

Классификация текстового контента

Александр Смирнов и Феодор Жилкин

17.05.2019г

Введение

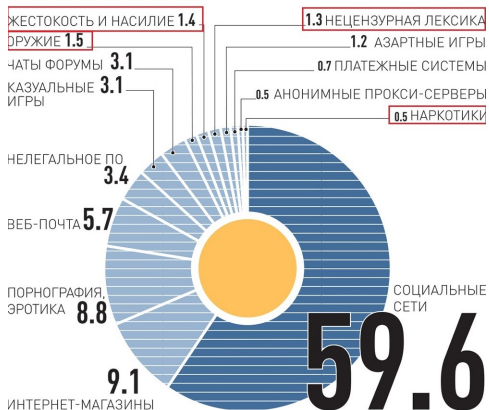


Рис.: Что интересует детей в интернете

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

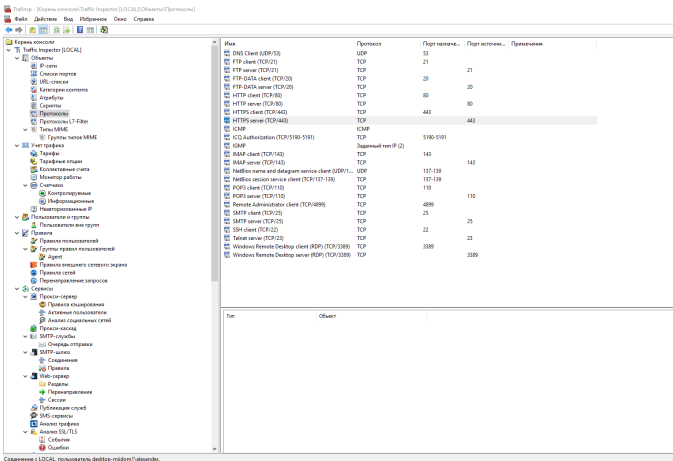


Рис.: Пример интерфейса схожей программы

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)

The screenshot shows a Chrome browser window with the address bar displaying `https://deti-online.com/stihi/stihi-agnii-barto/bukva-r/`. A content rating overlay from a Chrome extension is visible in the top right corner, showing the text: "Average result of given text is 0.93" and "This page is OK to be read by children". The website itself has a pink header with navigation tabs: "АУДИОСКАЗКИ", "ПЕСНИ", "РАСКРАСКИ", "РИСОВАНИЕ", "СТИХИ", "БАСНИ", "ЗАГАДКИ", and "МУЛЬТИКИ". On the left, there is a vertical sidebar with a list of categories: "исики", "го", "вршак", "дер", "халков", "ковский", "стишки", "маленьких", "мир", "времена года", and "ры". The main content area is titled "Агния Барто. Буква Р" and includes a subtitle: "Стихотворение Агнии Барто Буква Р из сборника лучших стихов Агнии Барто от Deti-Online.com." Below the title, a light blue box contains the text of the poem:

Пять лет Сереже в январе,
Пока — четыре, пятый,
Но с ним играют во дворе
И взрослые ребята.

А как на санках, например,
Он с гор летает смело!
Сереже только буква «р»
Немного портит дело.

На брата сердится сестра,
Ее зовут Марина.
А он стоит среди двора,
Кричит:— Ты где, Малина?

Она твердит:— Прижми язык,
Прижми покрепче к нёбу!—
Он, как прилежный ученик,
Берется за учебу.

Рис.: Допуск до контента

Библиотека на <https://pypi.org>

TalesParse 1.0.0

```
pip install TalesParse
```



Рис.: Библиотека

Датасет на <https://kaggle.com>

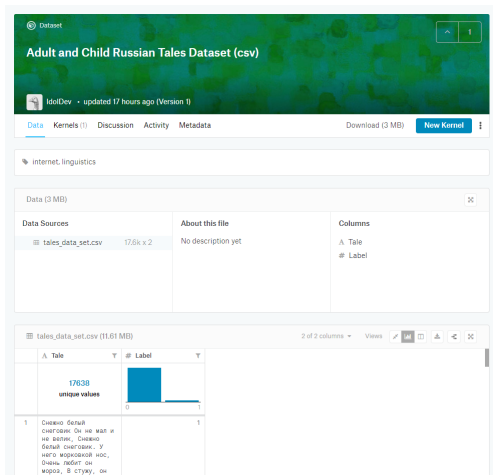


Рис.: Датасет

Как это всё работает

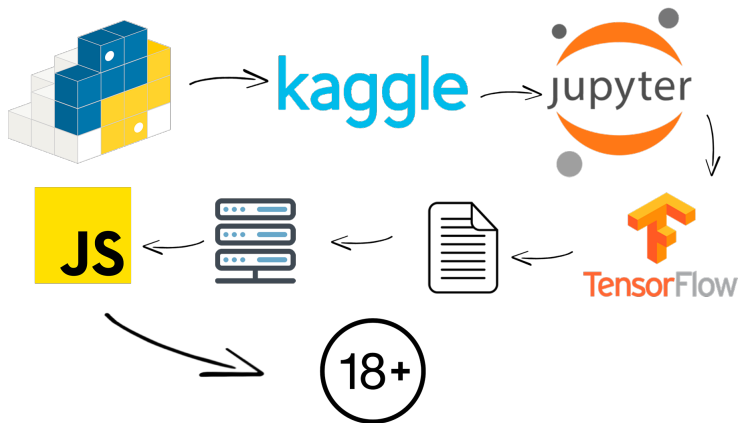


Рис.: Схема проекта

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>