Классификация текстового контента

Александр Смирнов и Феодор Жилкин

17.05.2019г

Введение

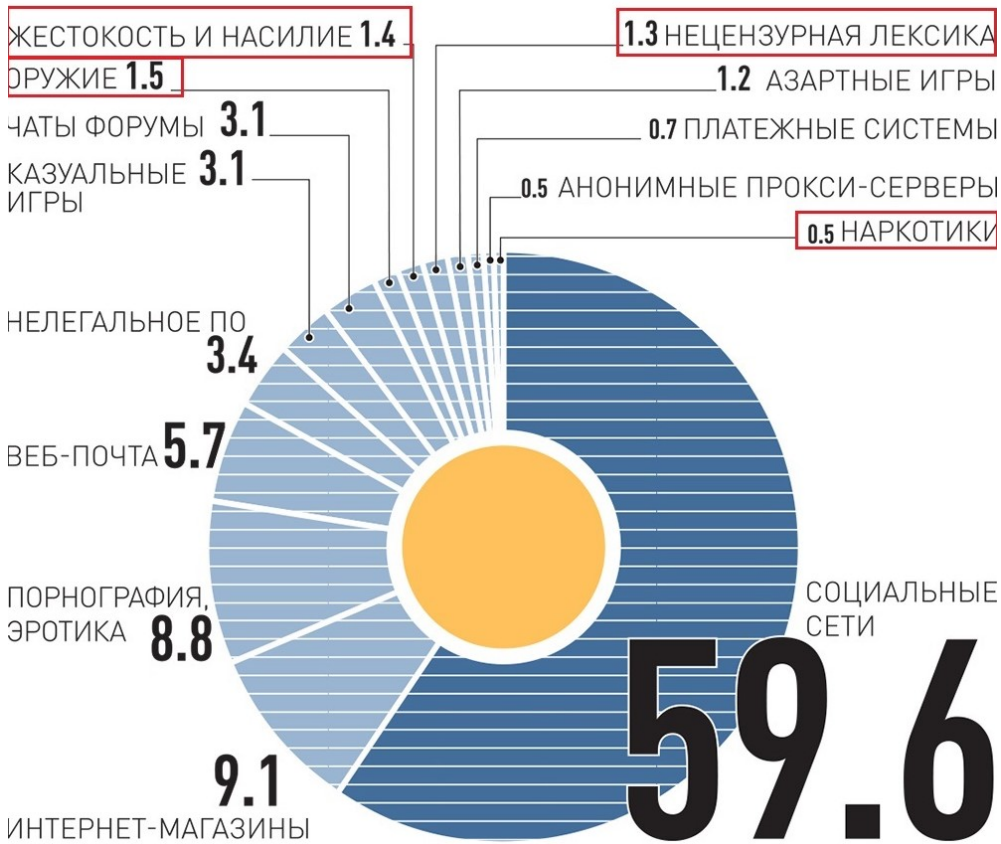

Рис.: Что интересует детей в интернете

Цели

- Ограничить детей от взрослого текстового контента
- Получение опыта
  - Нейросети
  - Python, Jupyter Notebook, JS
  - Майнинг датасета и составление csv-файла для загрузки на https://www.kaggle.com
  - Написание собственной Python-библиотеки
  - Написание расширения для Chrome
  - Написание Python-сервера для приёма запросов

Задачи

- Сделать расширение для Chrome
- Внести вклад в сообщество разработчиков
  - Датасет на https://www.kaggle.com
  - Python-библиотека

Сравнение с аналогами

- Ограничения на поиск
  - Семейный поиск Яндекс
  - Безопасный поиск Google
- Контентная фильтрация
  - Traffic Inspector
  - Интернет Цензор

Сравнение с аналогами

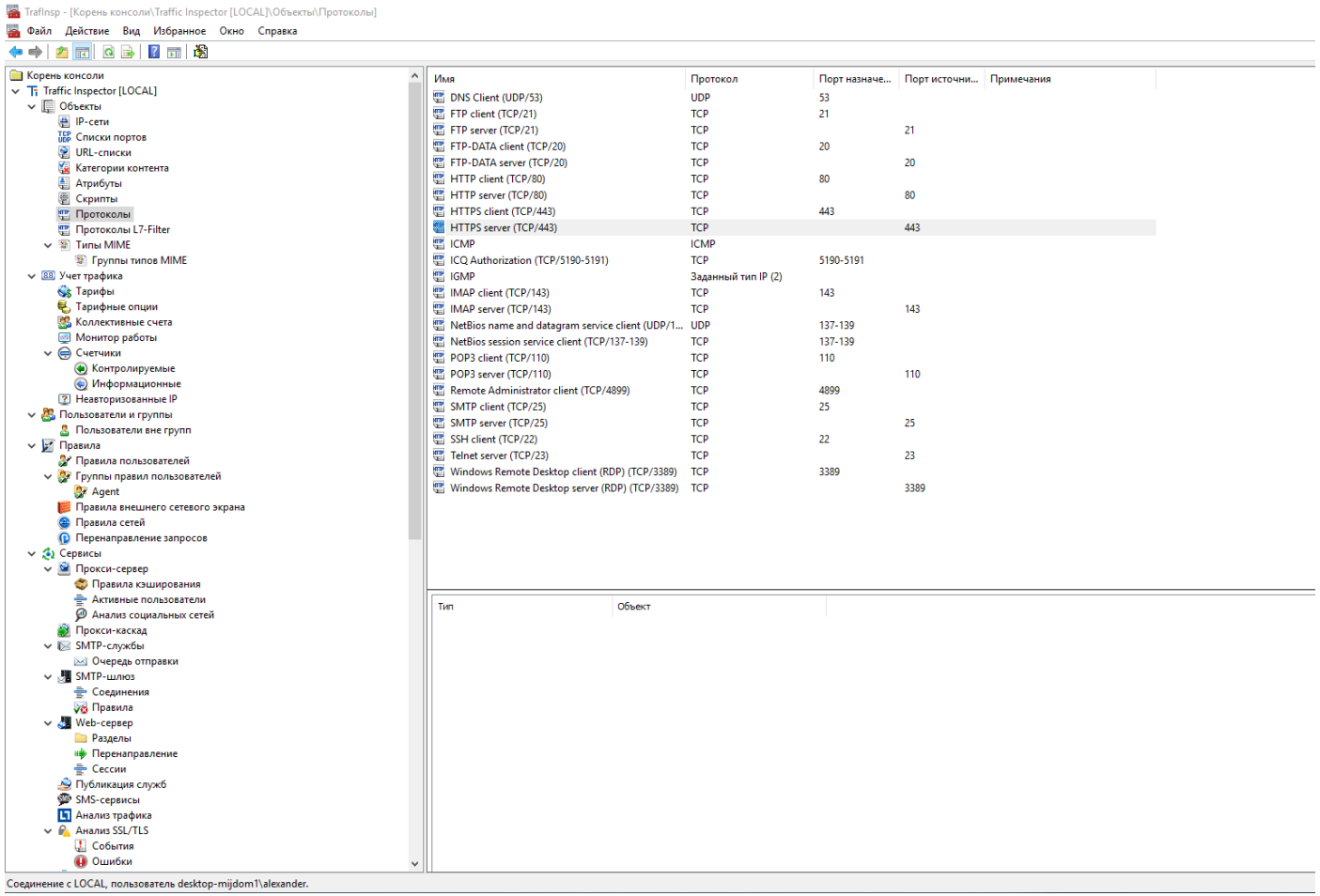

Рис.: Пример интерфейса схожей программы

Результаты

- Расширение для Chrome
- Библиотека на https://pypi.org/project/TalesParse/
- Датасет на https://www.kaggle.com/idoldev/adult-and-child-russian-tales-dataset-csv

Расширение для Chrome (1)

Рис.: Блокировка контента

Расширение для Chrome (2)

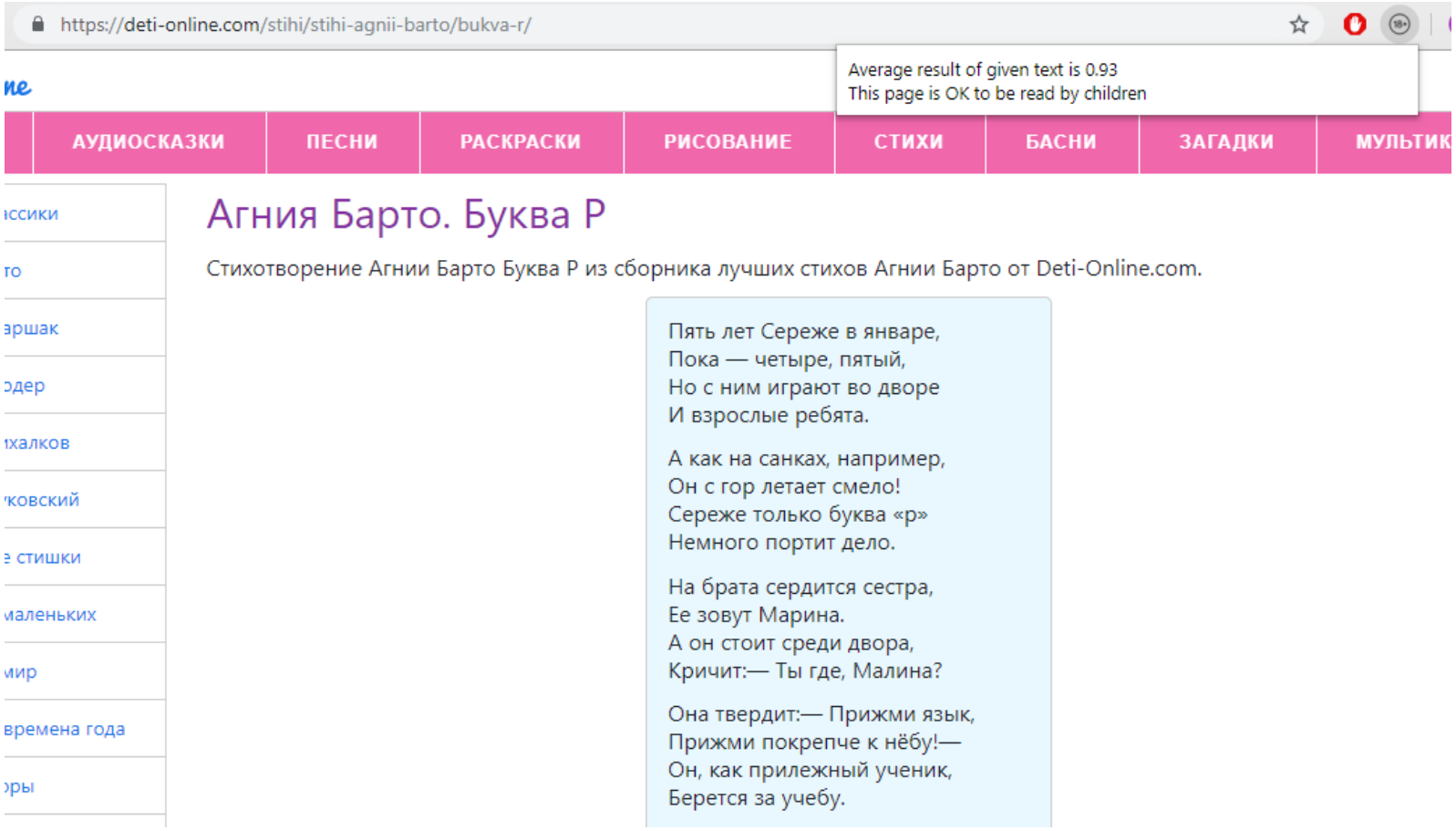

Рис.: Допуск до контента

Библиотека на https://pypi.org

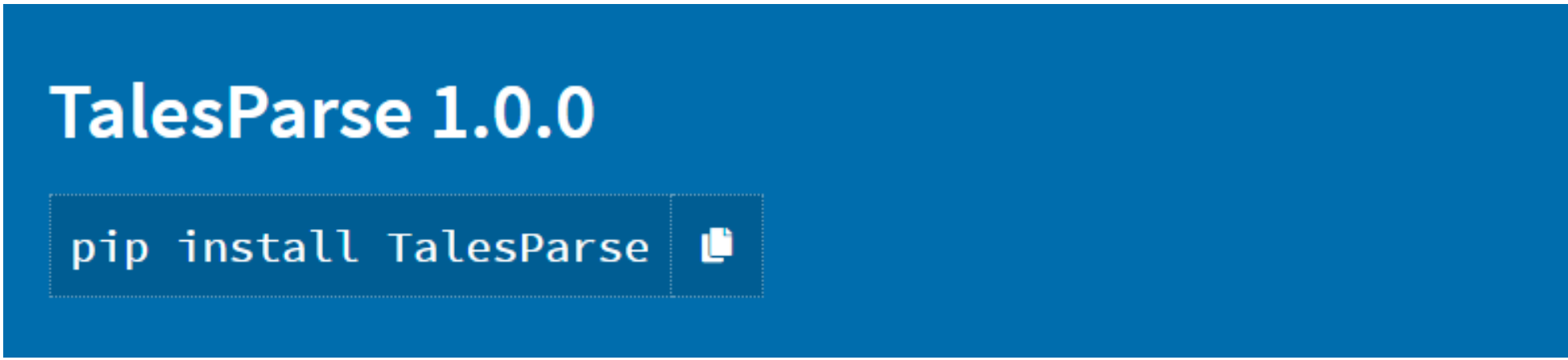

Рис.: Библиотека

Датасет на https://kaggle.com

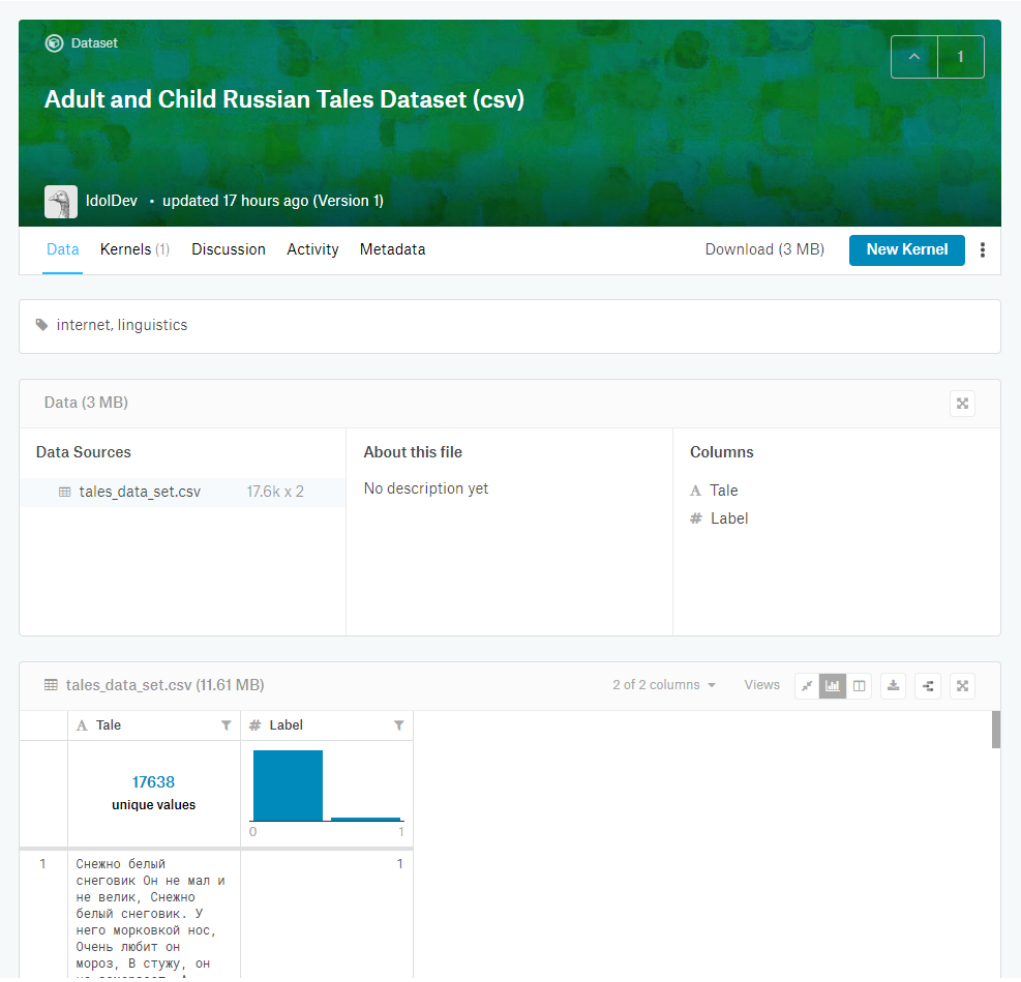

Рис.: Датасет

Как это всё работает

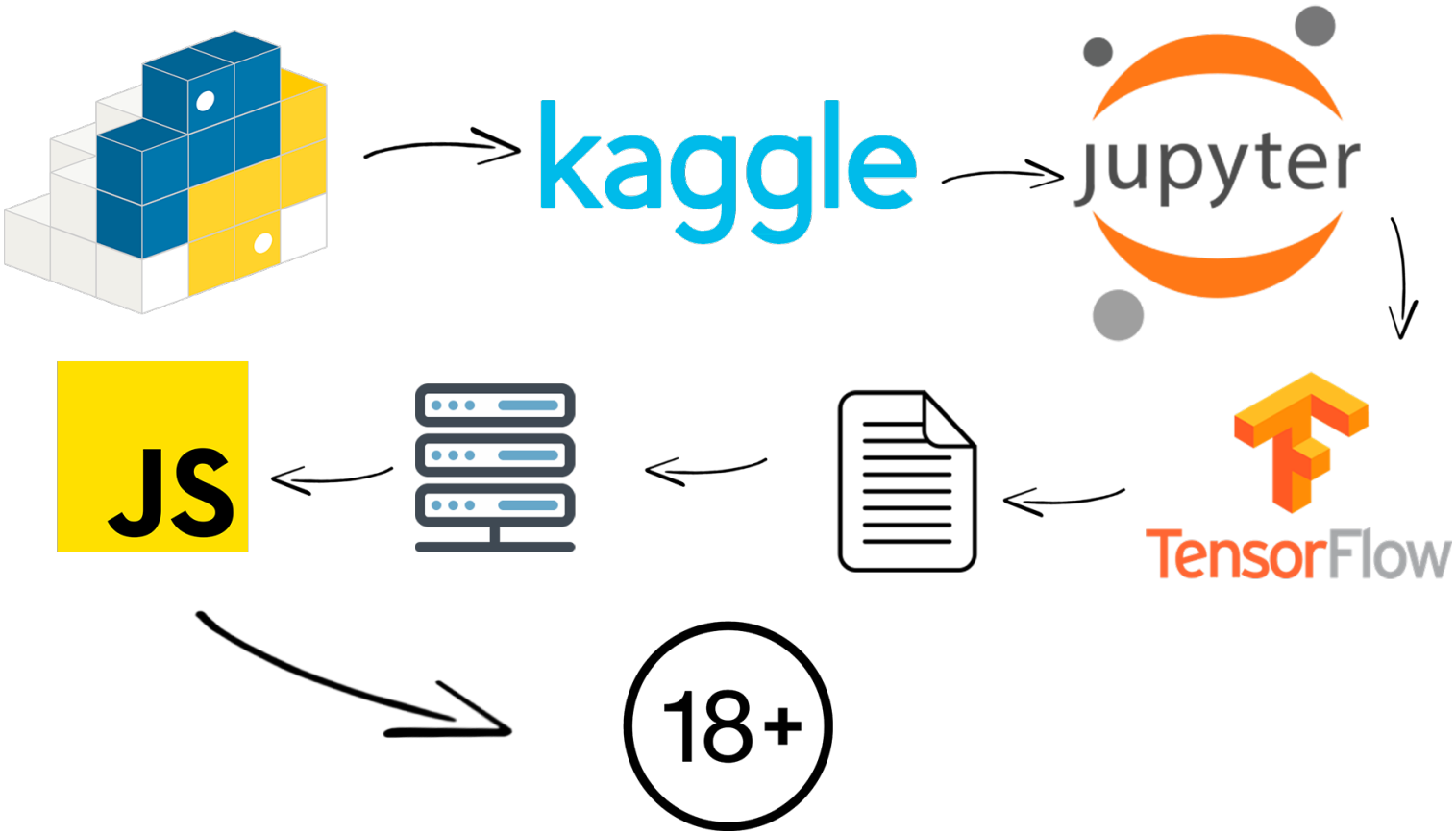

Рис.: Схема проекта

Итоги

- Феодор
  - Парсинг
  - Python-библиотека
  - Датасет
- Александр
  - Нейросеть
  - Сервер
  - Расширение

Результаты

- Проект - https://github.com/SmirnovAlexander/PoemClassifier
- Парсер - https://github.com/Feodoros/Scraping_Tales
- Библиотека - https://pypi.org/project/TalesParse/
- Датасет - https://www.kaggle.com/idoldev/adult-and-child-russian-tales-dataset-csv