

Классификация текстового контента

Александр Смирнов и Феодор Жилкин

17.05.2019г

Введение

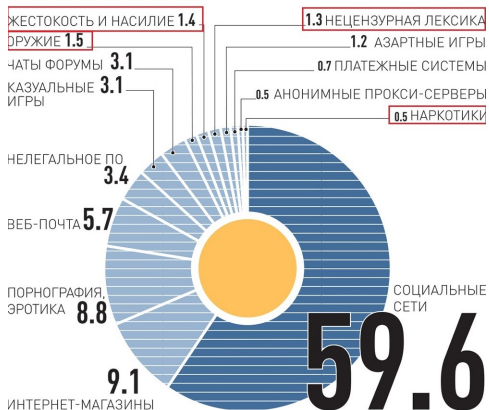


Рис.: Что интересует детей в интернете

Цели

- ▶ Ограничить детей от взрослого текстового контента
- ▶ Получение опыта
 - ▶ Нейросети
 - ▶ Python, Jupyter Notebook, JS
 - ▶ Майнинг датасета и составление csv-файла для загрузки на kaggle
 - ▶ 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 на <https://pypi.org>

Сравнение с аналогами

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

Сравнение с аналогами (2)

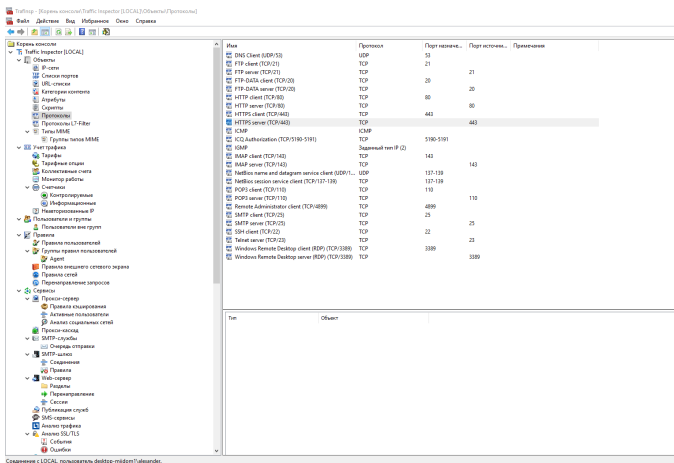
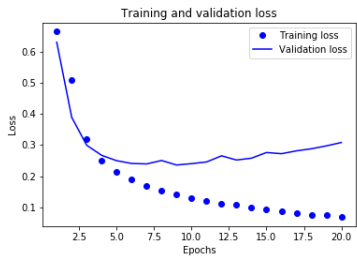
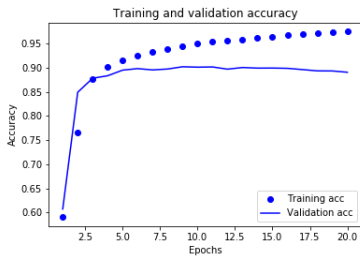


Рис.: Пример интерфейса схожей программы

Результаты

- ▶ Расширение для Chrome
- ▶ Библиотека на pyPi
- ▶ Датасет на kaggle

Результаты обучения



Расширение для Chrome (1)



Рис.: Блокировка контента

Расширение для Chrome (2)

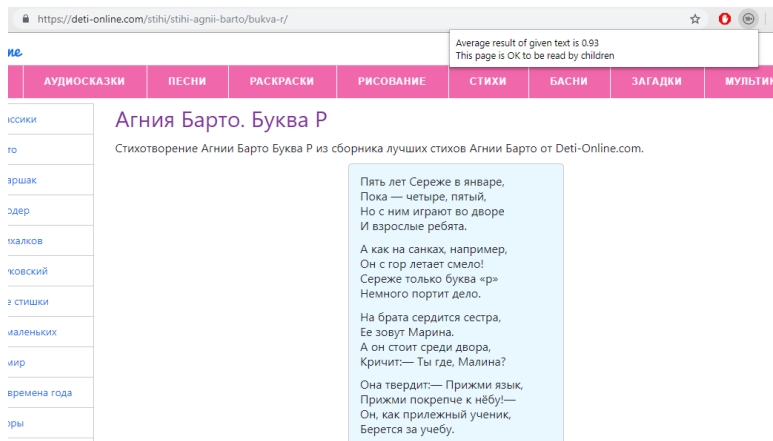


Рис.: Допуск до контента

Библиотека на рурі

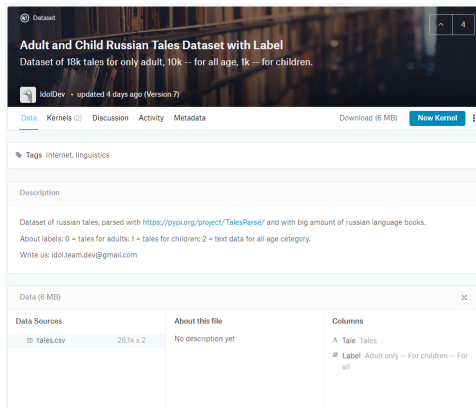
TalesParse 1.0.0

```
pip install TalesParse
```



Рис.: Библиотека

Датасет на kaggle



The screenshot shows the Kaggle dataset page for "Adult and Child Russian Tales Dataset with Label" by IdolDev. The dataset is described as containing 18k tales for adults, 10k for all ages, and 1k for children. It was updated 4 days ago (Version 7). The page includes tabs for Data, Kernels (2), Discussion, Activity, and Metadata. The description mentions the dataset is parsed with <https://pypi.org/project/TalesParse/> and includes a large amount of Russian language books. It also provides labels: 0 for tales for adults, 1 for tales for children, and 2 for text data for all age categories. The contact email is idol.team.dev@gmail.com. The data is 6 MB and consists of a single file named "tales.csv" (28.1k x 2). The columns are "Tale" and "Label".

Adult and Child Russian Tales Dataset with Label
Dataset of 18k tales for only adult, 10k -- for all age, 1k -- for children.

IdolDev • updated 4 days ago (Version 7)

Tags: internet, linguistics

Description

Dataset of russian tales, parsed with <https://pypi.org/project/TalesParse/> and with big amount of russian language books.

About labels: 0 = tales for adults; 1 = tales for children; 2 = text data for all age category.

Write us: idol.team.dev@gmail.com

Data (6 MB)

Data Sources	About this file	Columns
tales.csv 28.1k x 2	No description yet	A Tale Tales # Label Adult only -- For children -- For all

Рис.: Датасет

Как это всё работает

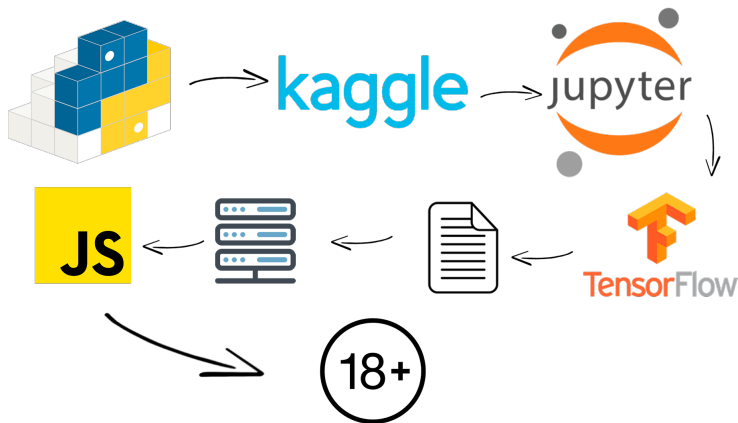


Рис.: Схема проекта

Итоги

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

Результаты

- ▶ Проект – <https://github.com/SmirnovAlexander/PoemClassifier>
- ▶ Парсер – https://github.com/Feodoros/Scrapping_Tales
- ▶ Библиотека – <https://pypi.org/project/TalesParse/>
- ▶ Датасет – <https://www.kaggle.com/idoldev/adult-and-child-russian-tales-dataset-with-label>