Анализ данных с КиноПоиска

Дмитрий Курносов*, Никита Лансков†, Михаил Нахатович‡[0000−0002−6279−1130] и Максим Смольский§
Высшая школа прикладной математики и вычислительной физики
Санкт-Петербургский политехнический университет
Санкт-Петербург, Россия
Email: *dima2202888@yandex.ru, †nl516@yandex.ru, ‡mish_1998@mail.ru, §mithridatus@mail.ru

***Аннотация*—Эта статья описывает процесс анализа данных с сайта кинопоиск. В рамках статьи рассмотрены процессы получения и хранения данных, а также их последующей обработки для решения поставленных задач.**

***Ключевые понятия*—Большие данные, распределённые вычисления**

I. Введение

Индустрия фильмов развивается с каждым годом и является важной частью в жизни каждого человека. Также растёт интерес к кинопрокату, ежегодно растёт оборот денежных средств в киноиндустрии, а также качество съёмки и число людей, задействованных в процессе работы над новыми фильмами.

КиноПоиск - крупнейший русскоязычный интернет-сервис о кино. Данный сервис предоставляет информацию о различных фильмах, актёрах, новостях кино и т.д.

В данной работе представлен анализ данных о фильмах: рассмотрены взаимосвязи между различными характеристиками фильмов и построены распределения фильмов по различным критериям. В рамках данной работы мы делали упор на статистические методы анализа данных.

II. Данные

A. Получение данных

Для скачивания данных использовался сторонний API для доступа к актуальной информации КиноПоиска. Так как данный API предоставляет информацию о фильме только по его идентификатору, для получения всех идентификаторов фильмов был выполнен парсинг самого сайта КиноПоиск. Коннектор написан на языке Python с использованием пакета PyMongo для работы с MongoDB из Python. Всего было выкачено 654165 фильмов. Объём данных составил 3.4 Гб.

B. Структура данных

Выгруженную информацию о фильмах можно поделить на блоки, представленные на Рис. 1. Каждый фильм содержит некоторые общие сведения такие как название, год производства, жанры и т.д. Также есть список создателей, то есть список всех актёров, режиссёров и т.д., задействованных в создании фильма. Различные рейтинги, в число которых входит рейтинг из базы IMDb. Рецензии зрителей и бюджет фильма, в который также входят сборы.

C. Обработка данных

Все вычисления производились при помощи системы распределённых вычислений - Apache Spark.

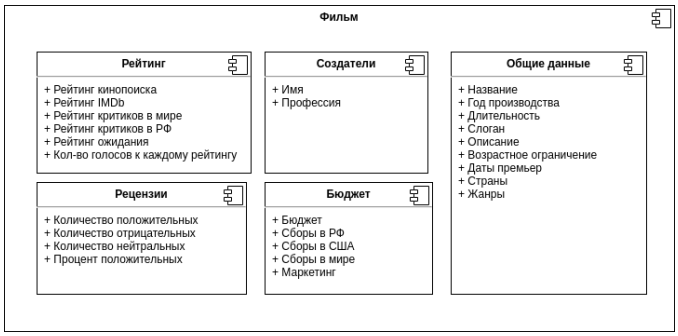

Рис. 1. Схема представления данных.

III. Статистические задачи

A. Корреляция оценок зрителей и критиков

В данной задаче проверялось наличие зависимости между оценками зрителей и критиков на основе корреляционного анализа. Ранговый коэффициент корреляции Спирмена между оценками получился равным 0.33, что говорит о наличии слабой зависимости между ними.

B. Корреляция рейтинга КиноПоиска и рейтинга IMDb

В данной задаче проверялось наличие зависимости между рейтингами фильмов с сайта КиноПоиска и рейтингами фильмов из базы IMDb. Ранговый коэффициент корреляции Спирмена между рейтингами получился равным 0.75, что свидетельствует о довольно хорошей корреляции между ними. Это означает, что зрители, как в России, так и за рубежом ставят схожие оценки фильмам.

C. Распределение фильмов по странам

Задача распределения фильмов по странам, в которых данные фильмы были сняты, позволяет получить представление о количестве фильмов, которые были сняты в конкретной стране, а также выделить те из них, количество снятых фильмов у которых наибольшее. Для реализации этой задачи используется информация из блока общих данных (Рис. 1), а именно иформация о странах, снявших конкретный фильм. Далее происходит суммирование полученной информации и сортировка в порядке убывания. Итогом работы является таблица, в которой отображены названия стран и соответствующие им значения снятых фильмов. Для удобства визуализации выбраны первые 15 позиций.

Представленный на Рис. 2 график демонстрирует, что первенство по количеству снятых фильмов с огромным отрывом достается США, второе и третье место занимают Великобритания и Франция соответственно.

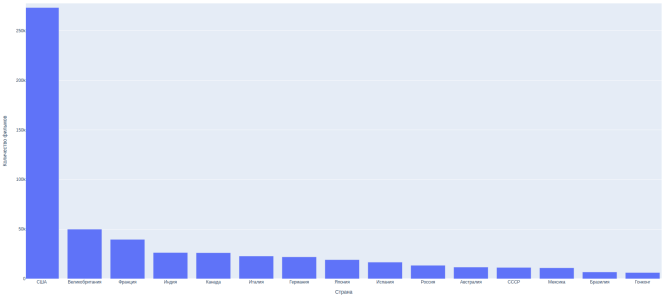

Рис. 2. Количество фильмов, снятых в конкретных странах

D. Распределение фильмов по прибыльности

В рамках данной задачи было построено распределение фильмов по их прибыльности. Прибыльность вычислялась по следующей формуле:

$$Benefits = BoxOffice - Budget$$

В результате, в топ 5 самых прибыльных фильмов вошли такие фильмы, как:

1) Аватар
2) Мстители: Финал
3) Звёздные войны: Пробуждение силы
4) Мстители: Война бесконечности
5) Титаник

А самыми неприбыльными фильмами стали следующие:

1) Ирландец
2) Мулан
3) Асура
4) Приключения Плуто Нэша
5) Остров головорезов

E. Распределение фильмов между странами по годам

В продолжение задачи распределения фильмов по странам было найдено распределение фильмов между странами по годам. Самым ранним годом в базе КиноПоиска оказался 1874 год, в который был снят лишь один фильм в Франции. По полученным данным в ранние годы фильмы снимались только в Франции, США и Великобритании. За 2020 год в топ-10 стран вошли США, Великобритания, Индия, Канада, Австралия, Франция, Испания, Россия, Германия и Италия. При этом как и в распределении фильмов по странам за всё время, США занимают первое место по количеству выпущенных фильмов почти в каждый год, в который они выпускали фильмы. Исключения составили лишь 1906 (Великобритания), 1905 (Великобритания), 1897 (Франция), 1896 (Франция) и 1893 (Российская империя) года.

F. Распределение фильмов по возрастным ограничениям и годам

В данной задаче рассматривалось распределение фильмов по возрастным ограничениям и годам. Рассматривались только фильмы с наличием возрастного ограничения. Всего имеется пять разных возрастных ограничений - 0, 6, 12, 16 и 18 лет. Из Рис. 3 видно, что до 1960-х годов фильмов с возрастным ограничением почти не было. Далее до 1990-х годов фильмов 16+ было существенно больше, чем всех остальных, а начиная с 1990-х годов стало больше фильмов 18+. В последнее время фильмов 16+ и 18+ стало примерно одинаково, фильмов 12+ в 2 раза меньше, фильмов 6+ в 4 раза меньше и фильмов 0+ в 8 раз меньше.

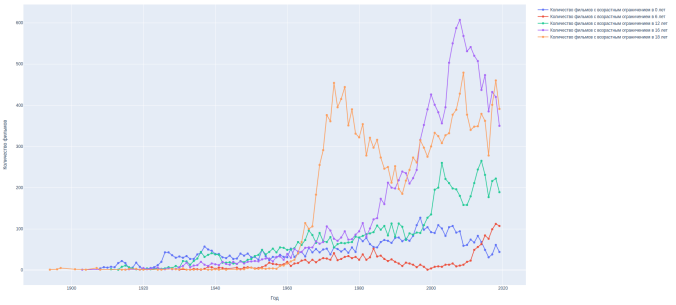

Рис. 3. Распределение фильмов по возрастным ограничениям и годам.

G. Средний рейтинг российских фильмов по годам

В число стран-производителей, предоставляемых КиноПоиском, входит не только Россия, но и СССР, и Российская империя. Поэтому было решено рассмотреть фильмы, выпущенные не только в России, но и в предшествующих странах. У фильмов, выпущенных в период гражданской войны, указана страна «Россия», поэтому нельзя точно определить страну-производитель, используя только данные КиноПоиска. На графике на Рис. 4 по оси абсцисс отложен год, а по оси ординат средний рейтинг по всем фильмам, выпущенным в этот год. Размер маркера прямопропорционален числу фильмов, выпущенных за конкретный год. Видно, например, что в СССР снимали меньше фильмов, но их рейтинг был лучше, чем у современных фильмов.

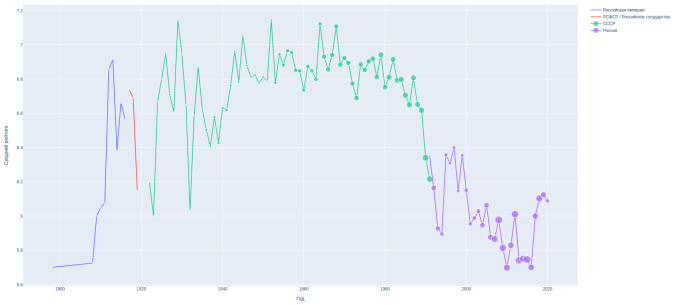

Рис. 4. Средний рейтинг российских фильмов по годам.

IV. Исследовательские задачи

A. Прогноз количества фильмов по жанрам на 10 лет

В данной задаче прогнозировалось количество фильмов по жанрам на 10 лет. Сразу стоит отметить, что рассматривались только фильмы до 2019-го года включительно, так как из-за пандемии коронавируса в 2020-ом году количество фильмов резко сократилось. Поэтому прогнозируется количество фильмов по жанрам, если бы в 2020-ом году не было пандемии коронавируса. В основном, для всех жанров прогнозируется продолжение роста количества фильмов - примером являются мультфильмы (см. Рис. 5). Но встречаются и редкие жанры, для которых на основе чередования роста и снижения количества фильмов в последнее время

спрогнозировалось продолжение этого чередования - примером является жанр фэнтези (см. Рис. 6).

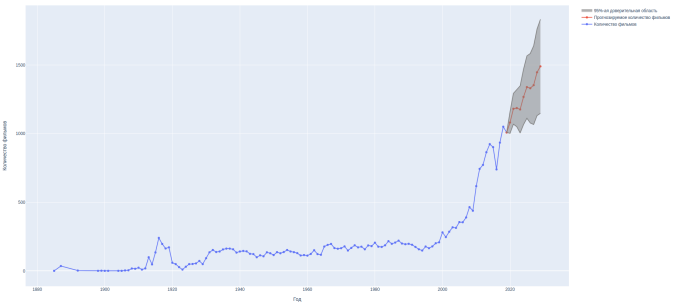

Рис. 5. Прогноз количества мультфильмов на 10 лет.

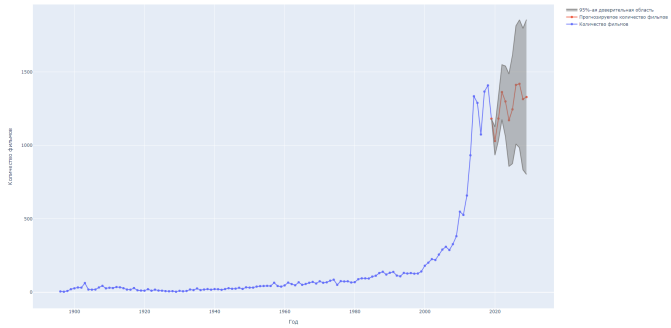

Рис. 6. Прогноз количества фильмов жанра фэнтези на 10 лет.

V. Дальнейшие исследования

Во-первых, мы хотим построить модель, с помощью которой можно предсказывать рейтинг фильма по его создателям, например, актёрам, режиссёрам и сценаристам, а также другим данным, таким как жанры или страны производители. Для этой задачи нужно понять, как эффективно связать категориальные признаки с итоговым числовым значением. Во-вторых, мы хотим выкачать сериалы и выполнить статистические задачи для них.

Список литературы

[1] Сайт «КиноПоиск». [Электронный ресурс]. Режим доступа: https://www.kinopoisk.ru/
[2] Сайт «MongoDB». [Электронный ресурс]. Режим доступа: https://www.mongodb.com/