 아래에 페이지를 변경해가며 URL의 특징을 파악하세요.





문제 1: 페이지를 이동할 경우 URL이 어떻게 변화하는지 서술하세요.

[Step 3] 수업시간에 배운 내용을 바탕으로 필요한 library를 호출하고, 제목, 저자, 평점, 평점의 수, 리뷰 개수, 요약문에 관한 변수를 선언하세요.

[Step 4] 이번 과제에서는 검색 키워드가 변화하더라도 마지막 페이지까지 Scraping을 수행할 수 있도록 코드를 구현하고자 합니다. 첫 번째 페이지를 기준으로 문제 1의 결과를 이용하여 아래 그림의 마지막 페이지인 100을 정수 타입으로 변수에 할당할 수 있는 코드를 작성하세요. (변수에 100을 직접 할당하는 방식은 허용되지 않습니다.)

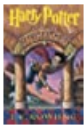
#Hint: length() 함수와 as.integer() 함수를 사용하세요!




[Step 5] Step 4의 결과를 바탕으로 처음 페이지부터 마지막 페이지까지 URL을 통해 접근할 수 있는 코드를 작성하세요.

[Step 6] 아래의 페이지에는 주어진 검색 결과에 대한 서적의 제목, 저자, 평점 등이 등장합니다. 하지만 본 과제에서는 서적의 제목을 클릭하여 이동되는 새로운 페이지에서 Data를 Scraping 합니다. 서적 제목의 링크로 이동할 수 있는 URL을 수집하여 보세요. 이후에 직접 링크를 클릭하여 URL을 확인하세요.


Page 1 of about 4940 results (0.11 seconds)



Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
by J.K. Rowling
★★★★★ 4.47 avg rating — 7,071,883 ratings — published 1997 — 791 editions
[Want to Read](#) [Get a copy](#)
Rate this book
★★★★★



Harry Potter and the Order of the Phoenix (Harry Potter, #5)
by J.K. Rowling, Mary GrandPré (Illustrator)
★★★★★ 4.50 avg rating — 2,515,626 ratings — published 2003 — 428 editions
[Want to Read](#) [Get a copy](#)
Rate this book
★★★★★



Harry Potter and the Goblet of Fire (Harry Potter, #4)
by J.K. Rowling
★★★★★ 4.56 avg rating — 2,602,650 ratings — published 2000 — 494 editions
[Want to Read](#) [Get a copy](#)
Rate this book
★★★★★

문제 2: 수집된 URL과 직접 클릭하여 이동한 페이지의 URL에 대한 차이점을 설명하세요.

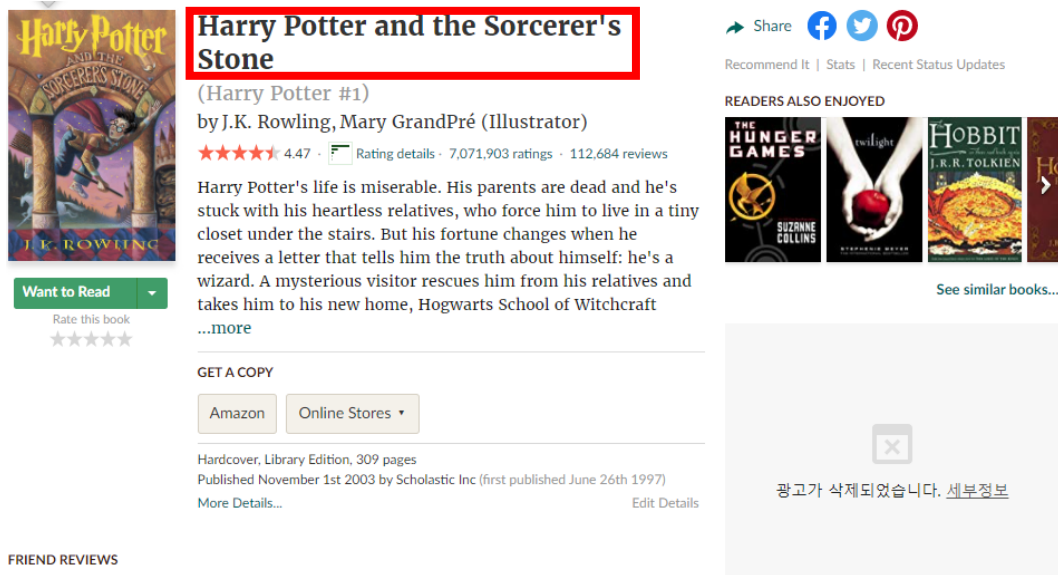
[Step 7] Step 6에서 수집된 URL의 개수를 이용하여 반복문을 작성하고, 이후 Step 6에서 수집된URL과 문제 2의 결과를 이용하여 서적 제목의 링크로 이동할 수 있는 URL을 변수에 할당하세요. 결과적으로 반복문을 수행하며 페이지 내의 모든 서적 제목에 대해 지정된 링크로 이동하는 URL을 변수에 할당할 수 있는 코드를 작성하세요.

#Hint: URL을 변수에 할당하기 위해 paste0 함수를 활용하세요!

#Hint: gsub('%2B', '+', tmp_url)를 한번 적용해 주세요!

[Step 8] 서적의 제목을 수집하는 코드를 작성하세요. 이 때, 서적의 제목에 두 글자 이상의 공백이 있다면 일괄적으로 한 칸의 공백으로 변경하세요. 이후 제목의 앞 뒤에 여백이 있다면 여백을 제거하세요.

#Hint: trim <- function (x) gsub("^\\W\\S+|\\W\\S+\$", "", x)



Harry Potter and the Sorcerer's Stone
(Harry Potter #1)
by J.K. Rowling, Mary GrandPré (Illustrator)
★★★★★ 4.47 · Rating details · 7,071,903 ratings · 112,684 reviews

Harry Potter's life is miserable. His parents are dead and he's stuck with his heartless relatives, who force him to live in a tiny closet under the stairs. But his fortune changes when he receives a letter that tells him the truth about himself: he's a wizard. A mysterious visitor rescues him from his relatives and takes him to his new home, Hogwarts School of Witchcraft ...more

GET A COPY
Amazon Online Stores ▾

Hardcover, Library Edition, 309 pages
Published November 1st 2003 by Scholastic Inc (first published June 26th 1997)
More Details... Edit Details

FRIEND REVIEWS

Share Recommend It | Stats | Recent Status Updates


READERS ALSO ENJOYED

See similar books...

광고가 삭제되었습니다. 세부정보

[Step 9] 서적의 저자를 수집하는 코드를 작성하세요. 이 때, 저자의 이름 앞 뒤에 여백 또는 쉼표가 있다면 제거하고 저자의 이름이 ' , '로 구분되도록 수집하세요. (아래 예시에서는 J.K. Rowling, Mary GrandPré (Illustrator)와 같이 수집)

#Hint: toString() 함수를 활용하세요!



Harry Potter and the Sorcerer's Stone
(Harry Potter #1)
by J.K. Rowling, Mary GrandPré (Illustrator)
★★★★★ 4.47 · Rating details · 7,071,903 ratings · 112,684 reviews

Harry Potter's life is miserable. His parents are dead and he's stuck with his heartless relatives, who force him to live in a tiny closet under the stairs. But his fortune changes when he receives a letter that tells him the truth about himself: he's a wizard. A mysterious visitor rescues him from his relatives and takes him to his new home, Hogwarts School of Witchcraft ...more

GET A COPY
Amazon Online Stores ▾

Hardcover, Library Edition, 309 pages
Published November 1st 2003 by Scholastic Inc (first published June 26th 1997)
More Details... Edit Details

FRIEND REVIEWS

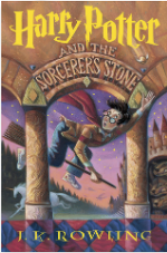
Share Recommend It | Stats | Recent Status Updates


READERS ALSO ENJOYED

See similar books...

광고가 삭제되었습니다. 세부정보

[Step 10] 서적의 평점을 수집하는 코드를 작성하세요. 평점에 공백이 존재할 경우 제거하여 수집하도록 코드를 작성하세요.



Harry Potter and the Sorcerer's Stone
(Harry Potter #1)
by J.K. Rowling, Mary GrandPré (Illustrator)
★★★★☆ 4.47  Rating details · 7,071,903 ratings · 112,684 reviews




Harry Potter's life is miserable. His parents are dead and he's stuck with his heartless relatives, who force him to live in a tiny closet under the stairs. But his fortune changes when he receives a letter that tells him the truth about himself: he's a wizard. A mysterious visitor rescues him from his relatives and takes him to his new home, Hogwarts School of Witchcraft ...more

Want to Read
Rate this book
★★★★☆

GET A COPY
Amazon Online Stores ▾


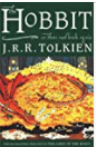


Hardcover, Library Edition, 309 pages
Published November 1st 2003 by Scholastic Inc (first published June 26th 1997)
More Details... Edit Details

FRIEND REVIEWS

Share   

Recommend It | Stats | Recent Status Updates


READERS ALSO ENJOYED



See similar books...

광고가 삭제되었습니다. 세부정보

[Step 11] 서적의 총 평점 개수를 수집하는 코드를 작성하세요. 평점의 개수는 정수의 형태로 수집되도록 코드를 작성하세요. (아래 예시에서는 7071903)



Harry Potter and the Sorcerer's Stone
(Harry Potter #1)
by J.K. Rowling, Mary GrandPré (Illustrator)
★★★★☆ 4.47  Rating details · 7,071,903 ratings · 112,684 reviews



Harry Potter's life is miserable. His parents are dead and he's stuck with his heartless relatives, who force him to live in a tiny closet under the stairs. But his fortune changes when he receives a letter that tells him the truth about himself: he's a wizard. A mysterious visitor rescues him from his relatives and takes him to his new home, Hogwarts School of Witchcraft ...more

Want to Read
Rate this book
★★★★☆

GET A COPY
Amazon Online Stores ▾


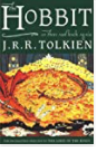


Hardcover, Library Edition, 309 pages
Published November 1st 2003 by Scholastic Inc (first published June 26th 1997)
More Details... Edit Details

FRIEND REVIEWS

Share   

Recommend It | Stats | Recent Status Updates

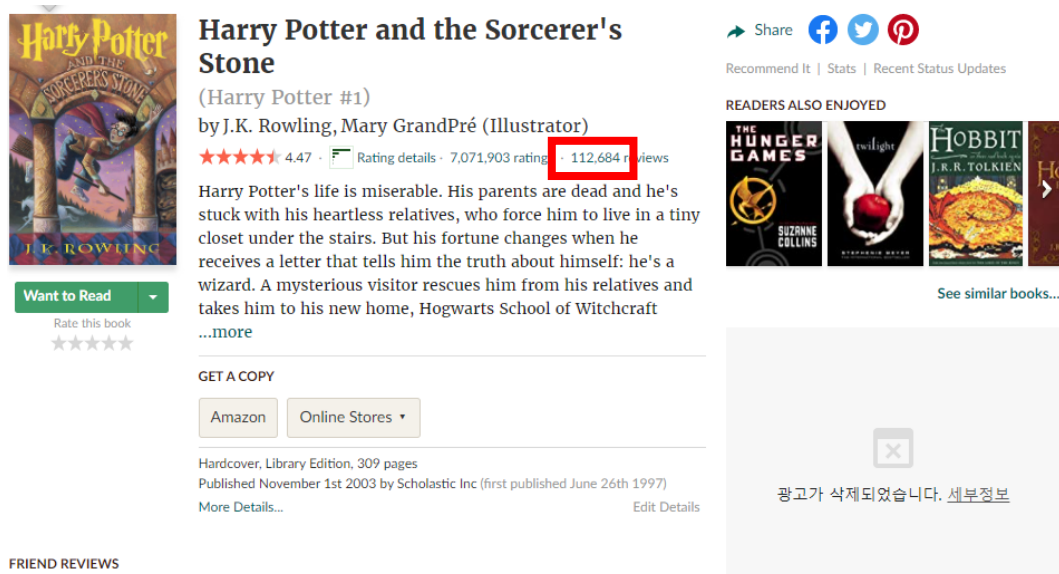
READERS ALSO ENJOYED



See similar books...

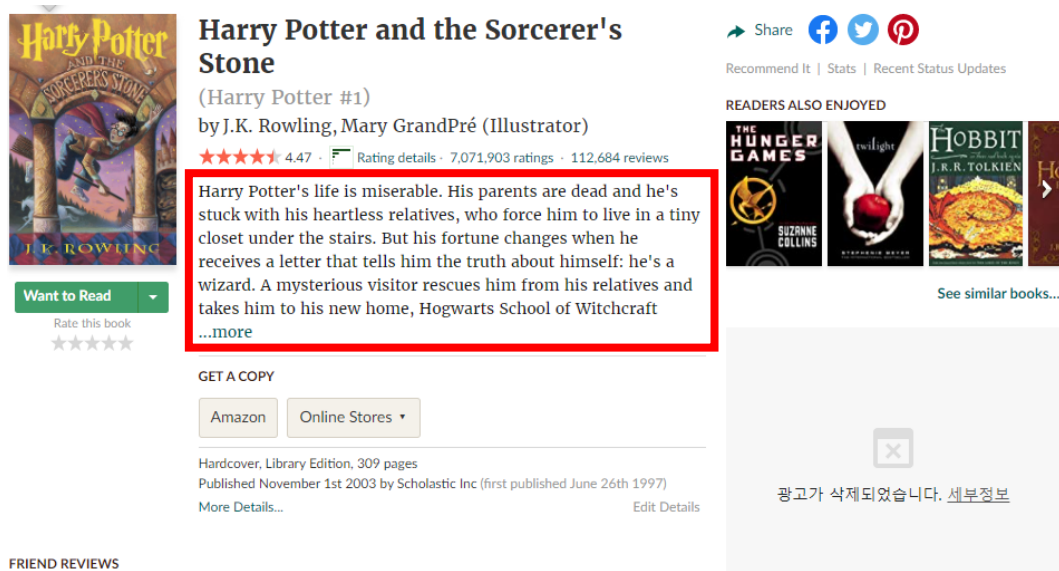
광고가 삭제되었습니다. 세부정보

[Step 12] 서적의 총 리뷰 개수를 수집하는 코드를 작성하세요. 리뷰의 개수는 정수의 형태로 수집되도록 코드를 작성하세요. (아래의 예시에서는 112684)



The screenshot shows the Amazon product page for 'Harry Potter and the Sorcerer's Stone'. The book cover is on the left. The title and author information are at the top. The rating is 4.47 stars with 112,684 reviews, where the number '112,684' is highlighted in a red box. The description follows, and there are buttons for 'Want to Read', 'Rate this book', and 'GET A COPY'. The 'READERS ALSO ENJOYED' section shows other book covers like 'The Hunger Games' and 'Hobbit'.

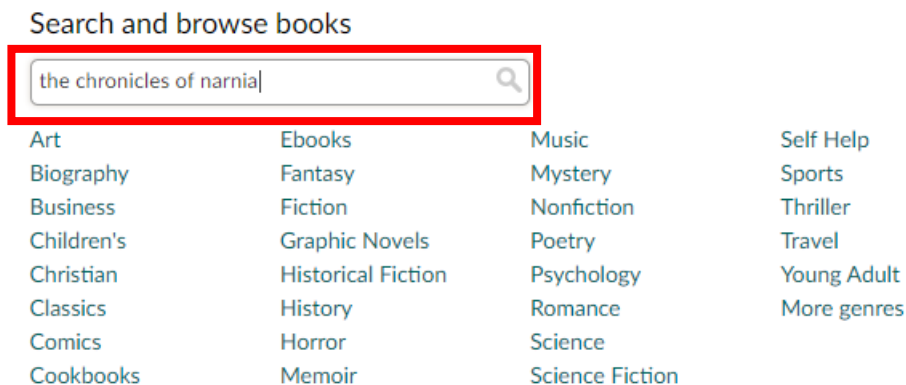
[Step 13] 서적의 요약문을 수집하는 코드를 작성하세요. 페이지에 보이는 부분만을 수집하는 코드를 작성하면 됩니다. (아래 예시에서는 “...more”까지) 두 칸 이상의 공백이 존재할 경우 일괄적으로 한 칸 공백으로 변경하고, 앞 뒤에 여백이 있을 경우 제거하세요. 만일 요약문이 존재하지 않으면 한 칸 공백 (“ ”)를 수집된 요약문으로 간주하도록 코드를 작성하세요.



This screenshot is identical to the one in Step 12, but the synopsis text is highlighted with a red box. The synopsis reads: 'Harry Potter's life is miserable. His parents are dead and he's stuck with his heartless relatives, who force him to live in a tiny closet under the stairs. But his fortune changes when he receives a letter that tells him the truth about himself: he's a wizard. A mysterious visitor rescues him from his relatives and takes him to his new home, Hogwarts School of Witchcraft ...more'.

[Step 14] Step 8 ~ Step 13의 결과를 이용하여 수집된 Data를 DataFrame 형태로 변환한 뒤 csv파일로 저장하세요. 이 때, 파일 이름은 “books.csv”로 설정하세요. 해당 파일은 본 과제의 제출물 중 하나입니다.

[Step 15] Step 14의 결과는 정해진 1개의 Keyword인 “harry potter”로 검색했을 때 등장하는 모든 서적에 대한 제목, 저자, 평점, 평점의 수, 리뷰의 수, 요약문입니다. 이후의 Step에서는 복수 개의 Keyword에 대해 등장하는 모든 서적을 수집하고자 합니다. Good Reads 웹 사이트(<https://www.goodreads.com/>)로 돌아가서 중앙에 검색 키워드를 “the chronicles of narnia”, “the lord of the rings”로 변화시켜 가며 URL의 패턴을 파악하세요.



문제 3: 검색 Keyword (harry potter, the chronicles of narnia, the lord of the rings)에 따라 URL이 어떻게 변화하는지 설명하세요.

[Step 16] 문제 3의 결과를 참고하여, modify_url 함수를 사용해 검색 키워드에 따라 페이지에 접근할 수 있는 URL을 만들어 보세요.

#Hint modify_url(url, query = list(q = search_keyword))

[Step 17] 검색 키워드 “harry potter”, “the chronicles of narnia”, “the lord of the rings”에 대해 수행되는 반복문을 작성하고, Step 16의 결과와 Step 14까지 작성한 코드를 바탕으로 자동적으로 3개 키워드에 대해 검색된 서적의 제목, 저자, 평점, 평점의 수, 리뷰의 수, 요약문을 수집하는 코드를 작성하세요. 수집된 결과가 검색 Keyword로 구분될 수 있도록 별도의 Column을 만들어 DataFrame을 만들고, 결과를 “multi_books.csv”로 저장하세요. 해당 파일은 본 과제의 제출물 중 하나입니다.

#Hint: 단순 반복문으로도 Step 17을 수행할 수 있지만, 필요하다면 함수를 사용하세요.

#Hint: books.csv와 multi_books.csv의 예시는 각각 다음과 같습니다. 임의로 몇 개의 Row를 제거했으므로 수집 내용은 고려하지 마세요.

title	authors	rating	num_of_ra	num_of_re	text
Harry Pott	J.K. Rowlin	4.5	2515728	42557	There is a
Harry Pott	J.K. Rowlin	4.47	7072129	112687	Harry Pott
Harry Pott	J.K. Rowlin	4.57	2816672	55662	Harry Pott
Harry Pott	J.K. Rowlin	4.43	2735355	53035	Ever since
Harry Pott	J.K. Rowlin	4.56	2602732	46711	Harry Pott
Harry Pott	J.K. Rowlin	4.57	2445101	39614	The war ag
Harry Pott	John Tiffar	3.62	704588	64951	Based on a
Harry Pott	J.K. Rowlin	4.62	99147	553	
Harry Pott	J.K. Rowlin	4.73	252437	7381	Over 4000
Harry Pott	J.K. Rowlin	4.62	2818422	65647	Harry Pott
Harry, a Hi	Melissa Ar	4.12	15007	709	THE HARR'

<Example of books.csv>

keyword	title	authors	rating	num_of_rating	num_of_reviews	text
harry potter	Harry Potter and the	J.K. Rowling, Mary GrandP	4.5	2515728	42557	There is a
harry potter	Harry Potter and the	J.K. Rowling, Mary GrandP	4.47	7072129	112687	Harry Pott
harry potter	Harry Potter and the	J.K. Rowling, Mary GrandP	4.57	2816672	55662	Harry Pott
harry potter	Harry Potter and the	J.K. Rowling, Mary GrandP	4.43	2735355	53035	Ever since
harry potter	Harry Potter and the	J.K. Rowling, Mary GrandP	4.56	2602732	46711	Harry Pott
the chronicles of	The Lion, the Witch	C.S. Lewis	4.22	2133080	21909	Narnia...the
the chronicles of	The Chronicles of N	C.S. Lewis, Pauline Baynes	4.26	519448	10392	Journeys to
the chronicles of	The Last Battle	C.S. Lewis	4.02	222760	6427	This edition
the chronicles of	The Silver Chair	C.S. Lewis, Pauline Baynes	3.95	239297	5358	Jill and Eus
the chronicles of	The Magician's Nep	C.S. Lewis	4.04	411967	13291	The secret
the lord of the rin	The Fellowship of th	J.R.R. Tolkien	4.36	2360694	22861	One Ring t
the lord of the rin	The Two Towers	J.R.R. Tolkien	4.44	722195	10784	Alternate C
the lord of the rin	The Return of the K	J.R.R. Tolkien	4.53	680771	10177	In the thir
the lord of the rin	The Lord of the Rin	J.R.R. Tolkien	4.5	567655	11615	One Ring t
the lord of the rin	J.R.R. Tolkien 4-Boo	J.R.R. Tolkien	4.6	110477	1785	This four-v

<Example of multi_books.csv>

출제된 문제는 3 문제이며, 제출 파일을 도출하기 위한 과정은 17 Step입니다. 작성한 R script를 포함하여 4개 파일을 제출하시기 바랍니다.

수고하셨습니다.