

Министерство науки и образования РФ
Федеральное государственное бюджетное учреждение
высшего образования
“Тверской Государственный Технический Университет”
(ТвГТУ)

Кафедра Программного обеспечения

Отчет по лабораторной работе №1
По дисциплине: «Анализ больших данных»
Тема: “Реляционные данные. Исследовательский анализ данных.
Построение визуализаций данных OLAP”

Выполнил:
студент группы
Б.ПИН.РИС-21.06
Миронов М.В.

Проверила:
старший преподаватель
кафедры ПО
Корнеева Е.И.

Тверь 2025 г.

Содержание

| | |
|--------------------------------------|---|
| 1. Задача | 1 |
| 2. Вариант задачи | 1 |
| 3. Ссылка на код | 1 |
| 4. Описание проделанной работы | 2 |
| 4.1. База данных | 2 |
| 4.2. Описание данных | 3 |
| 4.2.1. Таблицы | 3 |
| 4.2.2. Признаки | 3 |
| 4.3. Одномерный анализ | 3 |
| 4.3.1. views | 3 |
| 4.3.2. Виртуальные счетчики | 4 |
| 4.4. Многомерный анализ | 5 |
| 4.5. Средства | 6 |
| 5. Вывод | 6 |
| Список Литературы | 8 |

1. Задача

1. Произвести подключение к базе данных из python.
2. Описать данные, таблицы, признаки.
3. Провести одномерный анализ количественных признаков.
4. Многомерный анализ

Сложность “Well-done”

2. Вариант задачи

Kensho Derived Wikimedia Dataset [1]

3. Ссылка на код

Дамп таблиц не будет предоставлен, так как БД содержит слишком много данных для системы elearning. Взамен будет предоставлен DDL.

<https://github.com/NydusBorn/big-data>

4. Описание проделанной работы

4.1. База данных

Был произведен импорт данных из csv в БД postgres при помощи средств DataGrip. После создания связей между таблицами была получена следующая схема таблиц.

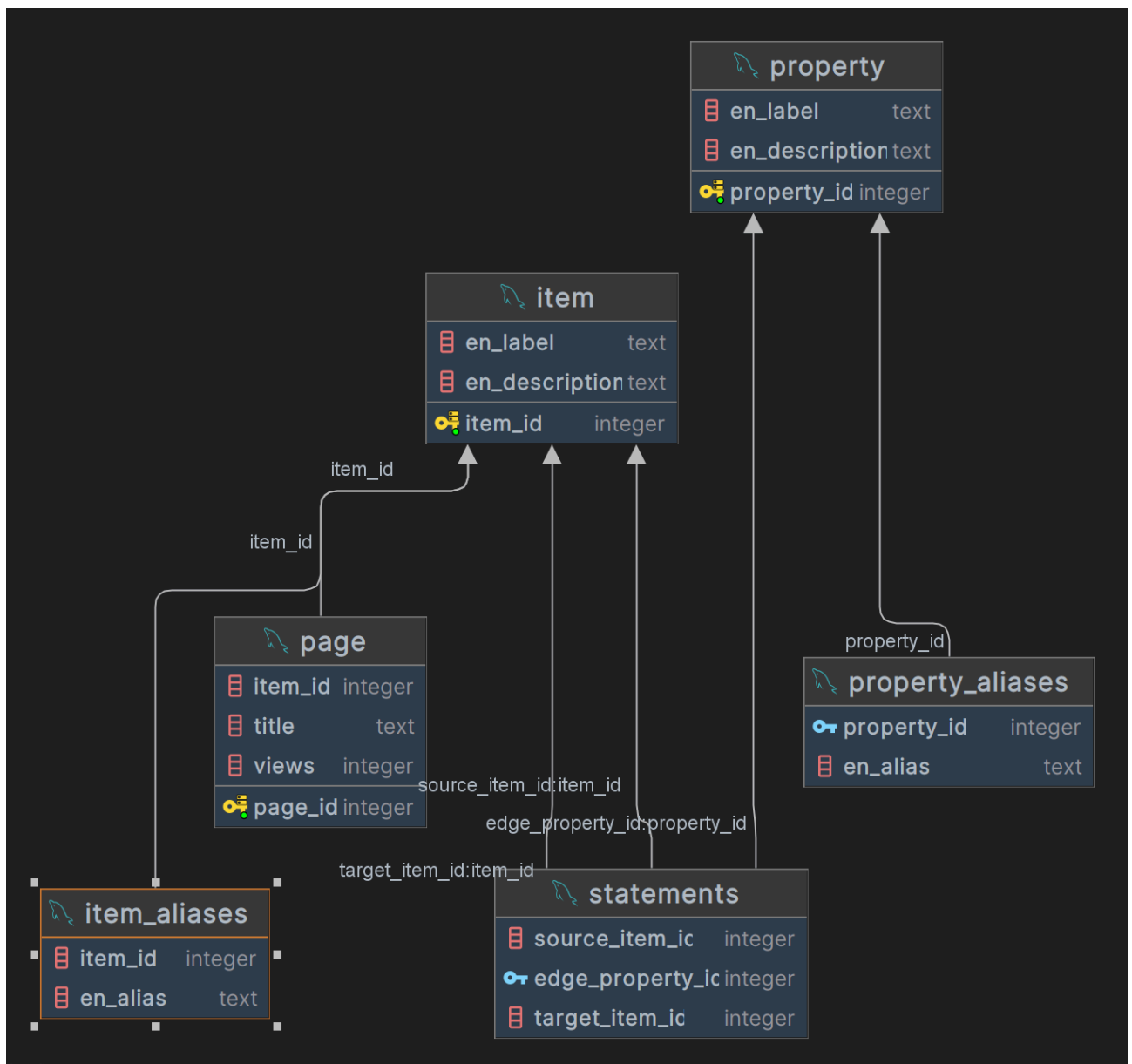


Figure 1: Схема связей в базе данных

Для подключения к базе данных были сгенерированы маппинги для репозитория при помощи встроенного в него инструмента `rwiz`.

4.2. Описание данных

4.2.1. Таблицы

В датасете 7 таблиц, но таблица “link_annotated_text” включена не была, так как не содержит анализируемых данных, и при том содержит 18 гб данных.

Основные данные в таблицах - связи и текст. Числовое поле только одно.

4.2.2. Признаки

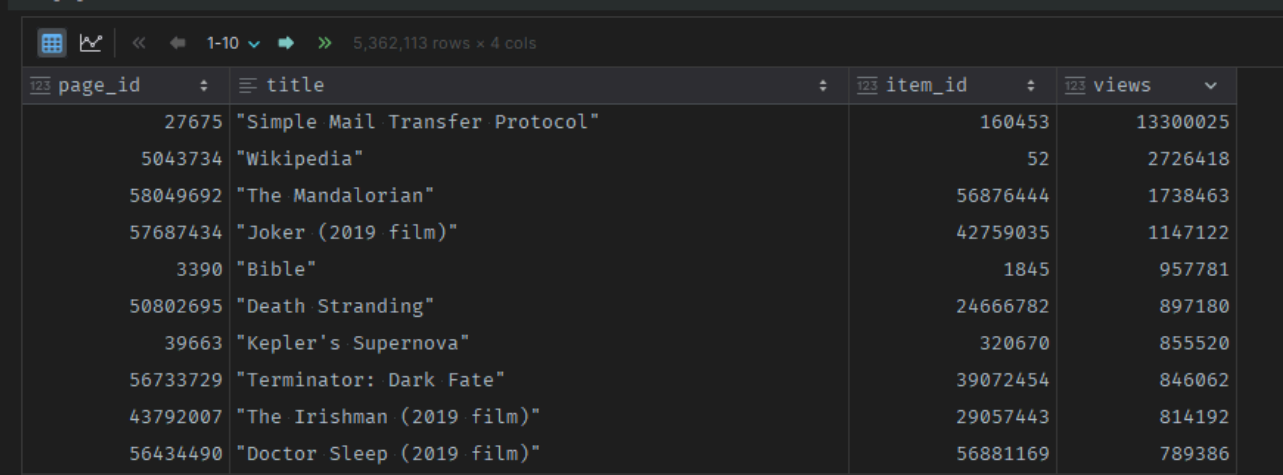
Количественные: page.views

Также можно выделить несколько виртуальных количественных признаков:
COUNT(item_aliases.item_id == item.item_id), COUNT(statements.source_item_id == item.item_id), COUNT(statements.target_item_id == item.item_id), COUNT(statements.edge_property_id == property.property_id)

Номинальные: property

4.3. Одномерный анализ

4.3.1. views



| page_id | title | item_id | views |
|----------|---------------------------------|----------|----------|
| 27675 | "Simple Mail Transfer Protocol" | 160453 | 13300025 |
| 5043734 | "Wikipedia" | 52 | 2726418 |
| 58049692 | "The Mandalorian" | 56876444 | 1738463 |
| 57687434 | "Joker (2019 film)" | 42759035 | 1147122 |
| 3390 | "Bible" | 1845 | 957781 |
| 50802695 | "Death Stranding" | 24666782 | 897180 |
| 39663 | "Kepler's Supernova" | 320670 | 855520 |
| 56733729 | "Terminator: Dark Fate" | 39072454 | 846062 |
| 43792007 | "The Irishman (2019 film)" | 29057443 | 814192 |
| 56434490 | "Doctor Sleep (2019 film)" | 56881169 | 789386 |

Figure 2: Топ 10 страниц по просмотрам

Как мы видим между страницами большой разрыв. Поэтому график будет представлен для страниц с количеством просмотров < 200 .

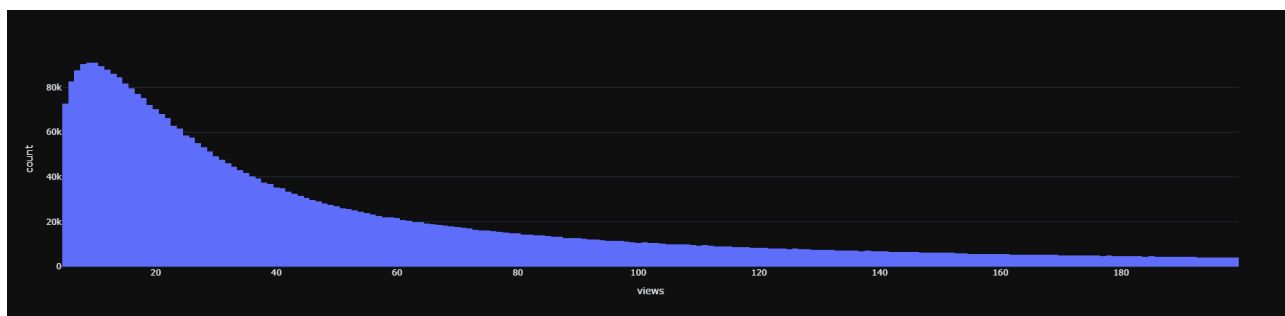


Figure 3: Страницы с менее чем 200 просмотров

Как можно увидеть значительная часть страниц имеет меньше 50 просмотров.

4.3.2. Виртуальные счетчики

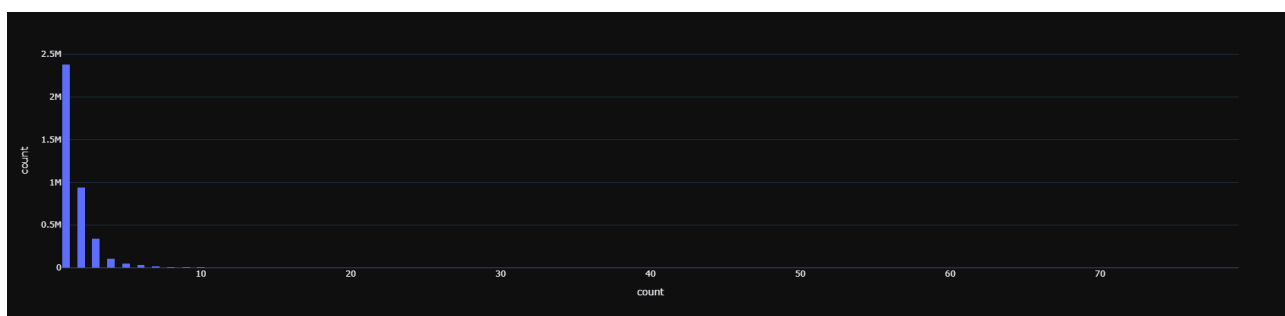


Figure 4: item_aliases

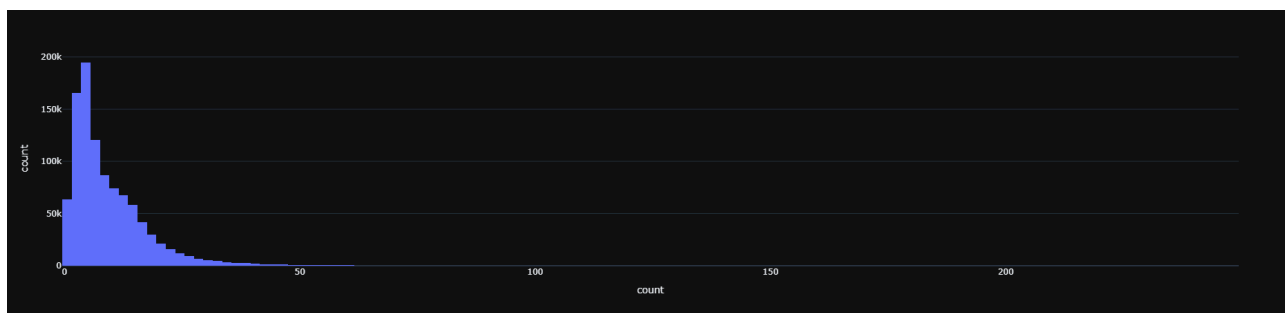


Figure 5: source_items

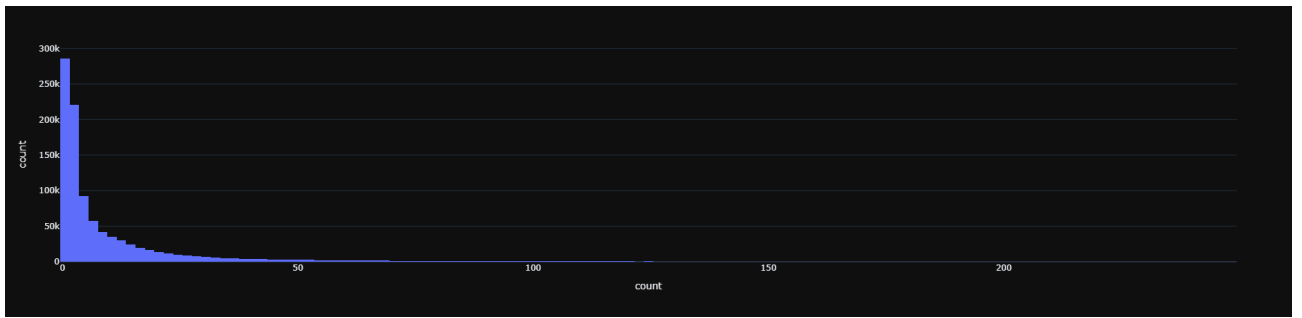


Figure 6: target_items

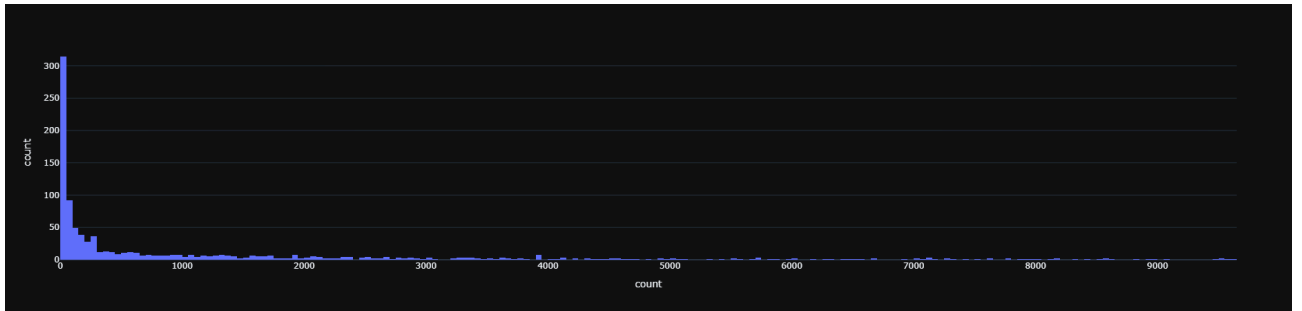


Figure 7: property_uses

Во всех графиках были выбросы, поэтому при формировании графиков были отфильтрованы экстремальные значения. Все графики кроме property_uses(Figure 7) приближены к нормальному распределению (с значительным отклонением в меньшую сторону). property_uses имеет распределение близкое к равномерному

4.4. Многомерный анализ

Так как есть только 2 количественные характеристики (которые можно как то интерпретировать), то будет 1 график - views x item_alias_counts.

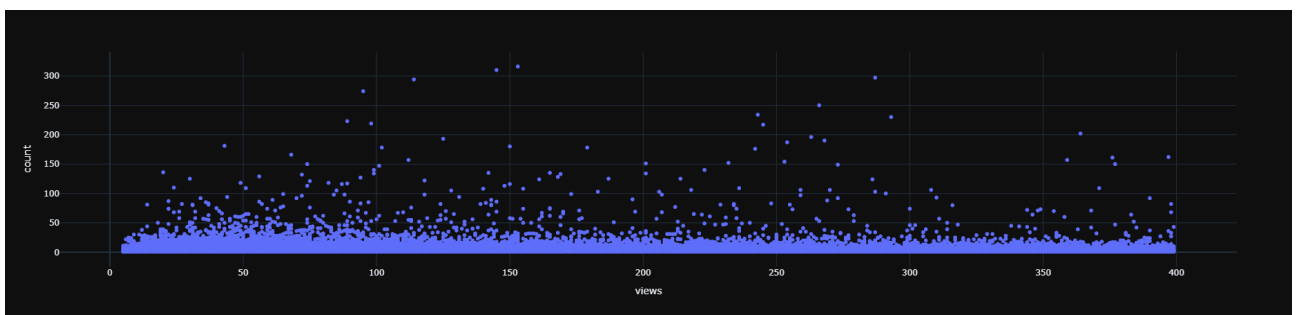


Figure 8: x - views, y - alias count

Как мы видим, зависимостей нет.

4.5. Средства

Для выполнения задачи использовались в том числе средства datagrip и ru-charm, позволяющие проводить анализ напрямую на таблицах sql и датафреймах из python.

Зачем что то делать когда можно не делать? Поэтому для задачи визуализации предлагается использовать metabase. (можно развернуть следующей командой: `docker run -d -p 32019:3000 metabase/metabase`)



Figure 9: Интерфейс приложения metabase с одним из графиков

5. Вывод

Была проведена работа по анализу датасета [1], в результате было изучено несколько новых инструментов. Из датасета не удалось получить какие либо очевидные знания, поэтому стоит предположить что:

1. Просмотры в значительной мере зависят от контента страницы, в частности популярности темы (датасет был создан ~2020 году, и 3 из 10 в топ 10 - фильмы вышедшие в 2019).
2. Влияние остальных входных незначительно по сравнению с 1.

Список Литературы

- [1] kaggle, “Kensho Derived Wikimedia Dataset.” Accessed: Jan. 19, 2025.
[Online]. Available: <https://www.kaggle.com/datasets/kenshoresearch/kensho-derived-wikimedia-data>