*Облакова Кристина*

**Итоговый проект**

Ссылка на [Кинопоиск](https://www.kinopoisk.ru/lists/movies/?b=films&b=released&b=foreign&page=263)

Ссылка на исходные данные в [Google таблице](https://docs.google.com/spreadsheets/d/1AiGnRCZ0eVHQm8UG_7WlFyFaXJZBQRJF/edit?usp=drive_link&ouid=117082241032387795397&rtpof=true&sd=true)

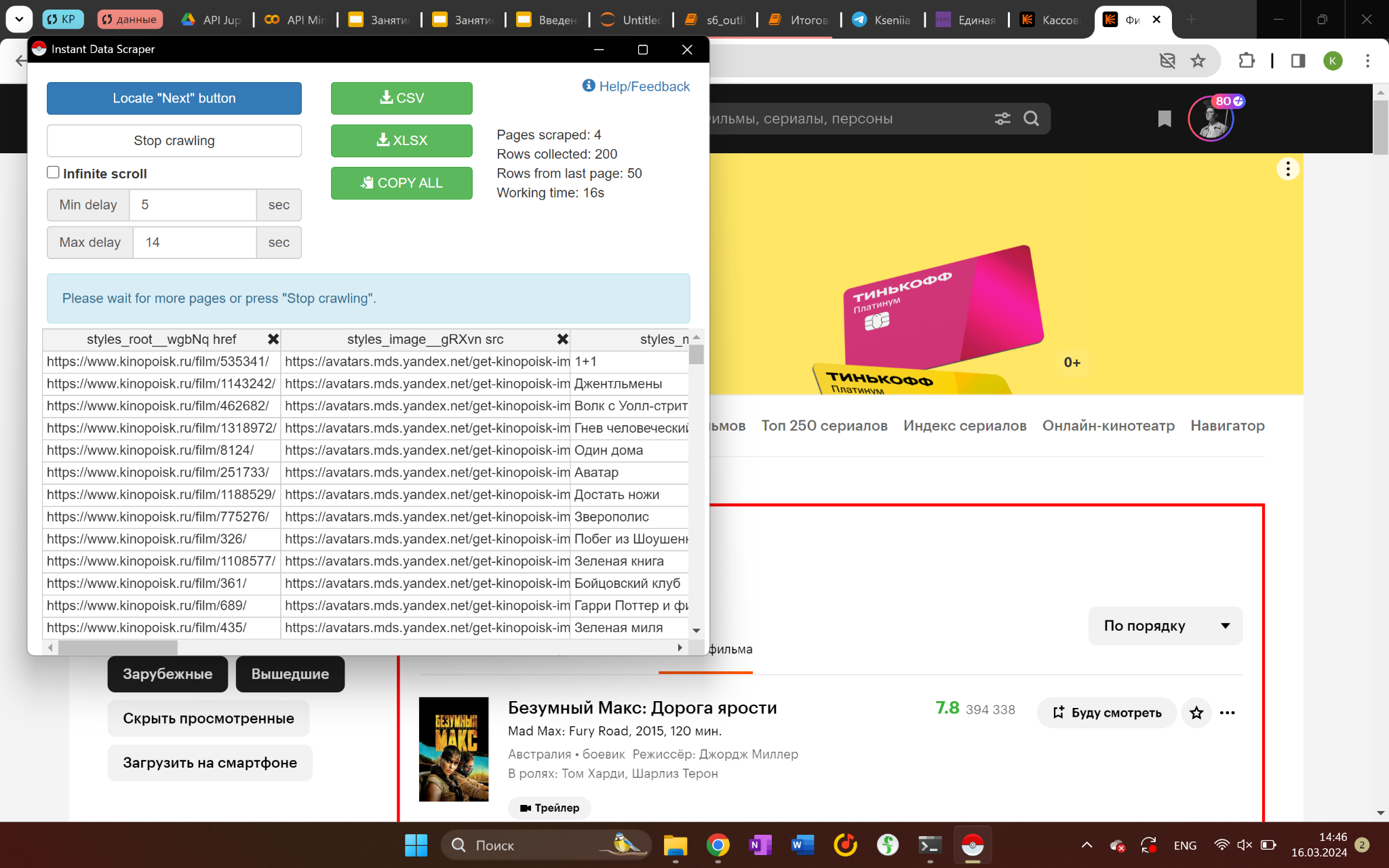
Ссылка на [ноутбук](https://drive.google.com/file/d/1HoghdlB8QuoJX3MOE5FbwW-FDl5rM7S5/view?usp=drive_link)

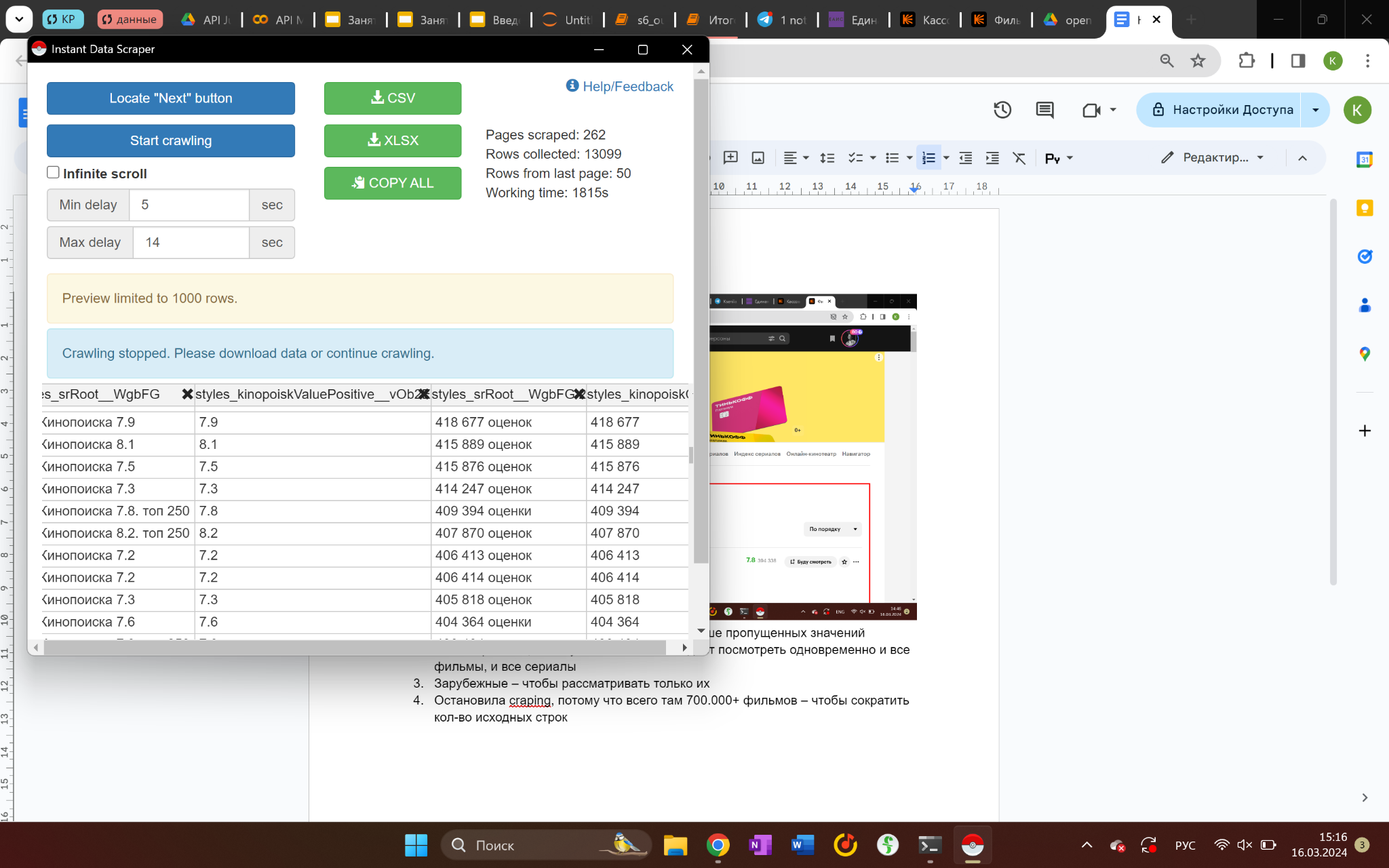
Ссылка на [промежуточную Google таблицу](https://docs.google.com/spreadsheets/d/1K2I9GIqUwAsM53bxjI-uOSdr8r3lMYpq/edit?usp=drive_link&ouid=117082241032387795397&rtpof=true&sd=true)

Ссылка на [папку](https://drive.google.com/drive/folders/1fg4e2ra5WdxTlbtt4z2rTjOwRFtHb7zt?usp=drive_link) со всеми файлами проекта

Для выполнения итогового проекта я выбрала **Кинопоиск** как источник для датасета, так как он является перечнем и сборником фильмов и сериалов, и на нем их собрано наибольшее число по сравнению с другими платформами. Источник хорошо **структурирован**, это известный **модерируемый** ресурс, поэтому информация на нем достоверная.

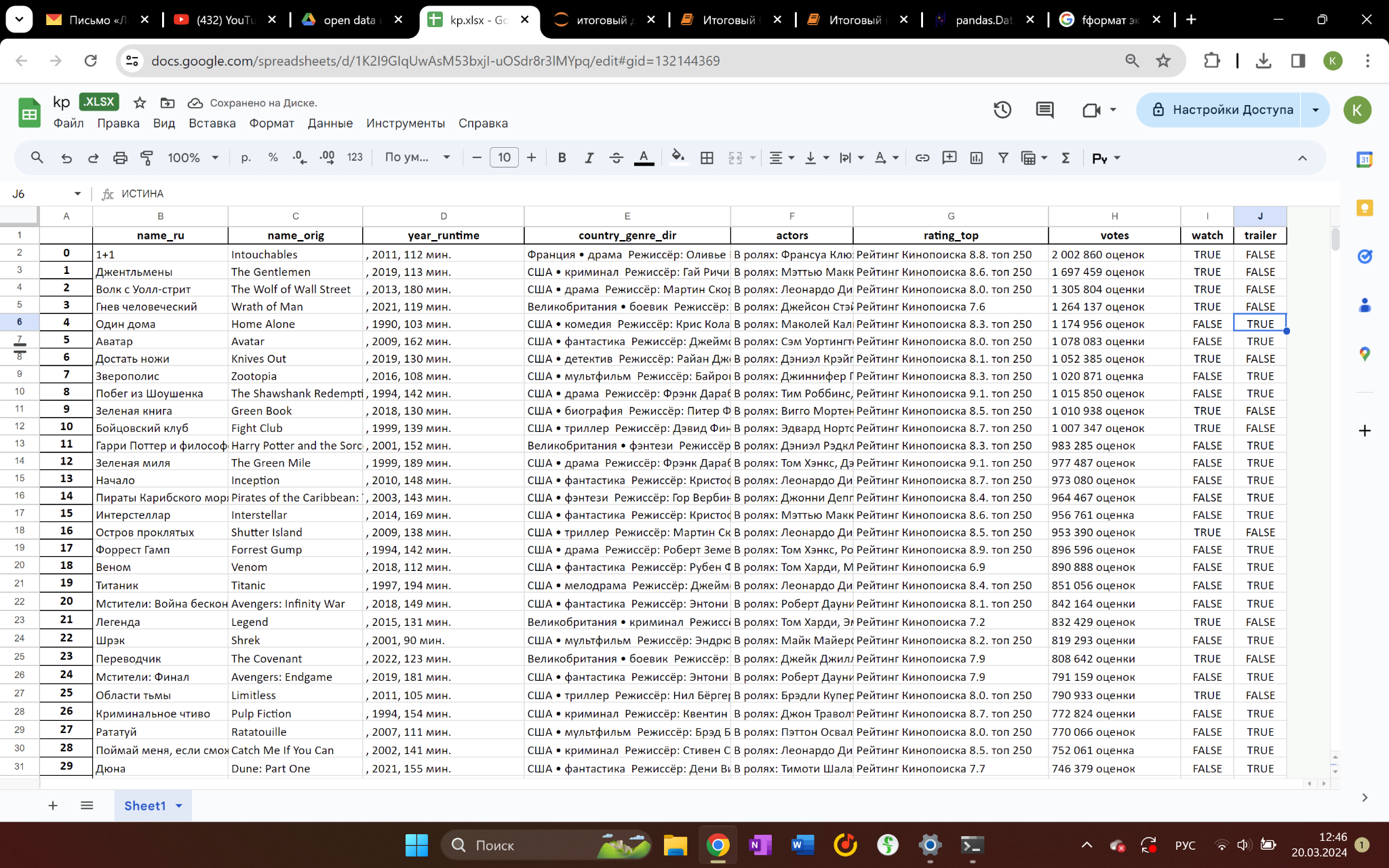
Я решила анализировать все **зарубежные фильмы**, размещенные на Кинопоиске. Чтобы получить данные, я с помощью **Instant Data Scrape**r собрала данные. Перед этим выставила несколько **ограничений**:

1. **Вышедшие** – чтобы потенциально было меньше пропущенных значений
2. Только **фильмы** и только **зарубежные**, потому что это цель анализа
3. **Остановила scraping** на 13.099 строках, потому что всего там размещено 700.000+ фильмов



Далее сохранила файл xlsx и csv. Оба [загрузила](https://docs.google.com/spreadsheets/d/1AiGnRCZ0eVHQm8UG_7WlFyFaXJZBQRJF/edit?usp=drive_link&ouid=117082241032387795397&rtpof=true&sd=true) в папку с проектом на Google диске (на всякий случай, как резервные копии), csv перенесла в **Jupiter** для обработки: [**ссылка на ноутбук**](https://drive.google.com/file/d/1HoghdlB8QuoJX3MOE5FbwW-FDl5rM7S5/view?usp=drive_link). Там начала чистить датасет, описание очистки там – внутри блокнота.

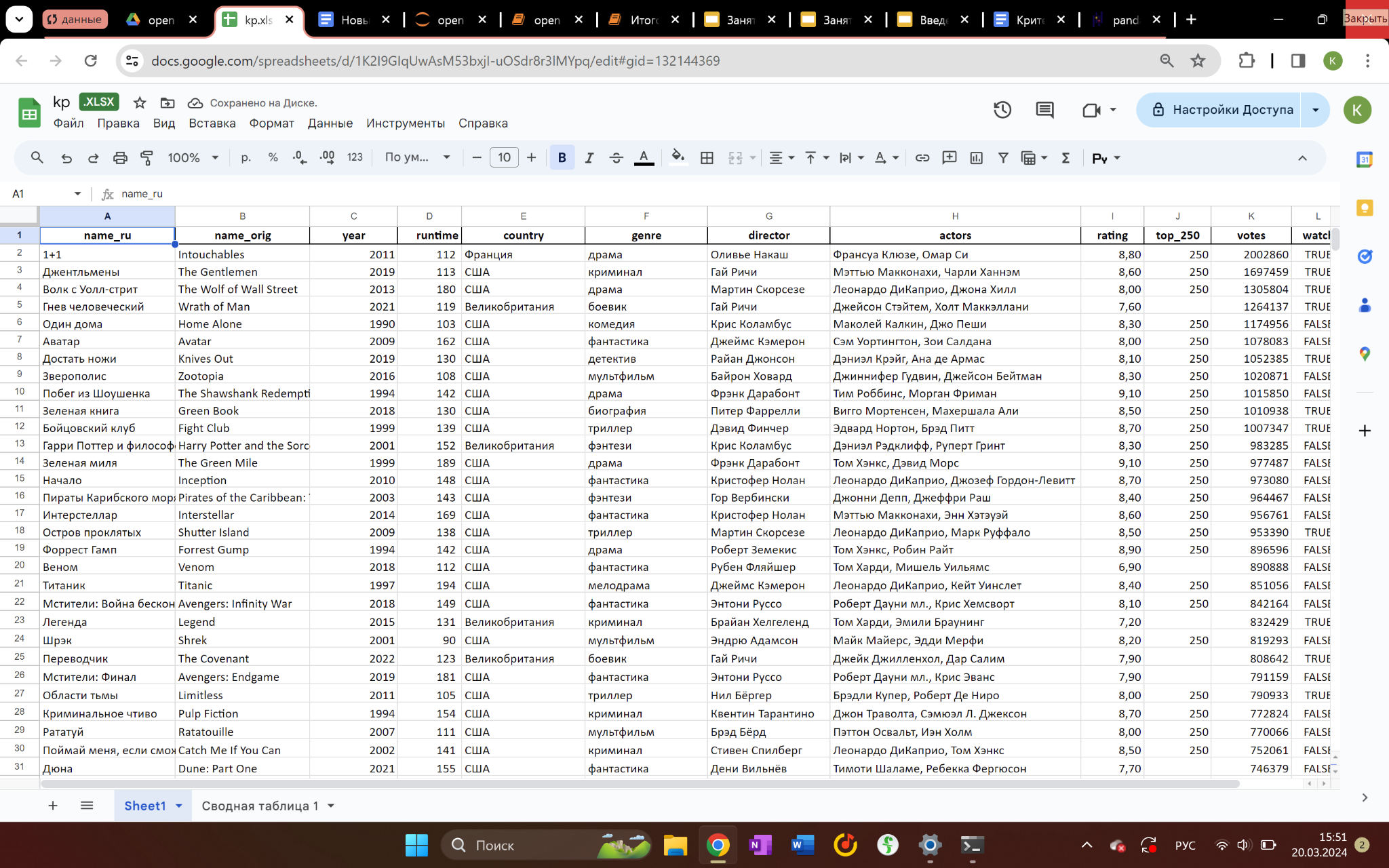
На этапе обработки некоторые действия легче было сделать через **Google таблицы**. Вот так выглядел датасет после Jupiter:



В нем я:

1. Разделила **year\_runtime**
2. Разделила **country\_genre\_director**, убрала слово “Режиссёр”. Там остались пробелы в начале, их можно убрать в pandas/
3. Убрала “В ролях: “ из столбца **actors**
4. Разделила **rating** и **top**, в рейтинге заменила точки на запятые, чтобы получились числовые значения
5. Посмотрела в сводной таблице, что из списков-топ есть только **топ-250**, решила сделать это буллевым значением позже
6. Убрала слова “оценка”, “оценок”, “оценки” из **votes**, убрала пробелы, чтобы значения были числами

Вот так он стал выглядеть после:



После этого я вернулась в **Jupiter** для завершения обработки и дальнейшего анализа данных.