Учреждение образования

“БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ ”

Кафедра интеллектуальный информационных технологий

Отчет по лабораторной работе №2

по курсу: “Естественно-языковой интерфейс интеллектуальных систем”

Выполнил студент группы 921702: Баранов А.Д.

Проверил: Крапивин Ю.Б.

Минск

2022

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

**Вариант 8:** реализовать распознавание испанского и немецкого языка методами N-грамм, алфавитным, нейросетевым.

**Требования к разрабатываемой системе:**

* на входе – текстовые документы одинакового размера (например, 1 страница формата А4), содержащие тексты на естественных языках согласно варианту;
* на выходе – активная ссылка на документ, и результат идентификации отдельного текста – язык текста; сводная статистика по всем текстам из тестовой коллекции.
* наличие средств сохранения в файл и распечатки полученной на выходе информации;
* интерфейс системы - доступный для пользователей любого уровня, содержащий help-средства работы с программой.

**Теоретические сведения**

**Метод N-грамм**

В основе этого метода лежит Закон Зипфа – эмпирическая закономерность распределения частоты слов естественного языка: если все слова языка (или просто достаточно длинного текста) упорядочить по убыванию частоты их использования, то частота *n*-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру *n* (так называемому *рангу* этого слова). Например, первое по используемости слово встречается примерно в два раза чаще, чем второе, и в три раза чаще, чем третье.

Алгоритм метода заключается в нахождении частот N-грамм для всех тестовых документов, для которых известен язык, а также для каждого документа, язык которого пытаемся определить. После этого среди всех тестовых документов находим тот, для которого расстояние от его N-граммной статистики до статистики тестируемого документа минимально. После этого языком тестируемого документа считается язык найденного тестового документа.

Расстояние между статистиками подсчитывается следующим образом: все N-граммы сортируются в порядке убывания частоты их появления, затем для каждой N-граммы вычисляется разница её позиций в отсортированном списке N-грамм тестового и тестируемого документов. Расстояние между статистиками определяется как сумма разниц позиций каждой N-граммы. Если N-грамма (например, “ED” на рис.1) отсутствует в профиле категории (языка) то ей назначается максимальная величина оценки несовпадения позиций N-грамм.

**Алфавитный метод**

Данный метод позволяет определить язык текста, основываясь на алфавите известного языка. Определение языка сводится к сравнению известного алфавита и уникальных символов текста.

**Нейросетевой метод**

Нейросетевой метод можно разбить на следующие подзадачи:

1. Поиск размеченного датасета с текстами на необходимых языках с явным указанием языка
2. Базовая реализация одного из видов нейронной сети и подготовка к обучению
3. Обучение нейронной сети на подготовленном датасете
4. Тестирование нейронной сети тестовыми данными и оценка точности распознавания

**Реализация**

Система состоит из двух частей: серверной и клиентской. В серверной части находится код на языке Python, который реализует выше перечисленные методы.

**Пользовательский интерфейс:**

Приложения написано на языке TypeScript с использованием фреймворка Angular.

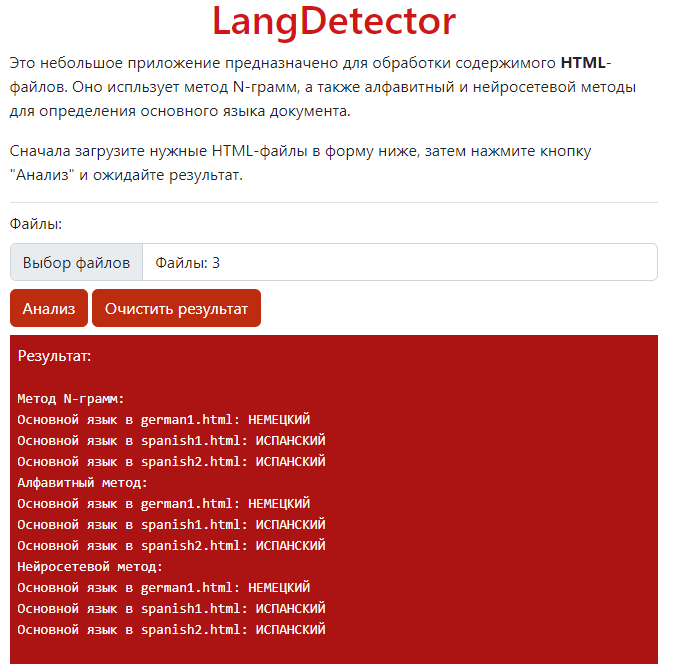


Рис. 2. Пользовательский интерфейс

Пользователь может выбрать несколько файлов для обработки.

Главный компонент приложения - MainComponent. Именно здесь выбранные пользователем файлы отправляются на дальнейшую обработку.



Рис. 3. Главный компонент

**Обработка текста:**

Функция, представленная на рисунке, очищает текст от лишних символов, оставляя только символы алфавита и пробелы.

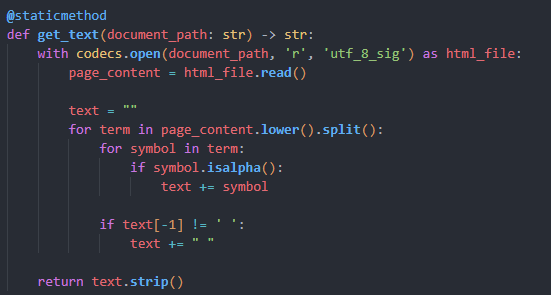


Рис. 4. Функция очистки текста.

Следующая функция, нормализует все полученные слова.

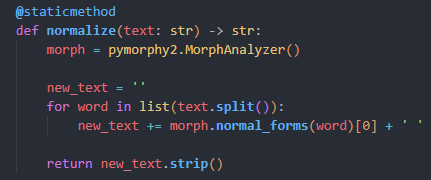


Рис. 5. Функция нормализации слов.

**Реализация метода N-грамм:**

Функция, представленная на рисунке 7, формирует словарь грамм из очищенного текста. Если длина слова является четное, получатся все граммы по 2 символа, иначе по одна грамма будет длины 3.

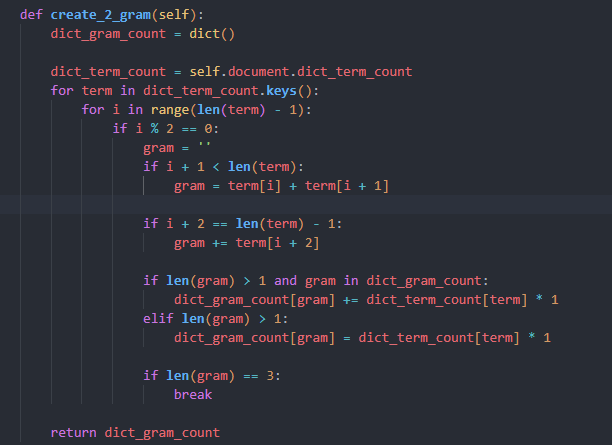


Рис. 6. Функция создания словаря грамм.

Функция, представленная на рисунке 8, сортирует словарь грамм, для точного нахождения расстояния между граммами поискового запроса и документа.

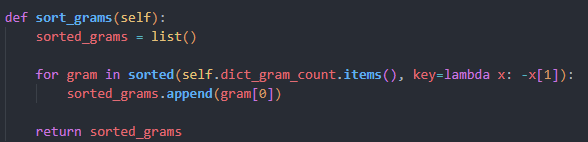


Рис. 7. Функция сортировки словаря грамм.

Ниже представленная функция, находит расстояние между документом из обучающей выборки и текстовым запросом.

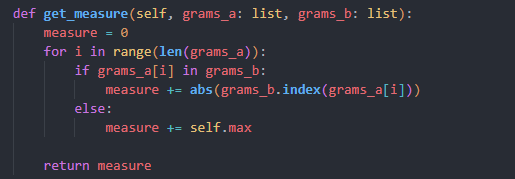


Рис. 8. Функция нахождения расстояния.

Далее определяется язык текстового запроса.

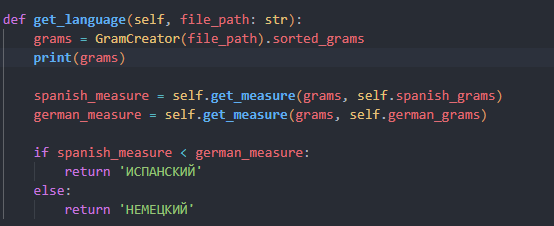


Рис. 9. Функция определения языка текстового запроса.

**Реализация алфавитного метода:**

Функция, представленная на рисунке 11, получает список всех символов.

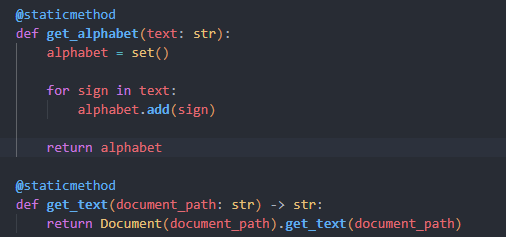


Рис. 10. Функция получения буквенных символов.

Далее, аналогично с предыдущим методом, определяет язык текстового запроса.

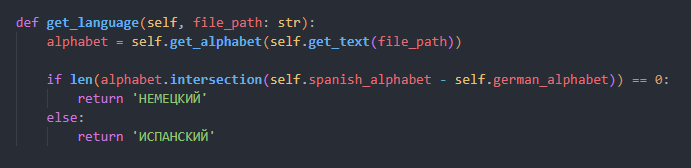


Рис. 11. Функция определения языка текстового запроса.

**Реализация нейросетевого метода:**

Для реализации нейросетевого метода была использована уже реализованная нейросеть из открытой библиотеки **spacy**.

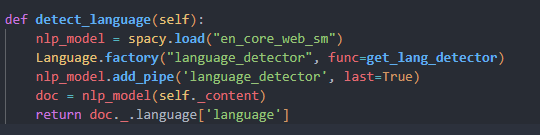


Рис. 12. Фрагмент реализации нейросетевого метода.

**Вывод:**

В ходе работы были изучены и отработаны практические навыки применения методов автоматического распознавания языка текстовых документов. Были реализованы методы распознавания языков с помощью средств языка Python, был разобран принцип работы нейронной сети для нейросетевого метода, были разобраны методы обучения нейронной сети на распознавание текстовых данных.