

CS 839 - Data Science Project - Stage 2

Team Members

- Aravind Soundararajan (soundararaj2@wisc.edu)
- Krishnan Rajagopalan (krajagopalan@wisc.edu)
- Palaniappan Nagarajan (pnagarajan3@wisc.edu)

Web Sources

- [Goodreads](#) - Goodreads is a "social cataloging" website that allows individuals to freely search its database of books, annotations, and reviews. Users can sign up and register books to generate library catalogs and reading lists. They can also create their own groups of book suggestions, surveys, polls, blogs, and discussions.
- [BookDepository](#) - Book Depository is a UK-based online book seller with a large catalogue of books.

Extraction of Structured Data

Scrapy, the open source web scraping tool was used for developing the wrapper-based extractor for both the sources.

1) Goodreads

Gone Girl Book Title
by **Gillian Flynn** (Goodreads Author) Book Author

★★★★★ 4.04 · Rating details · 1,740,701 Ratings · 123,734 Reviews

On a warm summer morning in North Carthage, Missouri, it is Nick and Amy Dunne's fifth wedding anniversary. Presents are being wrapped and reservations are being made when Nick's clever and beautiful wife disappears. Husband-of-the-Year Nick isn't doing himself any favors with cringe-worthy daydreams about the slope and shape of his wife's head, but passages from Amy's dia ...more

goodreads choice 2012 WINNER

Want to Read

Rate this book

★★★★★

Preview Listen

Read Excerpt

GET A COPY

Kindle Store \$9.99 Amazon IN Stores Links

Paperback 415 pages

Published April 22nd 2014 by Broadway Books First published May 24th 2012

Original Title Gone Girl

ISBN 0307588378 (ISBN13: 9780307588371)

Edition Language English Language

Characters Nick Dunne, Amy Elliot Dunne, Margo "Go" Dunne, Rand Elliot, Marybeth Elliot...more

setting North Carthage, Missouri (United States) Missouri (United States)

Literary Awards Barry Award Nominee for Best Novel (2013), Anthony Award Nominee for Best Novel (2013), Romantic Times Reviewers' Choice Award (RT Award) for Suspense/Thriller Novel (2012), Shirley Jackson Award Nominee for Novel (Finalist) (2012), Edgar Award Nominee for Best Novel (2013) ...more

Other Editions (7)

...Less Detail

edit details

Share

Recommend It | Stats | Recent Status Updates

READERS ALSO ENJOYED

GENRES Genres

Mystery	8,638 users
Fiction	2,261 users
Mystery > Crime	2,239 users
Contemporary	1,702 users
Thriller	1,637 users

See top shelves...

ABOUT GILLIAN FLYNN

Gillian Flynn is an American author and television critic for Entertainment Weekly. She has so far written three novels, *Sharp Objects*, for which she won the 2007 Ian Fleming Steel Dagger for the best thriller; *Dark Places*; and her best-selling third novel *Gone Girl*.

The start_urls were generated to be all pages under “All Time Favorite Romance Novels” tag (genre). And xpath was used to filter out the individual links of books in the listing page. The CSS identifier of individual book links were provided to Scrapy to crawl through the various books in the results. Each of the book links were visited to crawl the details page of the corresponding book. All the information available about the book such as Author, Genres (that people have categorised it into), Publishers, Publishing date, number of pages were pulled out by using xpath identifiers for each of the attribute in the page. We have used the “Item Pipeline” by defining two pipelines. One component (BookWormsPipeline) in the pipeline uses the process_item method which performs basic cleansing of HTML data and the other component (FileWritePipeline) writes the cleaned data to CSV.

2) BookDepository

Categories: Contemporary Fiction | Crime | Crime Fiction | Romance | Romance Books | People & Places
Genres

45% off

Gone Girl

Book Title

★★★★★ 4.04 (1,718,134 ratings by Goodreads)

Paperback | Phoenix | English

By (author) Gillian Flynn

Book Author(s)

Share

Also available in
CD-Audio US\$12.52

'What are you thinking, Amy? The question I've asked most often during our marriage, if not out loud, if not to the person who could answer. I suppose these questions stormcloud over every marriage: What are you thinking? How are you feeling? Who are you? What have we done to each other? What will we do? Just how well can you ever know the person you love? This is the question that Nick Dunne must ask himself on the morning of his fifth wedding anniversary when his wife Amy suddenly disappears. The police immediately suspect Nick. Amy's friends reveal that she was afraid of him, that she kept secrets from him. He swears it isn't true. A police examination of his computer shows strange searches. He says they aren't his. And then there are the persistent calls on his mobile phone. So what really did happen to Nick's beautiful wife? And what was in that half-wrapped box left so casually on their marital bed? In this novel, marriage truly is the art of war

...

Product details

Number of Pages

Format: Paperback 480 pages

Language

Language: English

Publishing Date

Publication date: 08 Nov 2012

Publisher

Publisher: Orion Publishing Co

ISBN10

ISBN10: 1780221355

ISBN13

ISBN13: 9781780221359

Bestsellers rank

Bestsellers rank: 94

Imprint: Phoenix (an Imprint of The Orion Publishing Group Ltd

The start_urls were generated to be all pages under “Romance”, “Historical Romance”, “Adult-Contemporary-Romance”, “Erotic-Fiction” genres. Individual book links were identified were collected using xpath. HTTP requests to these links were triggered to collect details from within the individual book page using xpath. Also the components of the item pipeline are

common to both the spiders. The approach is quite similar to the extraction of book details from Goodreads.

Entity Type - Books

We have extracted entities of type “Book” from the two sources. Specifically, we have extracted information about books broadly in the “Romance” category. Since both the sources contained similar information about the entity (books), a common schema was decided upon before the extraction and we extracted the corresponding attributes from the two web sources.

Schema of the tables

Attribute Name	Datatype	Description
Title	String	Title of the book (including the Edition number) as given in the source from which the book details are taken.
Authors	String	Names of all the persons listed under the author’s section along with the authors. This includes the translator, the illustrator etc.
Genres	String	The category of books under which the book can be classified such as Horror, Comedy etc. A single book can be classified under multiple genres.
Publishing Date	String	Date of publishing.
Pages	Integer	Number of pages in the book.
Publisher	String	Name of the publisher(s) as listed in the source.
Language	String	Language.
ISBN	Integer	Unique identifier information (International Standard Book Number).
ISBN13	Integer	Unique identifier information (International Standard Book Number).
URL	String	URL of the HTML page from where the data was scraped.

Information about the tables

CSV Filename	Web Source	Number of tuples
bookdepository.csv	https://www.bookdepository.com	3968
goodreads.csv	https://goodreads.com	3794

Open Source tools used.

- [Scrapy](#) - Scrapy is a web crawling framework, written in Python. It is a general purpose web crawler and can also be used for web scraping and extracting data using APIs. Using Scrapy, we define spiders for each web source. And Scrapy provides the framework for processing the response for each request fired. Also, support for crawling the multiple links in the HTML is provided by Scrapy.
- The [lxml XML toolkit](#) (used internally by Scrapy) is a Pythonic binding for the C libraries libxml2 and libxslt. It is unique in that it combines the speed and XML feature completeness of these libraries with the simplicity of a native Python API, mostly compatible but superior to the well-known ElementTree API