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).

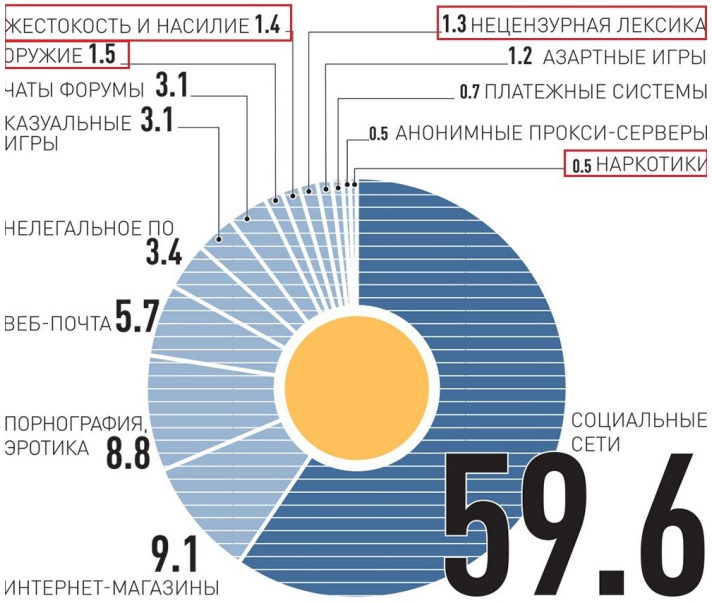

Рис. 1: Что интересует детей в интернете (Ист. – Лаборатория Касперского)

Можно видеть, что довольно большой процент потребляемой детьми информации относится к нежелательному контенту. Была поставлена задача – ограничить детей от подобного содержимого сайтов.

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

Цели и задачи

После исследования предметной области и поставленной проблемы были поставлены следующие цели и задачи.

- Цели:
  1. Собрать данные для обучения,
  2. Написать Python-библиотеку, собирающую рассказы с сайтов по определённому фильтру.
- Задачи:

1. Провести анализ возможных источников,
2. Собрать рассказы для взрослых и обычные рассказы,
3. Ознакомиться с областью:
   (a) пройти курс по датамайнингу [1],
   (b) 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.
- Парсер — это программное обеспечение для сбора данных и преобразования их в структурированный формат, чаще всего работа с текстовым типом информации.

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:

1. Наличие сцен насилия,
2. Наличие описания оружия, наркотиков;
3. Наличие ненормативной лексики.

К нейтральному контенту будем относить всё остальное.

Подходы к поиску материала.
Сбор текстов из:

1. Книги,
2. Журналы,
3. Пособия,
4. Детская литература,
5. Подслушано [1].

[1] Домашняя страница сервиса Подслушано, URL: https://ideer.ru/ (дата обращения: 09.10.2019)

2. Обзор существующих решений

На данный момент существуют многочисленные хранилища текстов на английском языке, что не подходит для решения задачи. Датасетов текстов на русском языке крайне мало, все они однотипны и очень скудны по своему содержанию. Большинство готовых решений представляют собой помошника для построения собственного парсера – обработка запросов, получение не обработанной HTML-разметки, избежание запретов IP и CAPTCHA [2]. Но также существует несколько инструментов, полностью решающих нашу задачу, однако воспользоваться ими можно только на коммерческой основе.

Готовые решения:

1. Датасеты:
   (a) Корпус коротких текстов на русском языке [3], односторонний по содержанию, содержит короткие тексты на русском языке
   (b) Датасет коротких предложений и словосочетаний на русском языке [4], достаточно большой датасет, но предложенные материалы слишком короткие и однотипны по содержанию, отлично подходит для тренировки чат-бота.
   (c) Большие по содержанию и тематикам датасеты «Text Classification - ChatBot»[5], «chatbot» [6] и «Building a chatbot»[7], но все они на английском языке, не подходят для решения задачи.
2. Реализованные решения по сбору текста:
   (a) ScrapingHub [8] позволяет собирать любую информацию с сайта. Хорошее решение, но собирает информацию только с одной страницы сайта. Так как наша цель – многостраничные сайты, необходимо использовать совместно с другим ПО, ко-

[2]Completely Automated Public Turing test to tell Computers and Humans Apart — полностью автоматизированный публичный тест Тьюринга для различения компьютеров и людей

[3]Раздел «Скачать корпус» на домашней странице сайта датасета, URL: https://study.mokoron.com/ (дата обращения: 09.10.2019)

[4]URL: https://github.com/Koziev/ (дата обращения: 09.10.2019)

[5]URL: https://www.kaggle.com/rahulvks/text-classification-chatbot (дата обращения: 09.10.2019)

[6]URL: https://www.kaggle.com/justdvnsh/chatbot/data (дата обращения: 09.10.2019)

[7]URL: https://www.kaggle.com/melkmanszoon/building-a-chatbot (дата обращения: 09.10.2019)

[8]Домашняя страница инструмента ScrapingHub, URL: https://scrapinghub.com/ (дата обращения: 09.10.2019)

торое будет поочередно передавать парсеру нужные страницы.

(b) Octoparse [9] и ParseHub [10] – ПО, позволяющее собирать необходимую информацию с сайтов. Удобный интерфейс позволяет пользователю выбрать необходимые блоки страницы и собрать оттуда текст. Доступна бесплатная пробная версия на несколько использований, с ограничением количества запросов.

[9] Домашняя страница инструмента Octoparse, URL: https://www.octoparse.com/ (дата обращения: 09.10.2019)

[10] Домашняя страница инструмента ParseHub, URL: https://www.parsehub.com/ (дата обращения: 09.10.2019)

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 [11]. Так как необходим большой объем информации, был написан сборщик текста с разного рода интернет-страниц – программа выполняет чтение HTML-разметки сайта, фильтрует блоки и сохраняет содержимое. Далее происходит очистка текста (удаление лишних HTML-тэгов, ненужных символов) и разбиение его на отдельные куски (100-128 слов), из которых будет составляться конечный датасет. После этого переходит на новую страницу и действия повторяются. Все тексты (один рассказ берем как один текст) с сервиса Подслушано из категорий, непристойных для чтения детьми, собираем и отмечаем пометкой «1» (к каждому отдельному тексту через запятую дописываем его пометку, иначе говоря, создаем табличку из двух столбцов, где каждая строчка в первом столбце – текст, во втором столбце – соответствующая ему пометка, такой формат файлов называется SCV[12]), собранные тексты с детских сайтов – пометкой «0». Поскольку разметка сайтов разная, то пришлось выбрать 2 основных сайта, с которых будем брать текстовую информацию и, соответственно, писать две различные процедуры чтения HTML-разметки страниц (одна процедура для сервиса Подслушано, вторая – для сайта детских стихов и рассказов). Также необходимо было создать программу, которая умела бы скачивать книги с онлайн-библиотек и разбивать их на отдельные тексты для того, чтобы финальный датасет был максимально широкий по своему содержанию. С онлайн-библиотек выкачивались книги с ограничением «16+» и обучающая литература, где точно не будет чего-либо непристойного, поэтому тексты, собранные с этих ресурсов отмечаем тоже пометкой «0». Книги с онлайн библиотек разбивались на куски примерно 100-128 слов. Далее был создан словарь непристойных слов и варжений для даль-

[11] URL: https://ideer.ru/ (дата обращения: 09.10.2019)

[12] Comma-Separated Values – текстовый формат, предназначенный для представления табличных данных

нейшей блокировки сайтов, где встречается слово из этого словаря. В конечном итоге было собрано более 28000 текстов на русском языке на различную тематику, которые доступны в открытом доступе на сайте https://www.kaggle.com и размеченных на «плохие» и «хорошие». Для того, чтобы решением пользовались другие разработчики, было принято решение выложить датасет на kaggle [13] и библиотеку на pypi [14]. С помощью готового датасета разработчики смогут решать похожие задачи, а с помощью библиотеки они смогут собирать любую текстовую информацию с сервиса Подслушано и сайта детской литературы[15].

Библиотека классов TalesParse.

Библиотека имеет 2 класса: Helper и Scraper. Использует процедуры встроенной библиотеки urllib для скачивания HTML-кода страницы, и библиотеки BeautifulSoup[16] для фильтрации блоков исходного кода страницы. В классе Helper реализованы основные методы для сбора информации: процедуры по чтению, фильтрации и очистки HTML-кода, а так же записи нужных текстов в файл. В классе Scraper с помощью вызова метода get_good_tales(x), где x – количество текстов, можем получить нужно нам количество текстов пригодных для детского чтения. С помощью вызова процедуры get_bad_categories() можем посмотреть с каких категорий сервиса Подслушано мы будем брать тексты, и выбрать конкретные.

С помощью метода get_bad_tales(bad_categories, x), где bad_categories – список категорий, x – количество текстов, можем получить нужное нам количество непристойных текстов.

[13]URL: https://www.kaggle.com (дата обращения: 09.10.2019)

[14]URL: https://pypi.org/project/TalesParse/ (дата обращения: 09.10.2019)

[15]URL: https://deti-online.com/ (дата обращения: 09.10.2019)

[16]URL: https://www.crummy.com/software/BeautifulSoup/ (дата обращения: 09.10.2019)

Пример использования библиотеки.

```
from TalesParse import Scraper as sc
def main():
    categories = sc.get_bad_categories()
    bad = sc.get_bad_tales(categories, 184)
    good = sc.get_good_tales(55)
    print(bad + good)
```

Заключение

В ходе данной работы были полностью выполнены поставленные задачи:

1. Проанализированы возможные пути решения задачи и готовые решения
2. Произведено ознакомление с областью датамайнинга
3. Сделана библиотека на pypi [17]
4. Собраны рассказы на kaggle [18]
5. Написан сборщик рассказов [19]

[17]URL: https://pypi.org/project/TalesParse/ (дата обращения: 09.10.2019)

[18]URL: https://www.kaggle.com/idoldev/adult-and-child-russian-tales-dataset-with-label (дата обращения: 09.10.2019)

[19]URL: https://github.com/Feodoros/Scraping_Tales (дата обращения: 09.10.2019)

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).