

Анализ данных с КиноПоиска

Дмитрий Курносов*, Никита Лансков*, Михаил Нахатович*[0000–0002–6279–1130] и Максим Смольский*

*Институт прикладной математики и механики

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия

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

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.

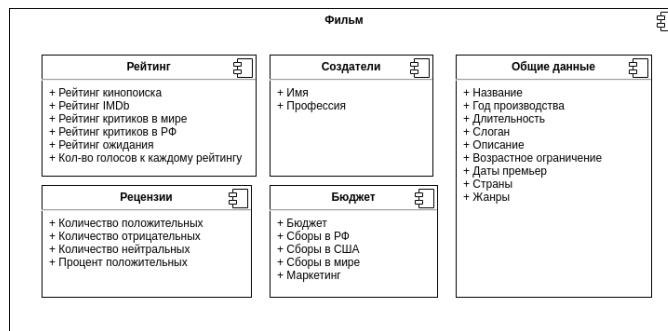


Рис. 1. Схема представления данных.

C. Обработка данных

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

III. Статистические задачи

- A. Корреляция оценок зрителей и критиков
- B. Корреляция рейтинга КиноПоиска и рейтинга IMDb
- C. Распределение фильмов по странам
- 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

- Е. Распределение фильмов между странами по годам
- Ф. Распределение фильмов по возрастным ограничениям и годам
- Г. Средний рейтинг российских фильмов по годам

IV. Исследовательские задачи

- А. Прогноз количества фильмов по жанрам 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/>