Санкт-Петербургский государственный университет

Направление Математическое обеспечение и администрирование информационных систем

Жилкин Фёдор Игоревич

Классификация текстового контента

Курсовая работа

Научный руководитель:
к. т. н., доц. Литвинов Ю. В.

Санкт-Петербург
2019

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems

Fedor Zhilkin

Classification of text content

Course Work

Scientific supervisor:
Associate Professor Yurii Litvinov

Saint-Petersburg
2019

Оглавление

Введение

Зачастую, находясь в интернете, можно наткнуться на контент, который был бы нежелателен к просмотру детьми. Обратимся к (1).

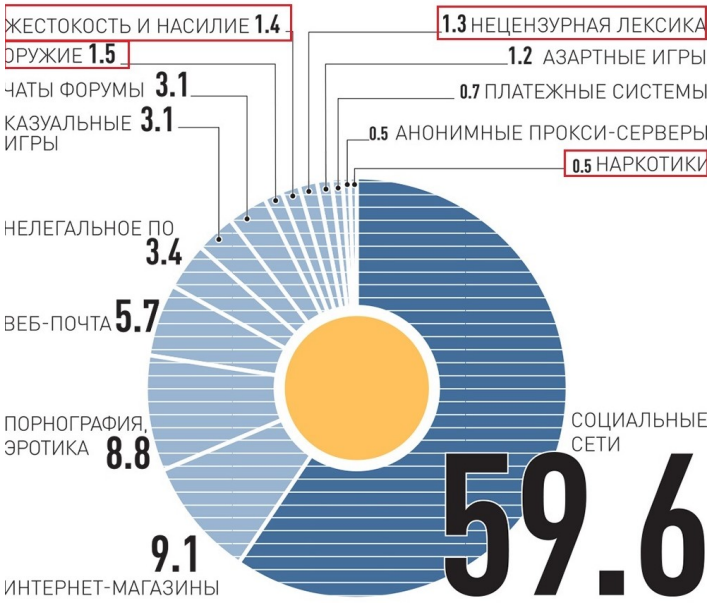

Рис. 1: Что интересует детей в интернете (Ист. – Лаборатория Касперского)

Можно видеть, что довольно большой процент потребляемой детьми информации относится к нежелательному контенту. Моя задача состоит в том, чтобы ограничить детей от подобного содержимого сайтов.

Данная работа будет о том, как найти и собрать примеры взрослого текстового контента и контента, подходящего для чтения детьми.

Цели и задачи

После исследования предметной области были поставлены следующие цели и задачи.

Цели

- Собрать данные для обучения.
- Написать Python-библиотеку, собирающую рассказы с сайтов по определённому фильтру.

Задачи

- Провести анализ возможных источников.
- Собрать рассказы для взрослых и обычные рассказы.
- Ознакомиться с областью:
  - Пройти курс по датамайнингу [1]
  - Чтение литературы [2]

Реализация данных задач позволит подготовить данные для обучения модели, которая впоследствие будет фильтровать сайты.

1. Основные понятия

Для прочтения данной работы требуются знания предметной области, поэтому введем некоторые понятия и определения.

- Датасет — набор данных
- Библиотека классов определяет типы и методы, которые могут быть вызваны из любого приложения

Характеристики сравнения качества материала.
Необходимо собрать тексты двух основных категорий: нейтральные и негативные. Негативное текстовое содержание страницы крайне опасно для просмотра в юном возрасте. Нейтральная информация не влечет плохих последствий. Таким образом, будем оценивать качество негативного контента по следующим критериям:

- Негативный контент:
  - Наличие сцен насилия
  - Наличие описания оружия, наркотиков
  - Наличие ненормативной лексики

  К нейтральному контенту будем относить всё остальное.

Подходы к поиску материала.
Сбор текстов из:

- Книги
- Журналы
- Пособия
- Детская литература
- Сайт ideer.ru

2. Обзор существующих решений

На данный момент существуют многочисленные хранилища текстов на английском языке, что не подходит для решения задачи. Датасетов рассказов на русском языке крайне мало, все они однотипны и очень скудны по своему содержанию.

Готовые решения:

- Корпус коротких текстов на русском языке, односторонний по содержанию, содержит короткие тексты на русском языке
- Большой по содержанию и тематикам датасет, но на английском языке, не подходит для решения задачи.

3. Описание предлагаемого решения

Необходимо собрать объемный, широкий по тематикам датасет на русском языке, поэтому было принято решение скачивать контент с различных онлайн библиотек, сайтов, журналов и сервиса Подслушано. В конечном итоге было собрано более 28000 текстов на русском языке на различную тематику.

Заключение

В ходе данной работы были полностью выполнены поставленные задачи.

- Сделана библиотека на pypi
- Собраны рассказы на kaggle
- Написан сборщик рассказов

Список литературы

[1] Intuit. Course Data Mining // web.iitd.ac.in. — 2017. — URL: https://www.intuit.ru/studies/courses/6/6/info (online; accessed: 10.06.2019).

[2] Wikipedia. Data Mining // Википедия, свободная энциклопедия. — 2011. — URL: https://ru.wikipedia.org/wiki/Data_mining (дата обращения: 05.06.2019).