

# Классификация текстового контента

Александр Смирнов и Феодор Жилкин

17.05.2019г

# Введение

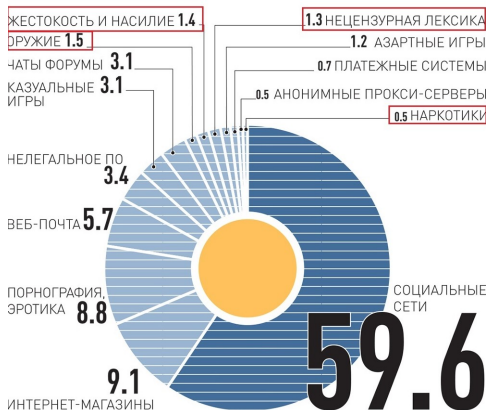


Рис.: Что интересует детей в интернете

# 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

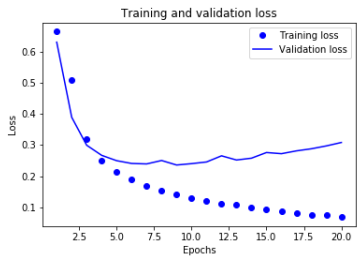
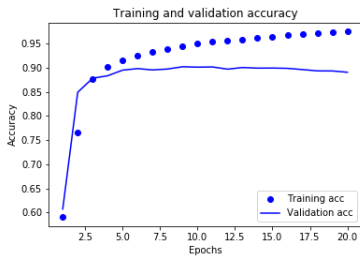
# Сравнение с аналогами

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

# Результаты

- ▶ Расширение для Chrome
- ▶ Библиотека на <https://pypi.org/project/TalesParse/>
- ▶ Датасет на <https://www.kaggle.com/idoldev/adult-and-child-russian-ales-dataset-with-label>

# 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", "дер", "халков", "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 нёбу!—  
Он, как прилежный ученик,  
Берется за учебу.

Рис.: Допуск до контента

# Библиотека на рурі

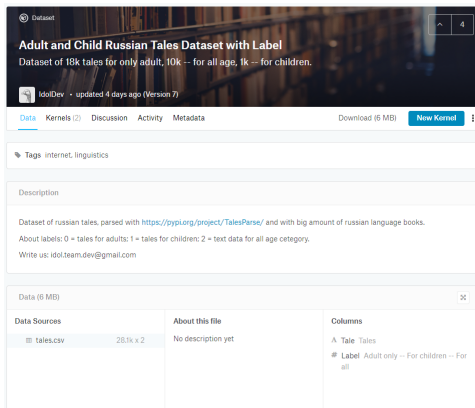
## TalesParse 1.0.0

```
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 header with the dataset title and a brief description: 'Dataset of 18k tales for only adult, 10k -- for all age, 1k -- for children.' Below the header, there are tabs for 'Data', 'Kernels (2)', 'Discussion', 'Activity', and 'Metadata'. The 'Data' tab is selected, showing a table with columns for 'Data Sources', 'About this file', and 'Columns'. The 'Data Sources' column lists 'tales.csv' (28.1k x 2). The 'About this file' column states 'No description yet'. The 'Columns' column lists 'Tale' and 'Label' (Adult only -- For children -- For all).

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)

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](mailto:idol.team.dev@gmail.com)

Data (6 MB)

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

Рис.: Датасет

# Как это всё работает

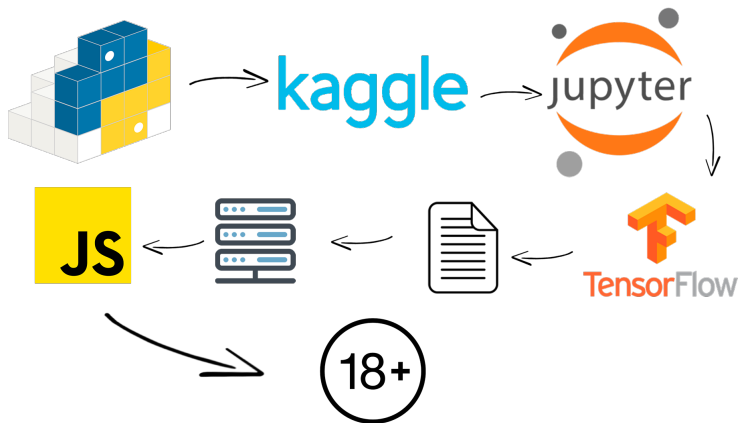


Рис.: Схема проекта

# Итоги

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

# Результаты

- ▶ Проект – <https://github.com/SmirnovAlexander/PoemClassifier>
- ▶ Парсер – [https://github.com/Feodoros/Scrapping\\_Tales](https://github.com/Feodoros/Scrapping_Tales)
- ▶ Библиотека – <https://pypi.org/project/TalesParse/>
- ▶ Датасет – <https://www.kaggle.com/idoldev/adult-and-child-russian-ales-dataset-with-label>