

# Классификация текстового контента

Александр Смирнов и Феодор Жилкин

17.05.2019г

# Введение

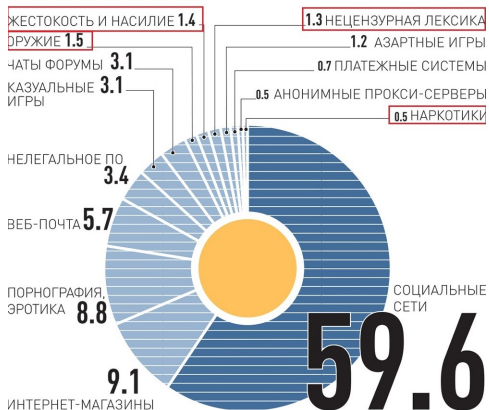


Рис.: Что интересует детей в интернете

# 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)

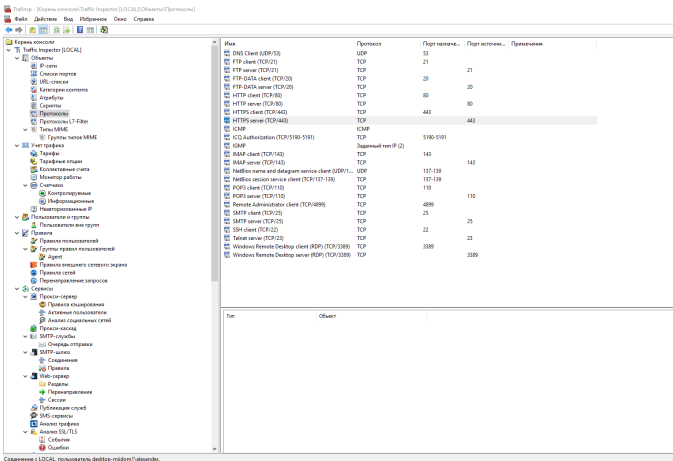


Рис.: Пример интерфейса схожей программы

# Результаты

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

# Анализ подходов

- ▶ Rule-based
- ▶ Machine Learning based
- ▶ Hybrid systems



# Характеристики сравнения эффективности

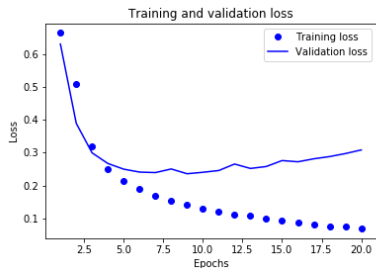
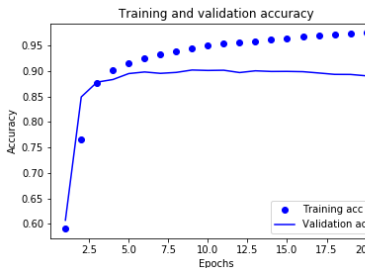
- ▶ Accuracy – общая точность классификатора
- ▶ Recall – отношение заблокированных взрослых сайтов к общему количеству взрослых сайтов (% классифицированных взрослых сайтов)
- ▶ Precision – отношение заблокированных взрослых сайтов к числу всех заблокированных сайтов (точность блокировки)
- ▶ F1 Score - 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)

The screenshot shows a Chrome browser window with the address bar displaying `https://deti-online.com/stihi/stihi-agnii-barto/bukva-r/`. A notification box in the top right corner displays the text: "Average result of given text is 0.93" and "This page is OK to be read by children". Below the address bar is a pink navigation bar with tabs: "АУДИОСКАЗКИ", "ПЕСНИ", "РАСКРАСКИ", "РИСОВАНИЕ", "СТИХИ", "БАСНИ", "ЗАГАДКИ", and "МУЛЬТИКИ". On the left side, there is a vertical menu with categories: "иски", "го", "зршак", "одер", "тхалков", "ковский", "стишки", "маленьких", "мир", "времена года", and "ры". The main content area features the title "Агния Барто. Буква Р" in purple. Below the title is a subtitle: "Стихотворение Агнии Барто Буква Р из сборника лучших стихов Агнии Барто от Deti-Online.com." To the right of the subtitle, 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 небу!—  
Он, как прилежный ученик,  
Берется за учебу.

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 page includes a title, a brief description, tags, a description section, and a data preview table.

**Dataset**

### 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)

[Data](#) [Kernels \(2\)](#) [Discussion](#) [Activity](#) [Metadata](#) [Download \(6 MB\)](#) [New Kernel](#)

**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												
<table border="1"><thead><tr><th>File</th><th>Size</th></tr></thead><tbody><tr><td>tales.csv</td><td>28.1k x 2</td></tr></tbody></table>	File	Size	tales.csv	28.1k x 2	No description yet	<table border="1"><thead><tr><th>Label</th><th>Adult only</th><th>For children</th><th>For all</th></tr></thead><tbody><tr><td>Tale</td><td>Tales</td><td></td><td></td></tr></tbody></table>	Label	Adult only	For children	For all	Tale	Tales		
File	Size													
tales.csv	28.1k x 2													
Label	Adult only	For children	For all											
Tale	Tales													

Рис.: Датасет

# Как это всё работает

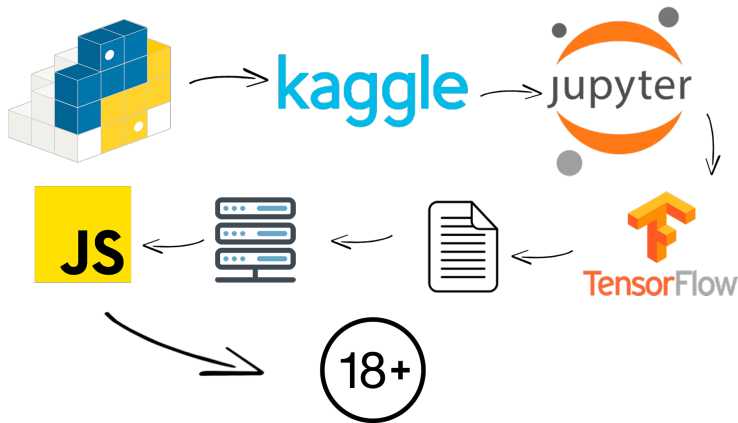


Рис.: Схема проекта



# Итоги

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

# Результаты

- ▶ Проект – <https://github.com/SmirnovAlexander/PoemClassifier>
- ▶ Парсер – [https://github.com/Feodoros/Scrapping\\_Tales](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>