Министерство образования Республики Беларусь

Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И УПРАВЛЕНИЯ

Кафедра интеллектуальных информационных технологий

Отчёт по лабораторной работе №2
по курсу «ЕЯзИИС» на тему:
«Методы автоматического распознавания языка текстового документа»

Выполнили студенты группы 921701: Василевский Артемий Дмитриевич
Драгун Владимир Андреевич
Орлов Максим Константинович
Пилипейко Валентин Игоревич

Проверил: Крапивин Юрий Борисович

Минск 2022

Содержание

1. Цель работы и вариант

Изучить и отработать практические навыки применения методов автоматического распознавания языка текстовых документов.

Вариант 7: Распознавание Французского и Английского языков методами N-грамм, алфавитным, нейросетевым. Формат файла, который содержит текст — HTML.

2. Информация о текстовой коллекции документов

Были использованы тексты из газет и классических произведений на соответствующем языке. После этого тексты были оформлены в виде html-документов и названы именем языка, который содержится в документе. На этом подготовка текстов была завершена.

В проекте можно найти следующие файлы в папке docs:

- english.html — содержит тренировочный текст на английском языке;
- french.html — содержит тренировочный текст на французском языке;

3. Описание системы, данных и алгоритмов

3.1. Описание структуры системы

Система состоит из модулей для работы с данными и для определения языка, на котором написан текст.

Подробнее о каждом модуле:

- main.py — содержит информацию о путях к файлам с текстами, а также ответственен за вывод информации пользователю
- GramsMethod.py — содержит реализацию алгоритма для определения языка текста с помощью метода N-грамм
- AlphabetMethod.py — содержит реализацию алгоритма для определения языка текста с помощью Алфавитного метода
- NeuralMethod.py — содержит реализацию алгоритма для определения языка текста с помощью Нейросетевого метода.

3.2. Описание типов данных

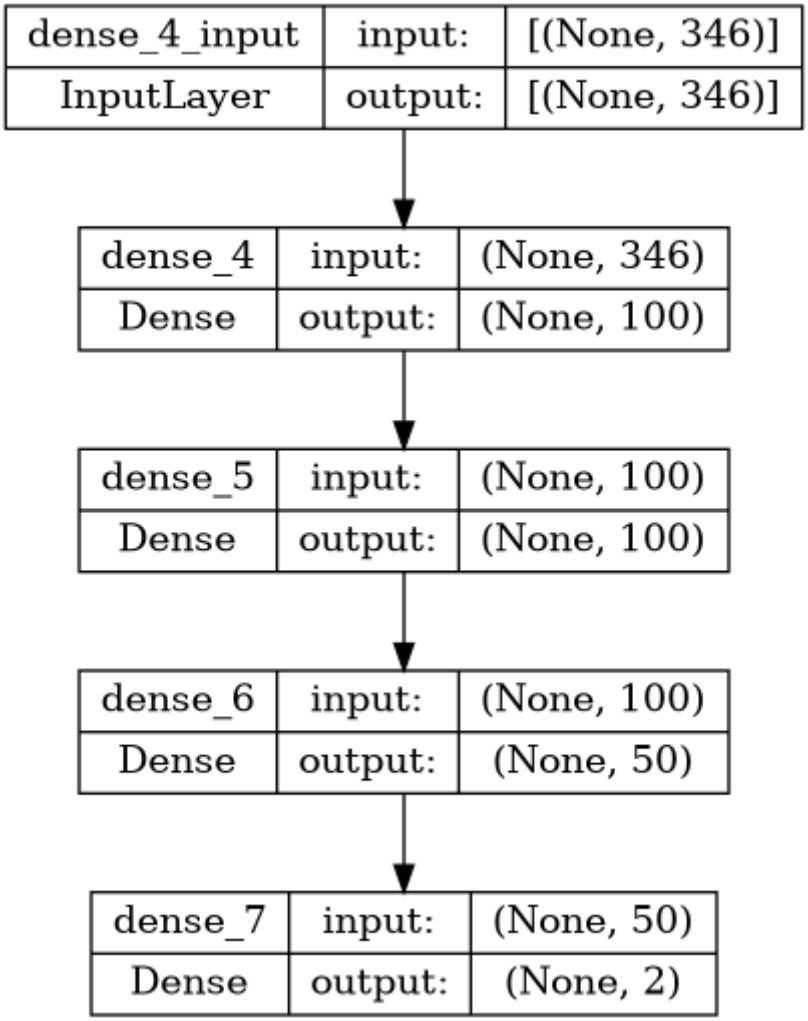
В данной системе мы использовали такие типы данных как строка, массив, ассоциативный массив.

3.3. Описание алгоритмов

Алгоритм N-грамм работает следующим образом. На первом этапе текст разбивается на n-граммы. N-грамма это подстрока размера N. После чего n-граммы собираются в ассоциативный массив, где ключом является n-грамма, а значением — количество встреченных таких n-грамм в тексте (то есть рейтинг). Затем данный массив сортируется по рейтингу. В итоге получается массив, где меньшему индексу соответствует более часто встречаемая n-грамма. Для определения того, в какой степени тестовый текст близок к языку, необходимо вычислить расстояние. Расстояние вычисляется как сумма индексов встреченных в тексте n-грамм. То есть, чем больше в тексте «популярных» для данного языка n-грамм, тем меньше будет расстояние, так как у n-грамм будет меньший индекс. Если n-грамма вообще не встречена в массиве, то расстоянию прибавляется большое значение (константа, порядок которой сопоставим с порядком размера массива n-грамм). В итоге, чем меньше расстояние от нашего текста до обучающей выборки, тем ближе язык.

Алфавитный алгоритм работает на том факте, что у разных языков зачастую разные алфавиты (или алфавиты с наличием некоторых специфических букв, которые есть только в данном языке).

Нейросетевой метод состоит в построении обучаемой модели, на входе которой могут быть слова или предложения, а на выходе степень близости к какому-либо языку (у нейросети столько выходов, сколько языков может быть распознано). На подготовительном этапе текст разбивается на триплеты. Все триплеты собираются в массив, дубликаты при этом удаляются. Этот массив триплетов подаётся на вход нейросети (с предварительной векторизацией). Для реализации была выбрана сеть с 5-ю слоями (один из слоёв является входным) и прямым распространением (сеть данного типа также называется DFF). Все слои состоят из нейронов, которые имеют полную связность с нейронами предыдущего и следующего слоёв. Функция активации всех слоёв, кроме последнего — relu. Функция активации последнего слоя — softmax. Первый слой содержит столько нейронов, сколько было найдено триплетов (значение может отличаться для разных выборок). Последний слой содержит 2 нейрона, так как по условию необходимо выбрать один из двух языков. Подробнее архитектуру нейросети можно рассмотреть на изображении.



3.4. Оценка быстродействия

Для оценки быстродействия было измерено время выполнение каждого метода. По итогам измерения методы показали схожий результат.

Конкретные значения:

1. Метод N-грамм — 0.06 секунд
2. Алфавитный метод — 0.054 секунды
3. Нейросетевой метод — 0.219 секунды

Как можно заметить, нейросетевой метод в 4 раза медленнее, чем остальные, однако время сопоставимо и неощутимо для пользователя. Поэтому можно сказать, что все алгоритмы имеют схожее время работы. Однако, для работы нейросетевого метода необходимо время на инициализацию программы, которое вносит ощутимый вклад во время запуска программы.

3.5. Результат тестирования системы

Пример взаимодействия с программой выглядит следующим образом

```
1. Grams method
2. Alphabet method
3. Neural method
4. Help
5. Exit

4
You are using the language detection system, please press one of the selected numbers to continue

1. Grams method
2. Alphabet method
3. Neural method
4. Help
5. Exit

3
1. Select file
2. Save results
3. Choose other method

1
/home/artrayme/PycharmProjects/EYAZIS/l2/dataset/english.html
1/1 [==============================] - 0s 87ms/step

file:///home/artrayme/PycharmProjects/EYAZIS/l2/dataset/english.html -- English

--- 0.23717355728149414 seconds ---
1. Select file
2. Save results
3. Choose other method
```

4. Использование библиотек

Для выполнения лабораторной работы использовались следующие библиотеки:

1. pandas
2. numpy
3. sklearn
4. tensorflow
5. keras
6. os
7. codecs
8. pymorphy2

Большая часть из представленных библиотек использовалась для нейросетевого метода, что является минусом нейросетевого метода, так как вынуждает создавать программу бОльшего размера, чем необходимо.

5. Вывод

В данной лабораторной работе была реализована система, которая позволяет определить язык текста одним из трёх алгоритмов: Алгоритм N-грамм, Алфавитный алгоритм, Нейросетевой алгоритм. После реализации система, были произведены тесты системы, по которым был сделан вывод, что система обладает достаточной точностью в сценариях работы, которые поставлены методическими указаниями. Наибольший интерес представляет алгоритм N-грамм и Алфавитный алгоритм из-за их высокой эффективности и скорости работы. Нейросетей алгоритм оказался не лучшим решением для подобного класса задач из-за большого потребления ресурсов ЭВМ, а также большим временем запуска из-за необходимости инициализации библиотек.