Министерство образования и науки Российской федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

Кубанский государственный технологический университет

(ФГБОУ ВО «КубГТУ»)

Институт компьютерных систем и информационной безопасности

Кафедра информационных систем и программирования

Отчет к

лабораторной работе №5:

«Машинный перевод»

по дисциплине «Информационный поиск и обработка естественного языка»

Выполнил студент

группы 19-КМ-ПИ1

Ручка Артем Алексеевич

**Задание:**

1. Пройти пятую часть курса на Stepik до пятого практического задания (https://stepik.org/users/338962299/courses?auth=registration)

2. Выполнить пятое практическое задание.

**Задача**: Построить систему машинного перевода (русский --> английский) на основе параллельного корпуса.

**Данные для обучения**: Мы рекомендуем использовать новостной параллельный корпус News Commentary v11. Корпус содержит примерно 200 000 пар предложений. Вы можете использовать дополнительные/другие данные для обучения.

**Инструменты**: Мы предлагаем использовать Moses или seq2seq.

**Данные для тестирования**: В качестве тестового набора вы получите несколько тысяч русских предложений (одно предложение в строке) и должны вернуть перевод этих предложений на английский (одно предложение в строке, с сохранением порядка предложений). Нельзя использовать существующие сторонние системы машинного перевода.

**Оценка**: Для оценки используется среднее значение BLEU среди переводов предложений из тестового набора.

**Ход работы:**

Для выполнения практического задания был выбран язык программирования Python. Данный язык является наиболее подходящим выбором для обработки естественного языка и машинного перевода текстов благодаря наличию большого количества библиотек, упрощающих выполнение задачи. В данном задании использовалась библиотеки numpy, pickle, re.

Загружаем данные и нормализуем их, затем токенизируем текст(рисунок 1).

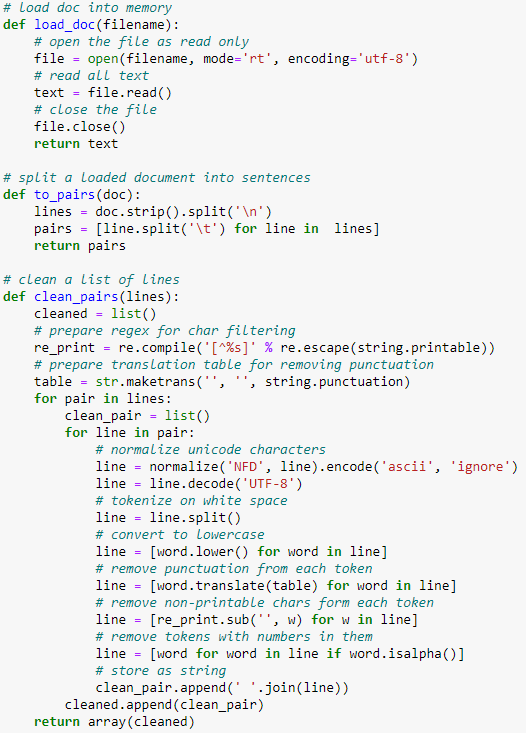


Рисунок 1 – Загрузка данных

Затем строим модель и сохраняем ее с помощью Pickle(рисунок 2).

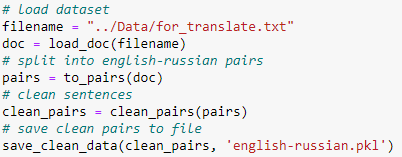


Рисунок 2 – Построение модели

Теперь переводим текст, используя полученную модель(рисунок 3).

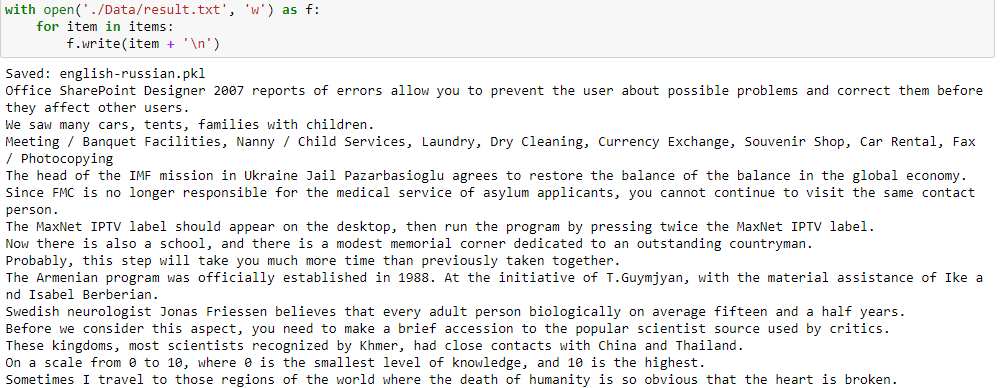


Рисунок 3 – Результат перевода

Проверяем полученный перевод средствами Stepik(рисунок 4).



Рисунок 4 – Проверка результата на Stepik

На этом курс «» окончен. В результате прохождения курса «Введение в обработку естественного языка» был получен сертификат с отличием(рисунок 5).

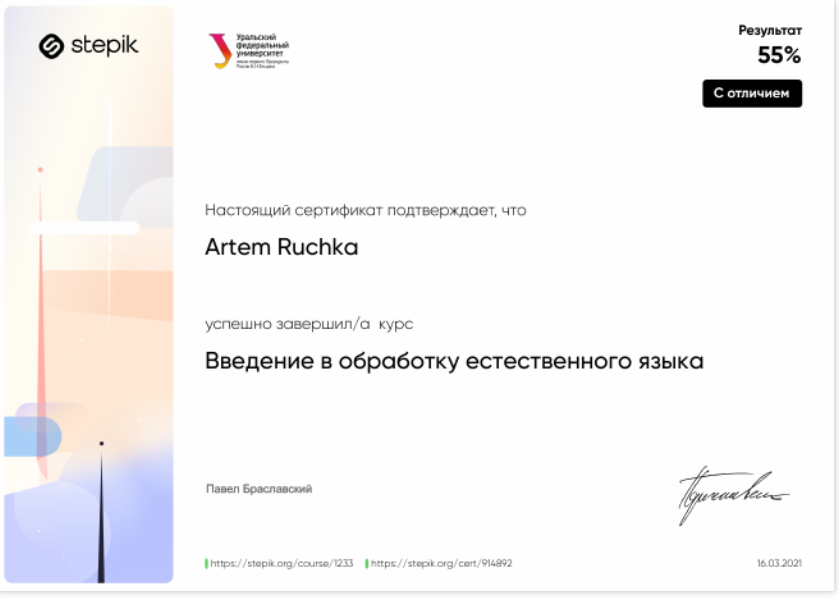


Рисунок 5 - Сертификат

**Заключение**

В результате выполнения данной лабораторной работы был изучен и проведен машинный перевод текстов методом построения системы машинного перевода на основе параллельного корпуса, а также был завершен курс «Введение в обработку естественного языка» и был получен сертификат с отличием.