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 №4:

«Извлечение информации»

по дисциплине «Информационный поиск и обработка естественного языка»

Выполнил студент

группы 19-КМ-ПИ1

Ручка Артем Алексеевич

**Задание:**

1. Пройти четвертую часть курса на Stepik до четвертого практического задания (https://stepik.org/users/338962299/courses?auth=registration)
2. Выполнить четвертое практическое задание.

**Задача**: выделить в тексте именованные сущности двух типов -- организация и персона.

**Ввод:** предложения, разделенные переносом строки.

**Вывод:** для каждого предложения из входных данных вывод в виде "индекс1 длина1 тег1 индекс2 длина2 тег2 ... индексN длинаN тегN EOL", где индексK -- индекс начала токена, длинаK -- длина токена, тегK -- тег токена . В качестве тегов могут выступать PERSON (персона) и ORG (организация). Если именованная сущность состоит из нескольких токенов, тег приписывается к каждому токену. Во входных предложениях могут отсутствовать именованные сущности этих типов. Разделитель в выводе в рамках одного предложения -- пробельный символ, между выводами для различных предложений -- перенос строки.

**Sample Input:**

Барак Обама принимает в Белом доме своего французского коллегу Николя Саркози.

О возможном включении благотворительного фонда в список "иностранных агентов" 7 мая написала газета «Ведомости».

**Sample Output:**

0 5 PERSON 6 5 PERSON 63 6 PERSON 70 7 PERSON EOL

22 18 ORG 41 5 ORG 93 6 ORG 101 9 ORG EOL

**Ход работы:**

Для выполнения практического задания была создана программа на языке Python(рисунок 1), которая выделяет в тексте именованные сущности. Данный язык был выбран, так как его инструментарий содержит все необходимые инструменты для автоматического реферирования документов. В частности, был использован пакет natasha.

Алгоритм обработки текста показан на рисунке 1.

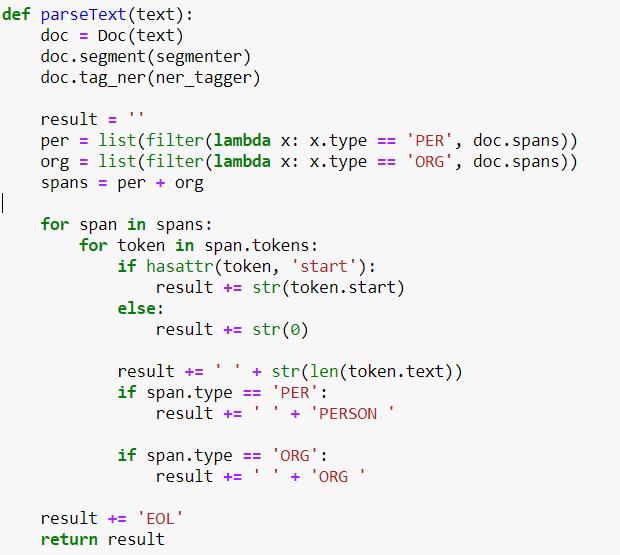


Рисунок 1 – Алгоритм обработки текста

В результате работы програмы получаем объект, в котором находятся обработанные данные(рисунок 2). Эти данные сохраняются в отдельный файл.

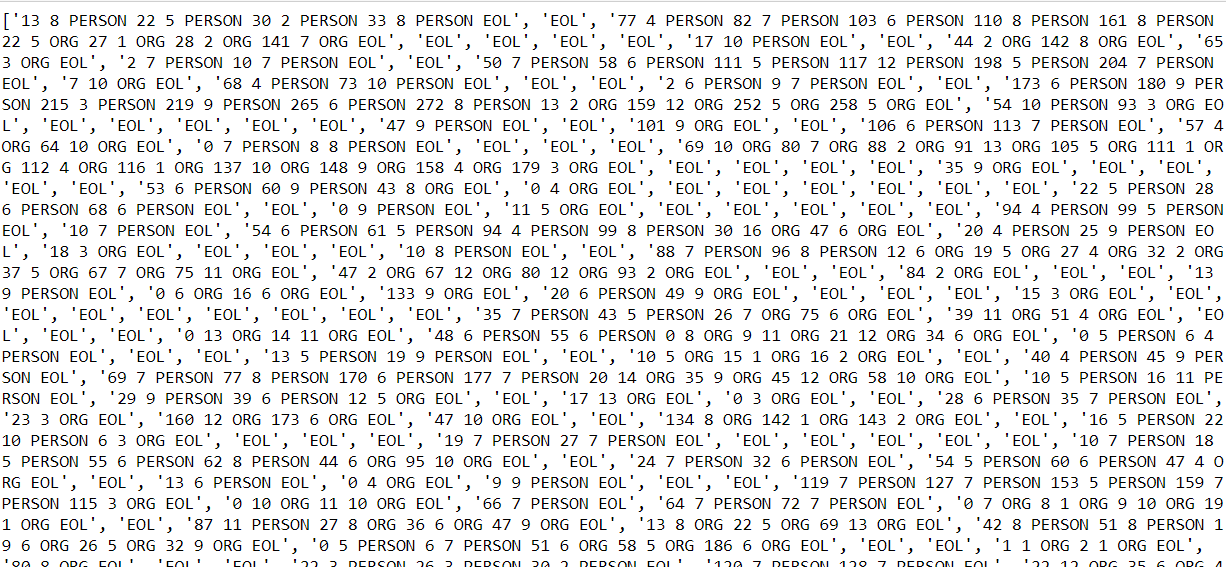


Рисунок 2 – Результат работы программы

Проверяем полученные данные(рисунок 3).

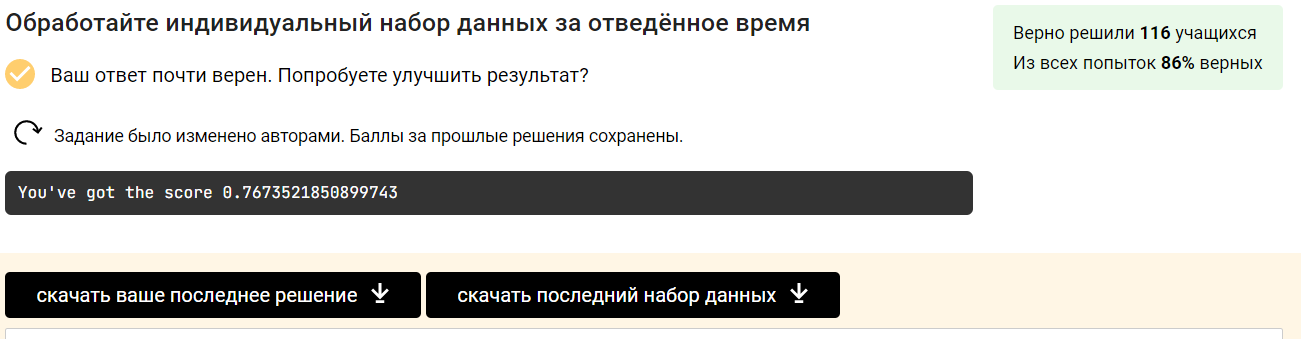


Рисунок 3 – Результат выполнения практического задания

**Заключение**

В результате выполнения данной лабораторной работы была написана программа на языке Python, которая выделяет в тексте именованные сущности, что и требовалось выполнить в четвертом практическом задании курса «Введение в обработку естественного языка».