Классификация текстового контента

Александр Смирнов и Феодор Жилкин

17.05.2019г

Введение

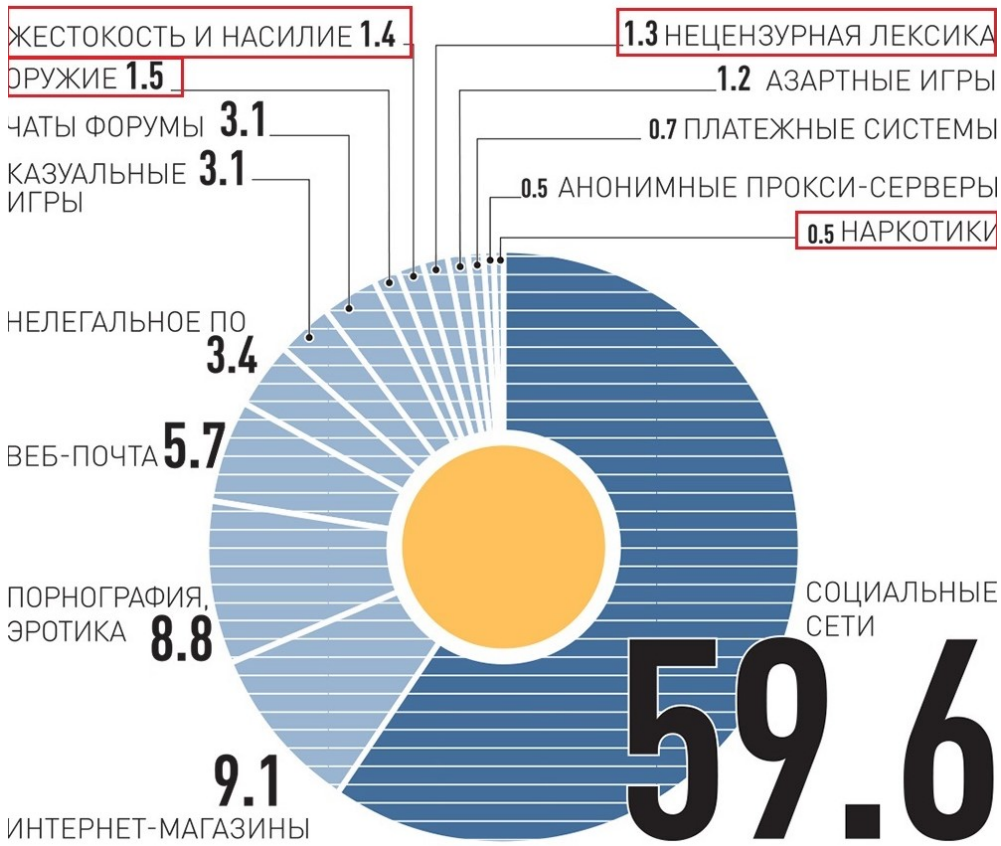


Рис.: Что интересует детей в интернете

Цели

- Ограничить детей от взрослого текстового контента в интернете
- Получение опыта
  - Нейросети
  - Майнинг датасета
  - Написание Python-библиотеки
  - Написание расширения для Chrome
  - Написание Python-сервера для приёма запросов

Задачи

- Провести анализ возможных решений для классификации текста
- Спарсить датасет с взрослыми и нормальными историями
- Написать Python-сервер, использующий обученную модель для ответа на запросы от расширения
- Сделать расширение для Chrome, анализирующее текстовый контент

Сравнение с аналогами

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

Сравнение с аналогами (2)

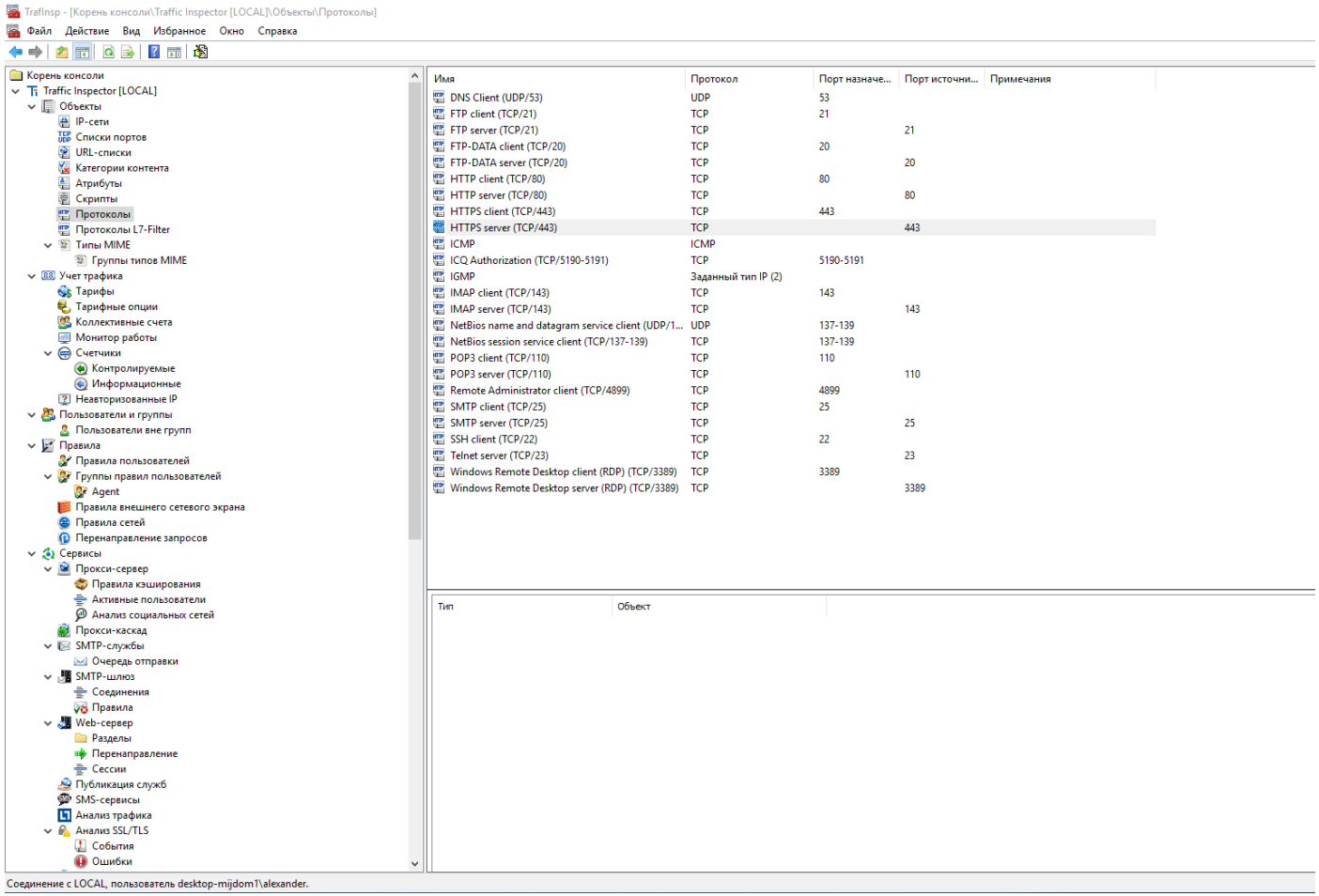


Рис.: Пример интерфейса схожей программы

Результаты

- Расширение для Chrome
- Библиотека на pypi
- Датасет на kaggle

Анализ подходов

- Rule-based
- Machine Learning based
- Hybrid systems

Характеристики сравнения эффективности

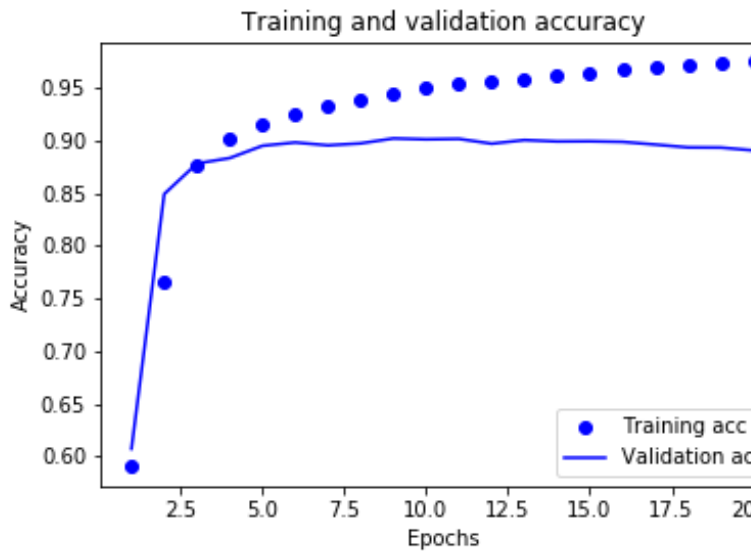
- Accuracy – общая точность классификатора
- Recall – отношение заблокированных взрослых сайтов к общему количеству взрослых сайтов (% классифицированных взрослых сайтов)
- Precision – отношение заблокированных взрослых сайтов к числу всех заблокиронных сайтов (точность блокировки)
- F1 Score - среднее гармоническое между Precision и Recall, для учёта и того, и другого в одной величине

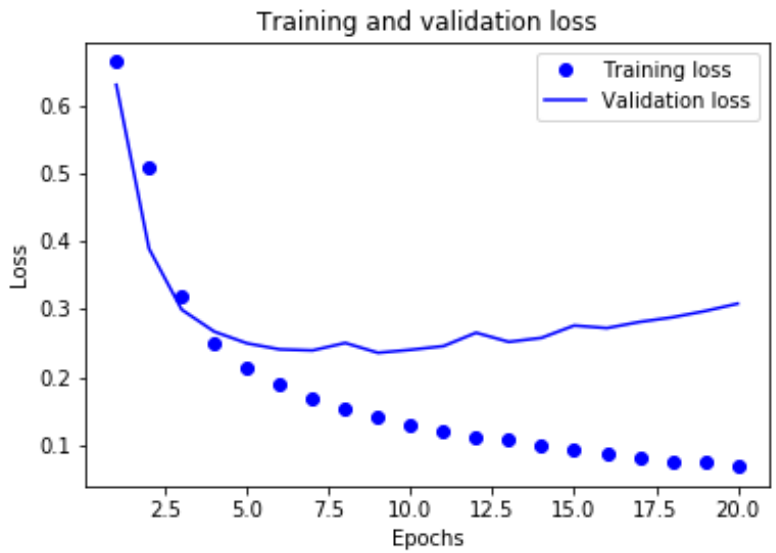
Сравнение

- Random model – случайная выборка блокировать/ не блокировать
- Rule-based model – блокируем по списку непотребных слов
- Classifier – 3-х слойная обычная сеть
- Upgraded Classifier – Classifier, из словаря которой были исключены самые частые слова и добавлена ненормативная лексика

	Accuracy	Recall	Precision	F1 Score
Random model	0.51	0.58	0.58	0.58
Rule-based model	0.41	0.03	1.0	0.06
Classifier	0.88	0.93	0.87	0.90
Upgraded Classifier	0.90	0.92	0.91	0.91

Результаты обучения





Расширение для Chrome (1)

Рис.: Блокировка контента

Расширение для Chrome (2)

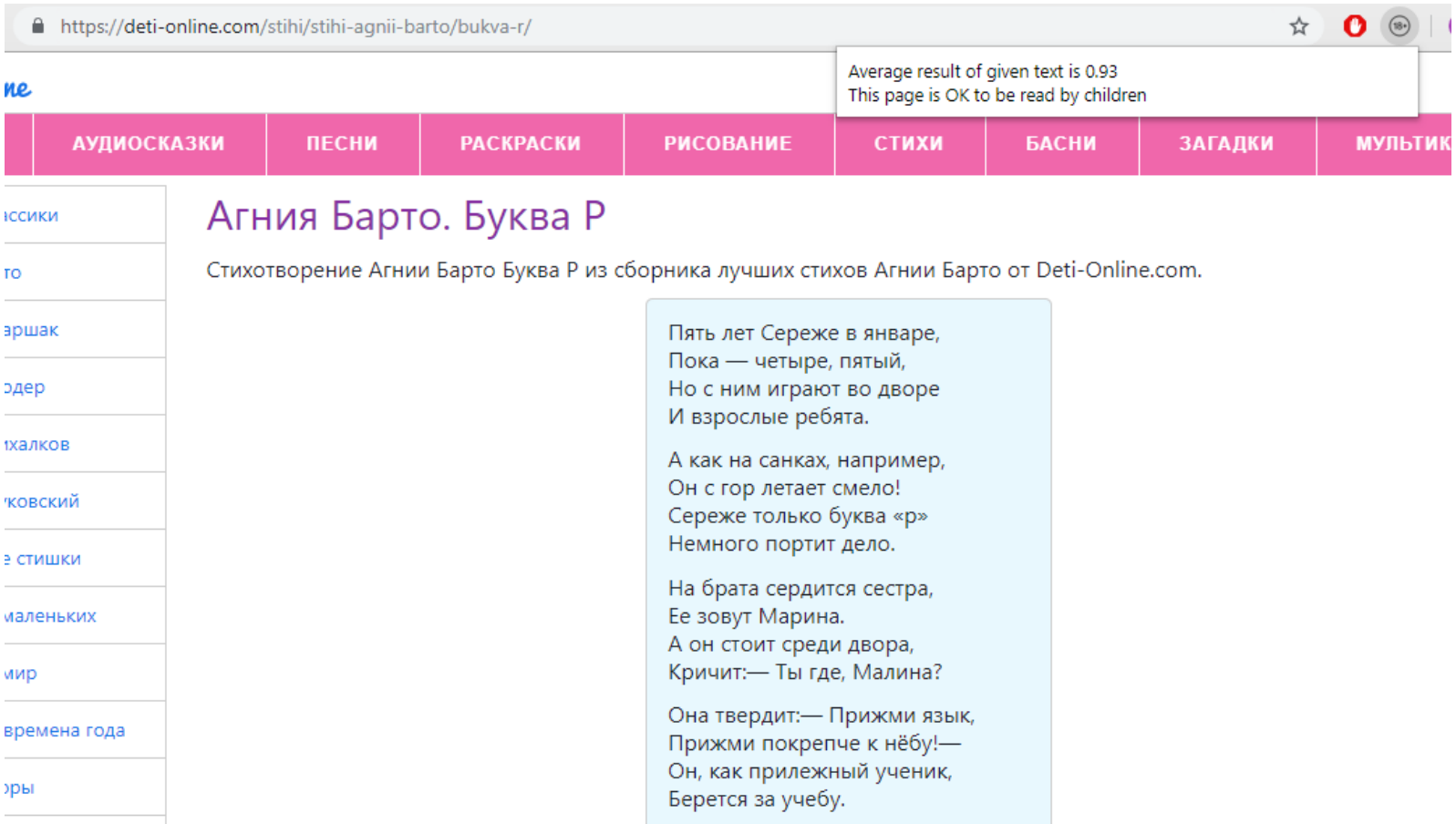


Рис.: Допуск до контента

Библиотека на pypi

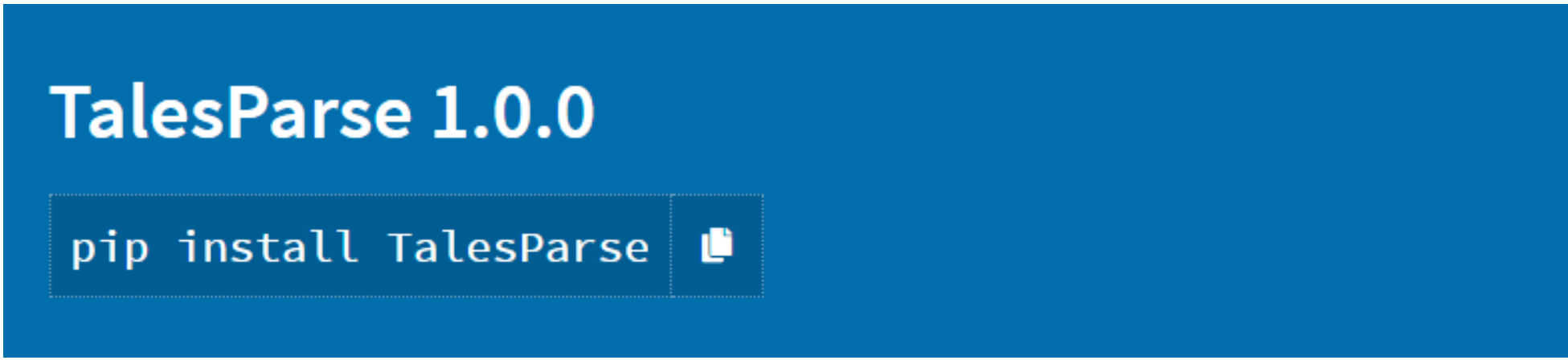


Рис.: Библиотека

Датасет на kaggle

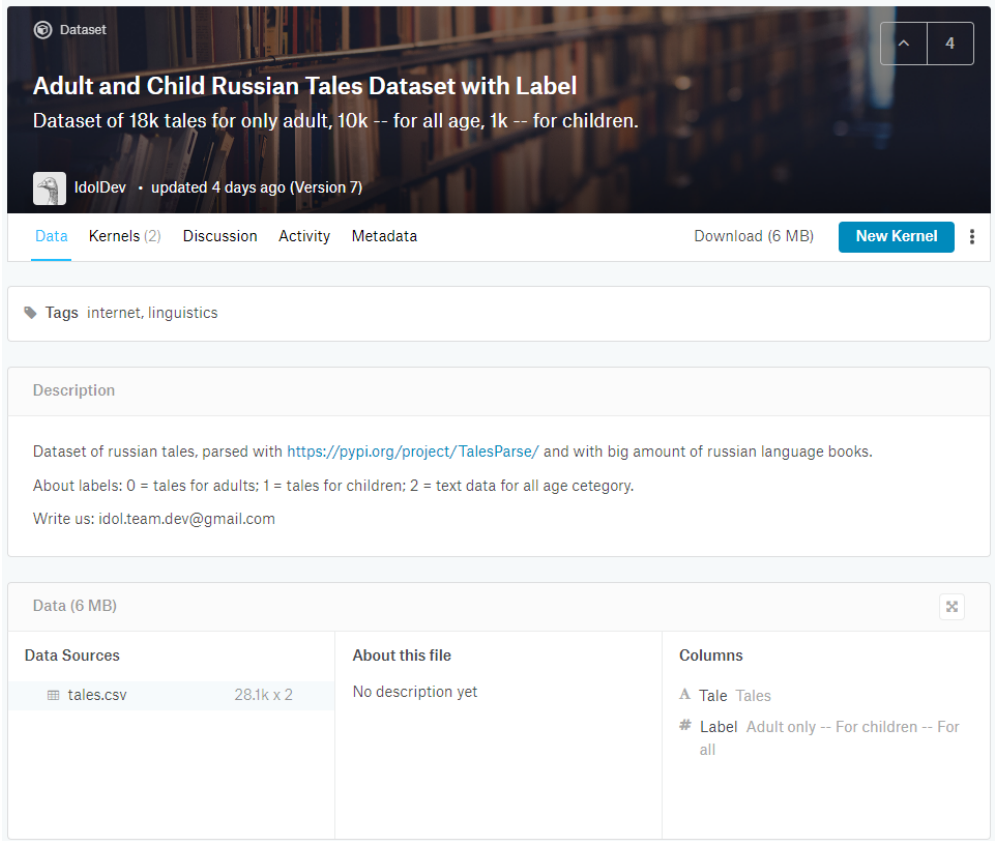


Рис.: Датасет

Как это всё работает

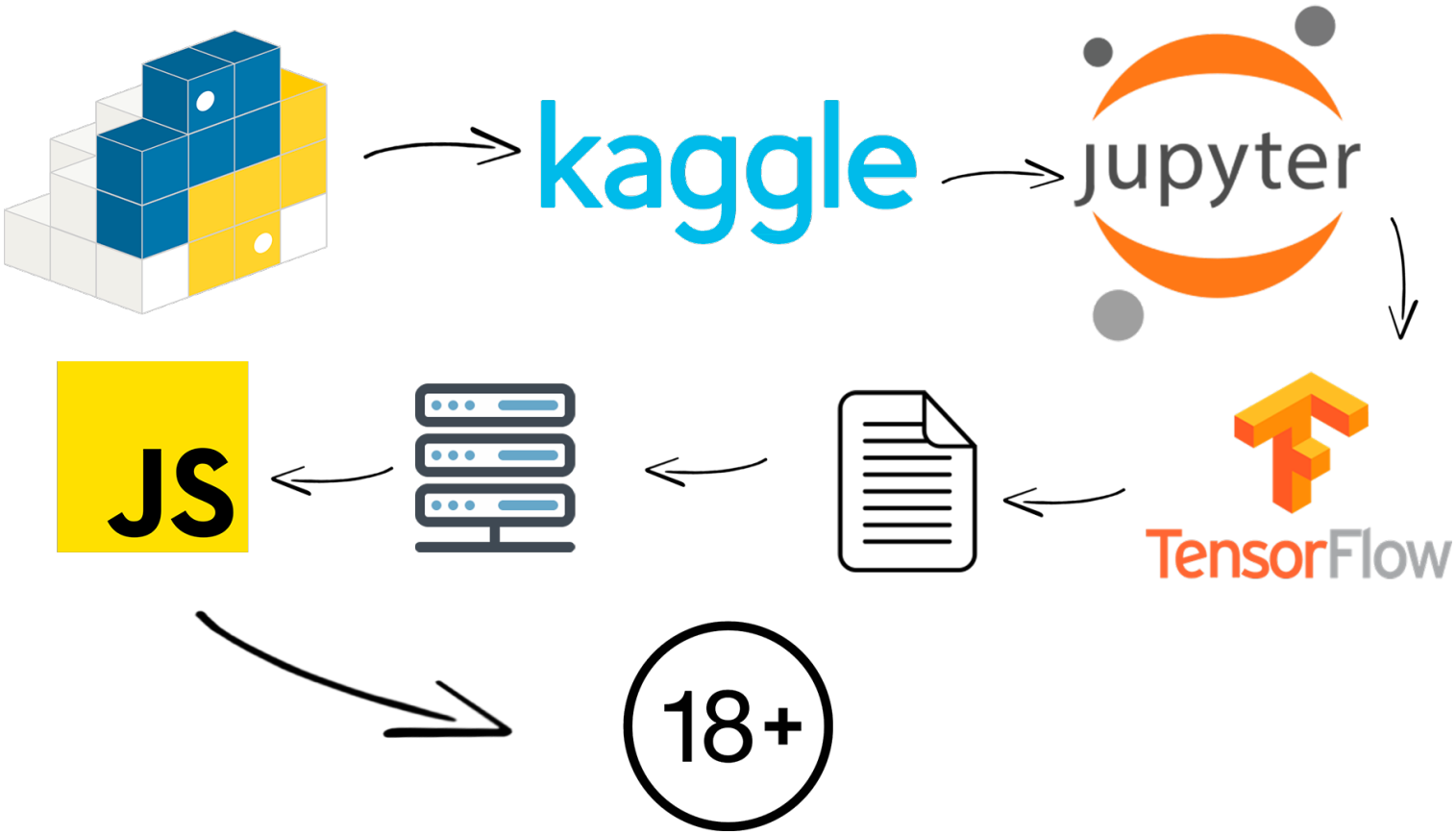


Рис.: Схема проекта

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-with-label