

Сбор, обработка и анализ данных с сайта «Кинопоиск»

В данном проекте реализуется сбор данных по фильмам с сайта «Кинопоиск» (<https://www.kinopoisk.ru>), размещение собранной информации в базах данных **MongoDB** (хранение), предобработка данных и их последующий анализ с помощью **Python** и **Tableau**. Рассмотрим данный проект по шагам.

Шаг №1. Сбор данных.

Сбор данных реализуем с помощью скриптов, написанных на Python, с использованием библиотек: **requests** (отправка GET запросов), **BeautifulSoup** (извлечение данных из файлов html и xml), **pymongo** (для взаимодействия с базами данных MongoDB).

Сбор данных осуществляется по следующей схеме:

- 1) Производится запрос к странице с фильмами с различными параметрами (год, номер страницы) на сайте «Кинопоиск» (на каждой странице располагаются 100 фильмов)
- 2) Далее из полученного ответа извлекаются данные по каждому фильму
- 3) Данные по каждому фильму записываются в словарь, из которых формируется список словарей за каждый год
- 4) Собираются данные по каждому году с 2000 по 2020
- 5) Данные отправляются в базу данных MongoDB для хранения

Скрипт данного шага находится в файле «Сбор данных Кинопоиск.ру»

Шаг №2. Хранение данных.

Хранение данных осуществляем с помощью **MongoDB**. Почему MongoDB? MongoDB – документоориентированная СУБД - не требует описания системы таблиц, записи хранятся в виде отдельных документов, поэтому мы можем легко подстраивать наши скрипты по сбору данных под изменения интернет-ресурсов, не боясь, что структура извлекаемых данных нарушится и мы не сможем записать эту информацию в базу данных.

Для каждого года создаем собственную базу данных. С помощью «**MongoDB Compass**» мы можем просмотреть полученные базы данных (**Рис. №1, Рис. №2**)

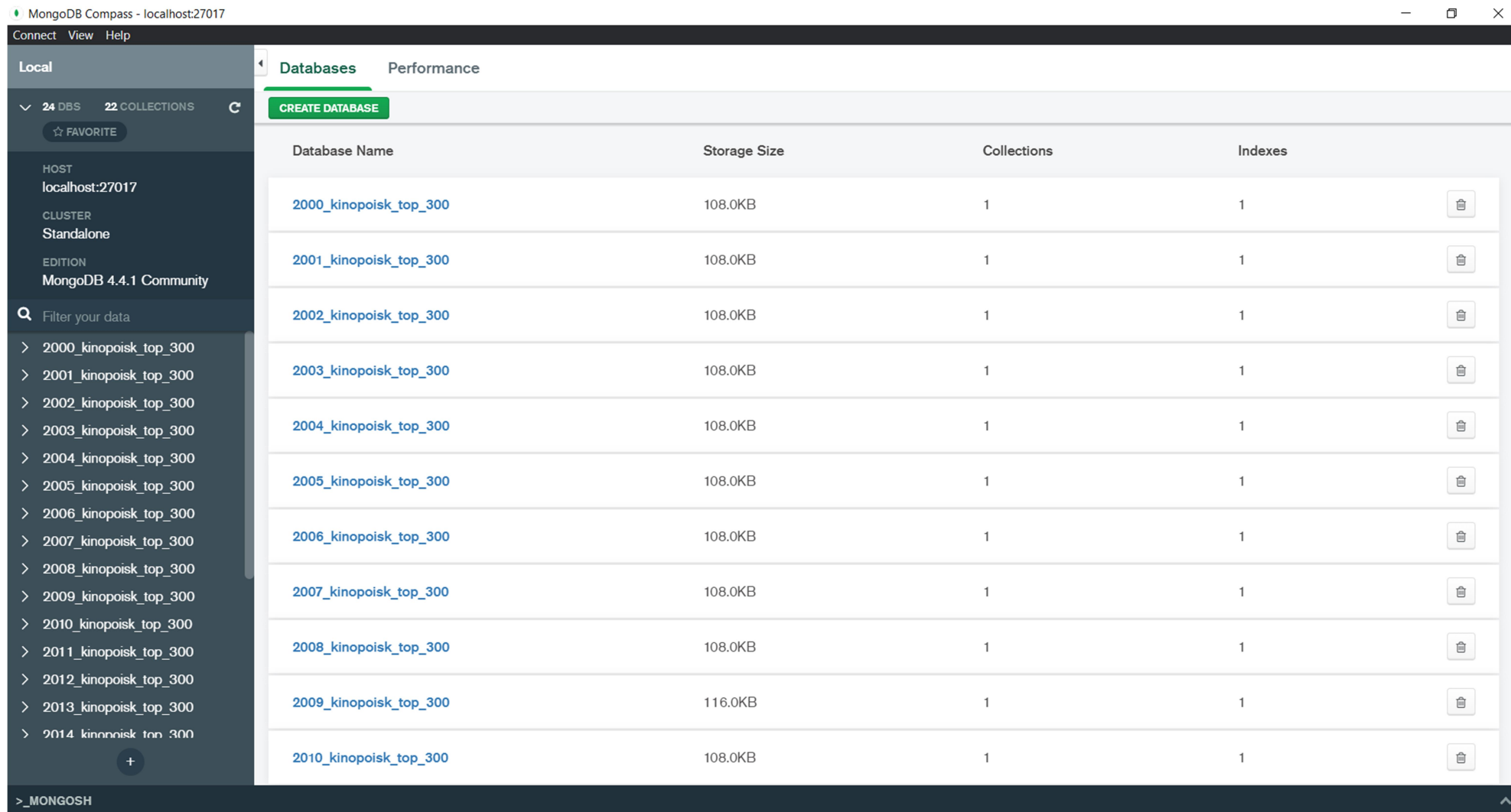


Рис. №1

MongoDB Compass - localhost:27017/2000_kinopoisk_top_300.films

Connect View Collection Help

Local

24 DBS 22 COLLECTIONS

☆ FAVORITE

HOST
localhost:27017

CLUSTER
Standalone

EDITION
MongoDB 4.4.1 Community

Filter your data

2000_kinopoisk_top_300

films

2001_kinopoisk_top_300

2002_kinopoisk_top_300

2003_kinopoisk_top_300

2004_kinopoisk_top_300

2005_kinopoisk_top_300

2006_kinopoisk_top_300

2007_kinopoisk_top_300

2008_kinopoisk_top_300

2009_kinopoisk_top_300

2010_kinopoisk_top_300

2011_kinopoisk_top_300

2012_kinopoisk_top_300

2000_kinopoisk_top_300 Documents

2000_kinopoisk_top_300.films

DOCUMENTS 300 TOTAL SIZE 110.7KB AVG. SIZE 378B INDEXES 1 TOTAL SIZE 20.0KB AVG. SIZE 20.0KB

Documents Aggregations Schema Explain Plan Indexes Validation

FILTER { field: 'value' }

ADD DATA VIEW

Displaying documents 1 - 20 of 300 REFRESH

	_id Int32	film_name String	film_time String	country String	main_roles String	genre String
1	1	"Российская империя (сериал)"	"60 мин"	"Россия, "	" Леонид Парфенов, Елизавета Ли	"документальный, и
2	2	"Metallica: S&M (ТВ)"	"145 мин"	"США, "	" Metallica, Джеймс Хетфилд, ..	"документальный, м
3	3	"Гладиатор"	"155 мин"	"США... "	" Рассел Кроу, Хоакин Феникс, .	"история, боевик, "
4	4	"Большой куш"	"104 мин"	"Великобритания... "	" Джейсон Стэйтем, Стивен Грэм,	"криминал, комедия
5	5	"Близкие друзья (сериал)"	"45 мин"	"Канада... "	" Гейл Харольд, Хэл Спаркс, ...	"драма, мелодрама)
6	6	"Первый шаг (сериал)"	"23 мин"	"Япония, "	" Стив Стейли, Ричард Эпкар, ..	"аниме, мультфильм
7	7	"Eminem: Stan (видео)"	"8 мин"	"США, "	" Эминем, Дайдо, ... "	"короткометражка, "
8	8	"Изгой"	"143 мин"	"США, "	" Том Хэнкс, Хелен Хант, ... "	"драма, мелодрама,
9	9	"Хроники Рэдволла: Маттimeo (се	"30 мин"	"Германия... "	" Энтони Бэкэнн, Вейн Бест, ...	"мультфильм, фэнте
10	10	"Книжный магазин Блэка (сериал)"	"25 мин"	"Великобритания, "	" Дилан Моран, Билл Бэйли, ...	"комедия) "
11	11	"О птичках"	"3 мин"	"США, "	" Ральф Эгглстон, ... "	"мультфильм, корот
12	12	"Влюблённые"	"216 мин"	"Индия, "	" Амитабх Баччан, Шах Рукх Кхан	"мюзикл, драма, ме
13	13	"Бандитский Петербург 2: Адвока	"366 мин"	"Россия, "	" Дмитрий Певцов, Ольга Дроздов	"драма, мелодрама,
14	14	"Аргай (сериал)"	"24 мин"	"Франция, "	" Мари-Кристин Адам, Кристиан А	"мультфильм, фанта
15	15	"BBC: Баллада о Большом Але (ми	"60 мин"	"Великобритания (документальный	" Кеннет Брана, Эйвери Брукс, .	"документальный) "

Рис. №2

Шаг №3. Обработка данных.

На следующем шаге извлекаем данные из MongoDB и соединяем в единый Датафрейм. Для этих целей и последующих используем модуль **Pandas**. Далее работаем с каждым столбцом в отдельности: удаляем записи с пропущенными значениями, форматируем значения в столбцах, изменяем типы данных и проч.

Скрипт данного шага находится в файле «**Kinopoisk_data_preparation.ipynb**»

Подготовленные данные записываем в файл «**kinopoisk_prepared.csv**»

Шаг №4. Анализ подготовленных данных.

На следующем шаге мы анализируем обработанные ранее данные. Проводим группировки, агрегации в различных разрезах. Строим несложные визуализации. Используемые инструменты: модули **Pandas, Matplotlib, Seaborn**.

Скрипт данного шага находится в файле «**Kinopoisk_Analysis.ipynb**»

Шаг №5. Анализ подготовленных данных (Tableau).

На заключительном шаге мы анализируем обработанные ранее данные в **Tableau**. Строим несложные визуализации.

Файл с визуализациями Tableau «**Анализ фильмов кинопоиск.twb**»

Заключение.

Данный проект имеет основной целью пройти по шагам цикл работы с данными.

Данный проект также возможно в будущем доработать, получив значительно больше информации по каждому фильму. Это можно сделать путем сбора информации отдельно со страницы фильма, к которой мы можем сделать GET-запрос по уникальной ссылке, которая у нас уже есть.