Учреждение образования

«Белорусский государственный университет информатики и радиоэлектроники»

Кафедра интеллектуальных информационных технологий

ОТЧЁТ

по лабораторной работе №2 по дисциплине «ЕЯзИИС»

на тему: «Методы автоматического распознавания языка текстового документа»

Выполнили студенты группы 821701: Поживилко П.С.
Витушко Л. Д.

Проверил: Крапивин Ю.Б.

Минск, 2021

Цель работы: изучить и отработать практические навыки применения методов автоматического распознавания языка текстовых документов.

Задание
Вариант 1.
Язык текста: русский, английский.
Формат документа: html.
Реализуемый метод: N-грамм, алфавитный, нейросетевой.

Ход работы.

Описание системы.

Система реализована на языке Python 3. Графический интерфейс системы создан с помощью - PyQt. Система может принимать тексты в 3-х режимах: чтение директорий, чтение одиночных файлов, чтение одиночных страниц из сети Интернет. Определять языки текстов система может 3-мя способами: алфавитным, N-граммным и нейросетевым. Присутствует возможность сохранения результатов работы в текстовый файл. Архитектура приложения представляет собой агрегирующий класс Application, который инкапсулирует в себе логику взаимодействия с пользовательским интерфейсом и перевод текста.

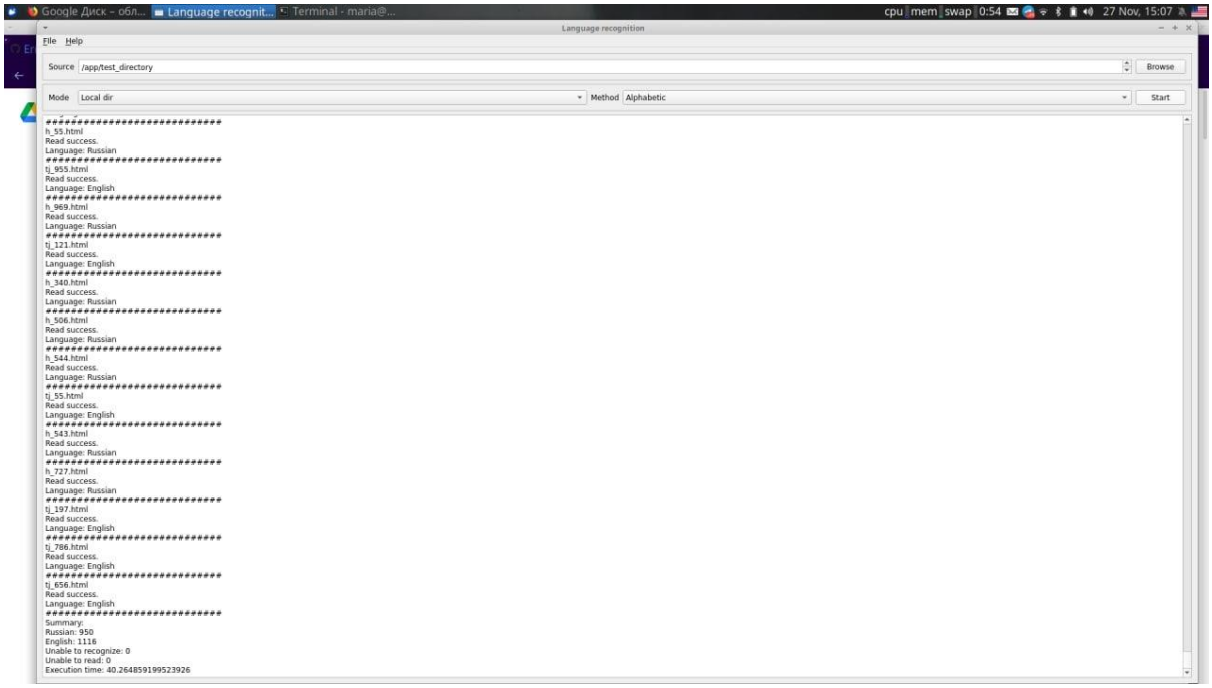


Рисунок 1. Интерфейс системы

Тестовый набор документов:

Тестовый набор документов включает в себя 2 тысячи html-страниц, из которых 1000 статей из хабрахабр.

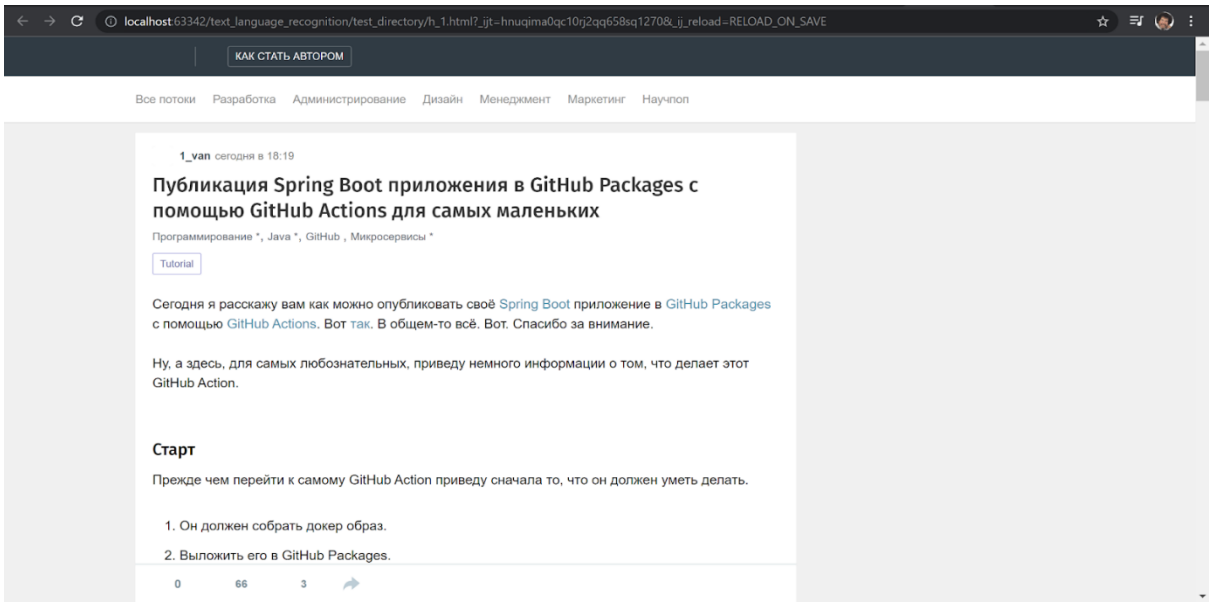
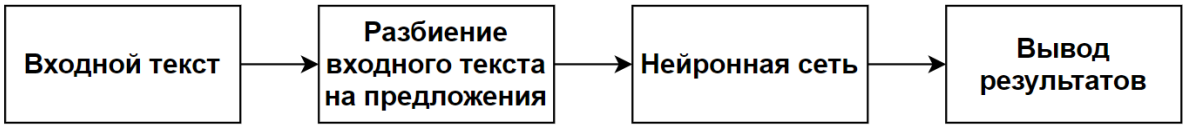


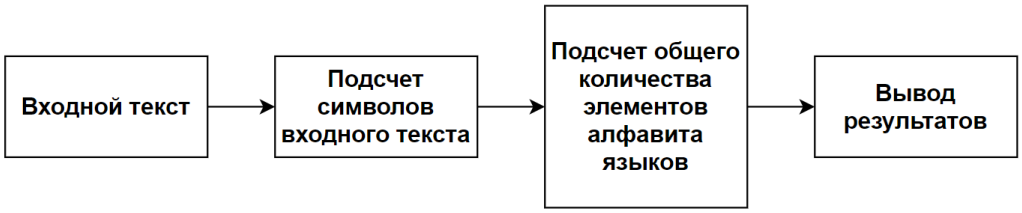
Рисунок 2. Пример страницы с сайта хабр

Схемы определения языка:

Нейросетевой метод:



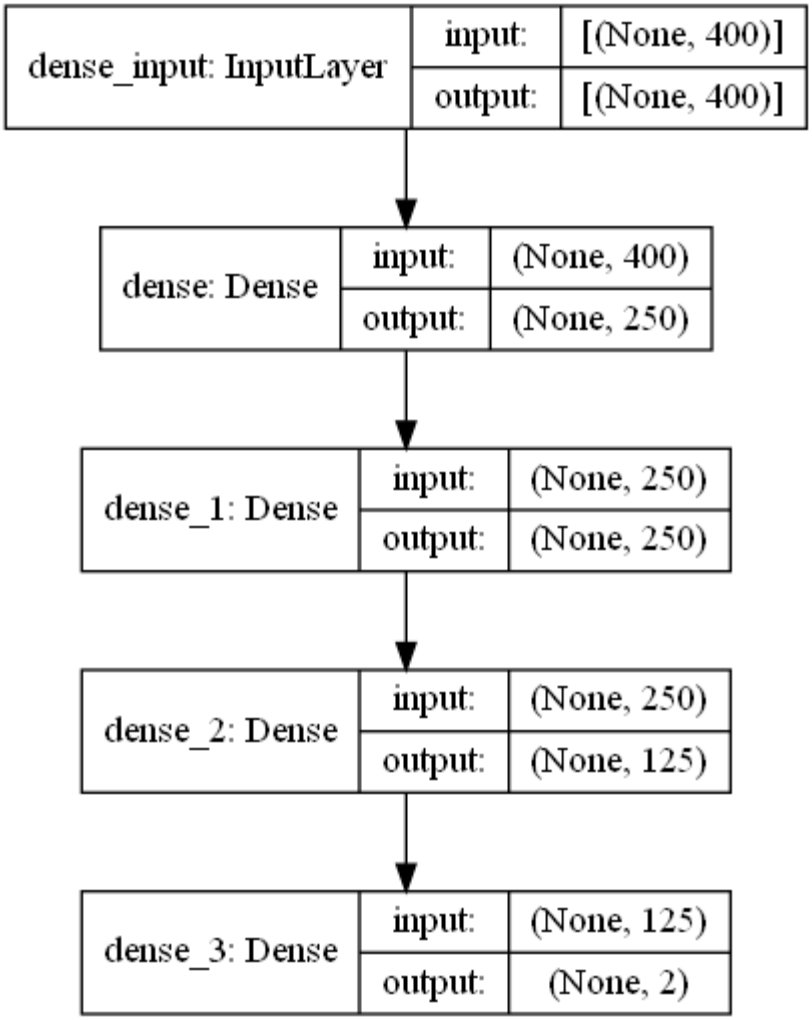
Алфавитный метод:



Метод N-грамм:

Нейронная сеть:

С помощью фреймворка Keras была составлена и обучена модель нейронной сети со следующей архитектурой:



Обучающая выборка включает в себя ~2.000.000 предложений на русском и английском языке. Тестовая выборка - ~500.000 предложений. Модель обучалась в течение 2х эпох, в качестве оптимизатора был выбран оптимизатор Адама, в качестве функции потерь - перекрестная энтропия, в качестве метрики выступала точность.

В результате обучения на тестовом множестве была достигнута точность 99.94%.

Результаты тестирования:

Метод N-грамм:

```
Summary:
Russian: 968
English: 1098
Unable to recognize: 0
Unable to read: 0
Execution time: 132.47048330307007
```

Нейросетевой метод:

```
Summary:
Russian: 966
English: 1096
Unable to recognize: 4
Unable to read: 0
Execution time: 639.2598187923431
```

Алфавитный метод:

```
Summary:
Russian: 950
English: 1116
Unable to recognize: 0
Unable to read: 0
Execution time: 81.65975713729858
```

Исходя из предположения, что все статьи с форума habr - русскоязычные, а остальные - англоязычные, получаем точность метода N-грамм - 99.61%, точность алфавитного метода - 98.74%, точность нейросетевого метода - 99.51%. Самым точным методом на данной выборке оказался метод N-грамм, самым быстрым - алфавитный.

Вывод:

В ходе выполнения лабораторной работы были изучены и применены на практике различные методы определения языка текста. В итоге самым точным оказался метод n-грамм, а самым быстрым - алфавитный. В нашем случае - когда языками являются совершенно разные по алфавиту русский и английский, наиболее эффективным по времени является алфавитный метод, а наименее эффективным - метод нейросетей. В то же время, если нужно будет классифицировать тексты на схожих по алфавиту языках, то алфавитный метод значительно уступит нейросетевому и n-граммному в точности.