Домашнее задание №2

Тема: «Краулинг книг с сайта chitai-gorod.ru».

Сроки: задание открывается 27.01.25. Дедлайн сдачи — 02.02.25.

Задание: компания N планирует организовать внутри отдела развития компетенций сотрудников систему автоматической закупки книг. Большинство сотрудников компании обожает читать книги в бумажном варианте, поэтому внутри здания была построена небольшая библиотека. Но появилась проблема — компании негде брать информацию о книгах.

На стороне отдела закупки был подготовлен API, который отправляет заявки на бронирование книг в магазинах, но, чтобы не использовать абстрактные «айдишники», было принято решение описывать книги в виде ISBN. Также был подготовлен фронтенд, через который сотрудники компании могут выбрать интересующие их книги. Схема данных для каждой книги выглядит так:

Название поля	**Описание**	**Обязательно для заполнения**
`title`	Название книги	Да
`author`	Автор	Нет
`description`	Описание	Нет
`price_amount`	Цена	Нет
`price_currency`	Валюта	Нет
`rating_value`	Средняя цена	Нет
`rating_count`	Количество оценок	Нет
`publication_year`	Год публикации	Да
`isbn`	ISBN	Да
`pages_cnt`	Количество страниц	Да
`publisher`	Издательство	Нет
`book_cover`	Обложка книги (ссылка на картинку)	Нет
`source_url`	Ссылка на источник данных	Да

Требований к скорости сбора нет — главное, уважать владельцев ресурса и не создать им зловредной нагрузки. По возможности следуй robots.txt и не превышай указанную там максимально возможную скорость сбора. Все посты должны писаться отдельным пайплайном в NoSQL СУБД MongoDB. Предполагается, что при минимальной конфигурации БД можно будет подключить к сервису FastAPI и получать книги по запросу ISBN-кодом.

Искать книги можно в следующих местах:

- в жанрах;
- авторах;
- издательствах;

- серии;
- Sitemap.

Про SitemapSpider можно узнать здесь: https://docs.scrapy.org/en/latest/topics/spiders.html#sitemapspider. Проверить данные можно через сервис FastAPI, находящийся в файле `fastapi_service_books.py`. Инструкция лежит в файле «`Как запустить сервис FastAPI.pdf`».

Система оценивания: десятибалльная.
Критерии оценивания: реализовать паука на Scrapy, собрать сэмпл данных, описать схему данных через scrapy.item, описать собираемые поля в XPath-выражениях, использовать Sitemap в реализации обхода/сбора источника, реализовать Pipeline для записи данных во внешний источник данных (MongoDB).

Формат сдачи: код проекта запушить в личный репозиторий GitHub, приложив ссылку на него в окне ниже. В корне проекта оставь сэмпл данных в формате .csv/.jsonlines, в объёме ±1000 объектов, а также пример запроса и ответа приложенного сервиса FastAPI.