**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**ІМЕНІ ТАРАСА ШЕВЧЕНКА**

Факультет комп’ютерних наук та кібернетики

Кафедра теорії та технології програмування

**Звіт**

до лабораторної роботи № 1

Виконав студент 4-го курсу

Групи ТТП-42

Ходаков Максим Олегович

Київ – 2025

**Короткий опис області**

Даний дата сет візуалізує список книжок наявних на книжковій платформі, разом з рейтингом та кількістю рецензій. Також для кожної книжки наявні автор, мова написання, дата випуску та видавництва.

**DataSet**

Посилання: <https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m>

Приблизно 10 000 000 книг доступні в архівах сайту, і ці набори даних збираються з них. для запитів через API ми використовували бібліотеку python Goodreads, набори даних оновлюватимуться кожні 2 дні.

Ці набори даних дуже добре підходять для виконання двох завдань:

1 . Створення системи рекомендацій книг на основі 10 млн книг 🥇

2 . Використання колонок опису для НЛП 🥈

Репозиторій на Github

Посилання на проект на github: <https://github.com/BahramJannesar/GoodreadsBookDataset>

Зміст

Приблизно 10 000 000 книг доступні в архівах сайту, і ці набори даних збираються з них. для запитів через API ми використовували бібліотеку python Goodreads, набори даних оновлюватимуться кожні 2 дні.

Приклади даних books в JSON:

{

"Id": "5107",

"Name": "The Catcher in the Rye",

"RatingDist1": "1:133165",

"RatingDist2": "2:224884",

"RatingDist3": "3:553476",

"RatingDist4": "4:808278",

"RatingDist5": "5:891037",

"pagesNumber": 277,

"RatingDistTotal": "total:2610840",

"PublishMonth": 30,

"PublishDay": 1,

"Publisher": "Back Bay Books",

"CountsOfReview": 44046,

"PublishYear": 2001,

"Language": "eng",

"Authors": "J.D. Salinger",

"Rating": 3.8,

"ISBN": "0316769177",

"Count of text reviews": 55539,

"Description": "The hero-narrator of The Catcher in the Rye is an ancient child of sixteen, a native New Yorker named Holden Caulfield. Through

circumstances that tend to preclude adult, secondhand description, he leaves his prep school in Pennsylvania and goes underground in New York City for

three days. "

}

Кількість файлів при завантажені архіву:

A screenshot of a computer

Description automatically generated

Приклади завантажених .CSV

Books:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **RatingDist1** | **pagesNumber** | **RatingDist4** | **RatingDistTotal** | **PublishMonth** | **PublishDay** | **Publisher** | **CountsOfReview** | **PublishYear** | **Language** | **Authors** | **Rating** | **RatingDist2** | **RatingDist5** | **ISBN** | **RatingDist3** |
| Harry Potter and the Half-Blood Prince (Harry Potter, #6) | 1:9896 | 652 | 4:556485 | total:2298124 | 16 | 9 | Scholastic Inc. | 28062 | 2006 | eng | J.K. Rowling | 4.57 | 2:25317 | 5:1546466 |  | 3:159960 |

UserRating:

A screen shot of a computer screen

Description automatically generated

Основні поля для роботи з книгами:

Name – назва книги

RatingDist1 – кількість оцінок 1 для книги

RatingDist2 – кількість оцінок 2 для книги

RatingDist3 – кількість оцінок 3 для книги

RatingDist4 – кількість оцінок 4 для книги

RatingDist5 – кількість оцінок 5 для книги

PublishYear – рік випуску

PublishMonth – місяць випуску

PublishDay – день випуску

CountsOfReview – кількість рецензій

Publisher – видавець

Language – мова книги

Authors – автори

**STAR схема для даного OLAP сховища**

A diagram of a data flow

Description automatically generated

Рис.1 – STAR схема

**ETL опис**

**Extract**: Спочатку всі дані було витягнуто з файлів за посиланням наведеним вище.

**Transform**:

Дані були об’єднанні з різних файлів в один а поті вже робилися подальші опрацювання. Дані в таблиці були трохи змінені, для відповідності правильним даним ( наприклад, день та місяць переплутані місцями або рейтинги перенести в числовий формат або прибрати символи “” зайві в іменах. Також необхідно було видалити не коректні дані (там де рік не коректний і тд.)

Таблицю розділено на декілька dimensions: дата, автор, видавництво, мова та таблицю книжок з measurements: рейтинг, кількість рецензій, кількість оцінок.

**Load**: Отримані та трансформовані дані можуть бути завантаженими до сховища даних (data warehouse) та попередньо ще трансформовані за бажанням та особливостями архітектури системи на яку завантажують.

**Pivot таблиці**

**Основна Таблиця**

# (Тут не всі колонки на скріншоті)

A screenshot of a spreadsheet

Description automatically generated Рис. 2 – Основна таблиця

A screenshot of a computer

Description automatically generated

Рис.3 – Pivot таблиця за середній рейтингом назвою та роком видання

A white sheet of paper with lines

Description automatically generated

Рис.3 – Pivot таблиця за максимальним рейтингом за датою та мовою англійською

A screenshot of a computer

Description automatically generated

Рис. 4 – Pivot таблиця з сумами кількостями рев’ю по роках з видавництвами та назвами безпосередньо книгами

**Висновок**

У ході виконання лабораторної роботи було проведено комплексне дослідження доступних джерел даних на платформі **Kaggle**, проаналізовано різні варіанти наборів даних, та обраний найбільш підходящий дата-сет для подальшого аналізу. Основна увага була приділена вибору даних, що містять інформацію про книги, їх авторів, видавців, кількість рецензій, рейтинги та дати публікації.

Після вибору набору даних було здійснено його **передобробку та очищення**, включаючи:

* Видалення дублікатів.
* Усунення відсутніх значень.
* Приведення типів даних до відповідного формату (цілі числа, рядки, десяткові числа).
* Уніфікацію значень у колонках, зокрема, заміна ком на крапки у числових значеннях.

На основі підготовлених даних була **розроблена і спроектована зіркова схема (Star Schema)**, яка є оптимальним рішенням для аналітичних запитів. Вона складається з факт-таблиці та декількох вимірних таблиць:

* **Факт-таблиця (DimBooks)**, яка містить ключові показники продуктивності, такі як рейтинг, кількість рецензій, розподіл оцінок тощо.
* **Вимірні таблиці**, що містять додаткову інформацію про видавців, авторів, мови та дати публікацій:
  + DimPublishers – з інформацією про видавців.
  + DimAuthors – з інформацією про авторів.
  + DimLanguages – з інформацією про мови книг.
  + DimPublishDates – для аналізу даних за часом.

Завдяки реалізації такої структури, була забезпечена можливість ефективного аналітичного аналізу та візуалізації даних у Pivot таблицях Microsoft Excel.

Загалом, виконана робота дозволила отримати цінний досвід у розробці аналітичних систем, обробці даних та побудові системи для бізнес-аналізу. В той же час робота показала що вибір джерела даних є важливим компонентом в бізнес аналітичних системах.