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 №1:

«Морфологический анализ»

по дисциплине «Информационный поиск и обработка естественного языка»

Выполнил студент

группы 19-КМ-ПИ1

Ручка Артем Алексеевич

**Задание:**

1. Пройти первую часть курса на Stepik до первого практического задания (https://stepik.org/users/338962299/courses?auth=registration)
2. Выполнить задание по лемматизации текста.

**Ход работы:**

1. Задание первое.

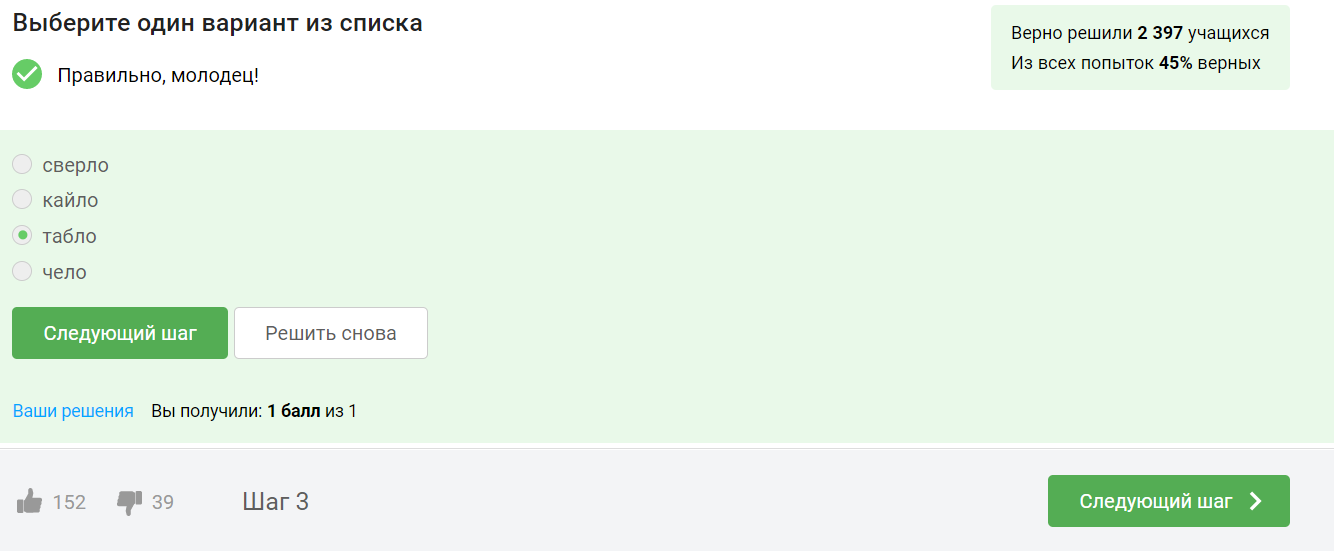


Рисунок 1 – первое задание

Задание второе.

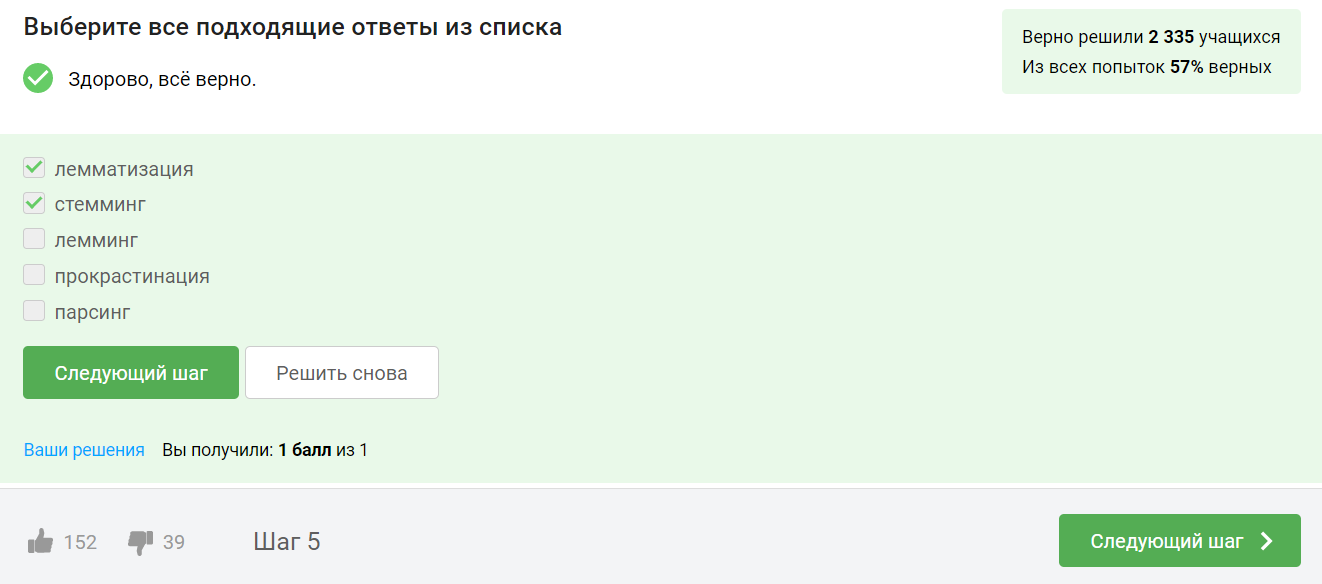


Рисунок 2 – второе задание

Задание третье решено неверно, правильный ответ *S,жен,неод=вин,ед*.

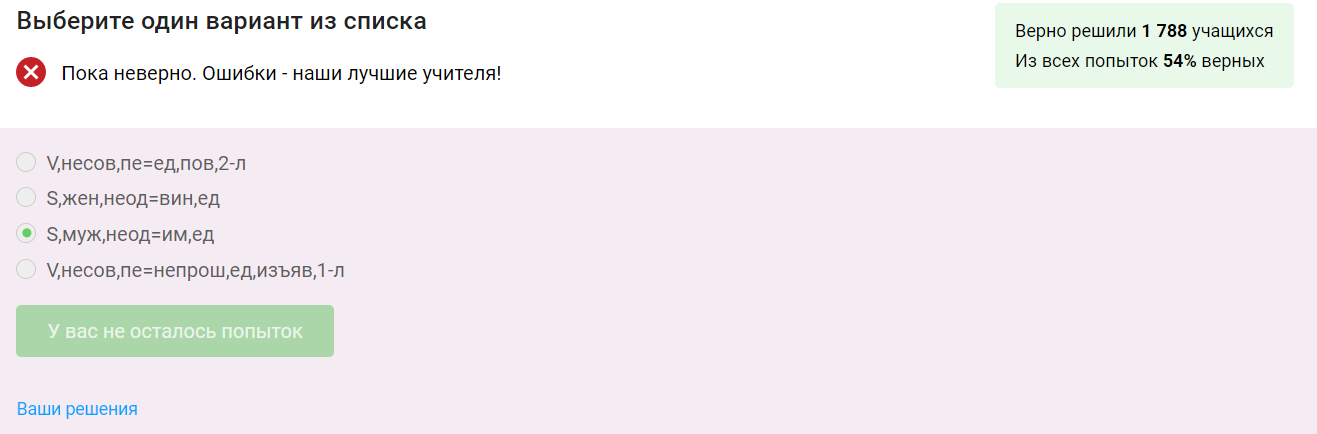


Рисунок 3 – третье задание

Задание четвертое.

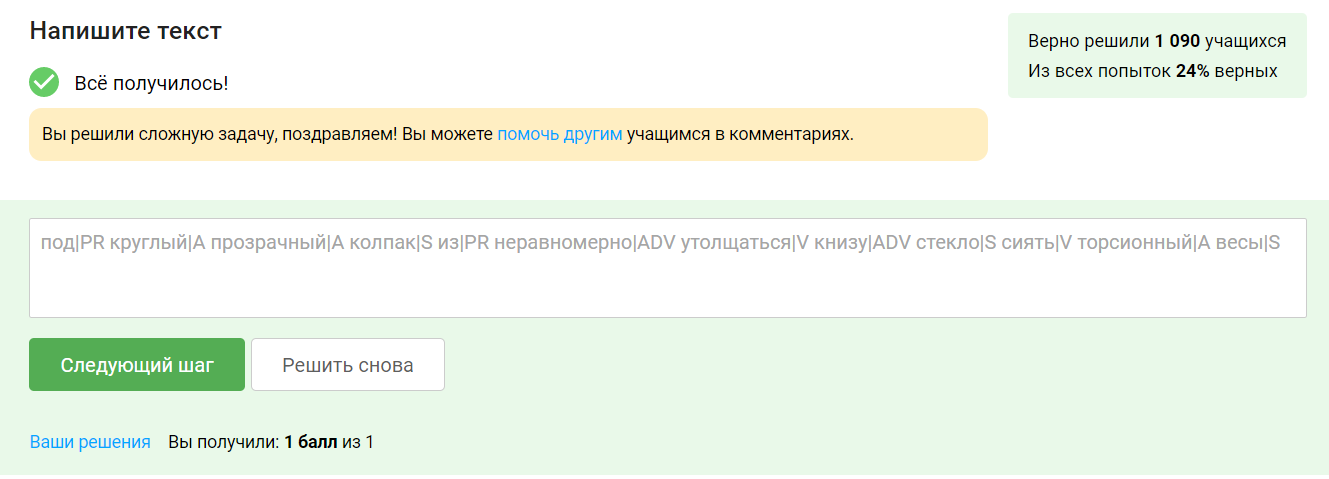


Рисунок 4 – четвертое задание

2) Для решения задачи можно использовать данные, которые упоминались в лекциях: например, словарь [oDict](http://odict.ru/" \t "_blank), разметку [OpenCorpora](http://opencorpora.org/?page=downloads" \t "_blank) и др. Ввод: предложения вида "токен1 токен2 ... токенN" с расставленными знаками препинания, разделенные переносом строки. Из знаков препинания в предложениях могут содержаться только запятая, точка, вопросительный и восклицательный знаки.

Вывод: для каждого предложения из входных данных вывод в виде "токен1{лемма1=тег1} токен2{лемма2= тег2} ... токенN{леммаN=тегN}" без исходных знаков препинания. Разделитель между токенами -- пробельный символ.  
 При лемматизации буквы е и ё, а также написание с прописной/строчной буквы признаются равноправными. Частеречные теги должны быть приведены к следующему инвентарю:

существительные (S),

прилагательные (A),

глаголы, в том числе причастия и деепричастия (V),

предлоги (PR),

союзы (CONJ),

сборная категория (ADV), включающая наречия, вводные слова, частицы, междометия.

В оценке не участвуют и могут быть размечены любым образом (в приведенном примере -- тег NI) местоимения (включая наречные и предикативные), числительные, а также составные предлоги и союзы (потому что, в течение и т.п.).

В результате выполненной работы была написана программа (рисунок 5) на языке Python. Данный язык был выбран, так как его инструментарий содержит все необходимые для обработки естественного языка инструменты. С помощью данной программы был обработан текст(рисунки 6 и 7) и получено решение. Результат проверки решения приведен на рисунке 8.



Рисунок 5 - Программа

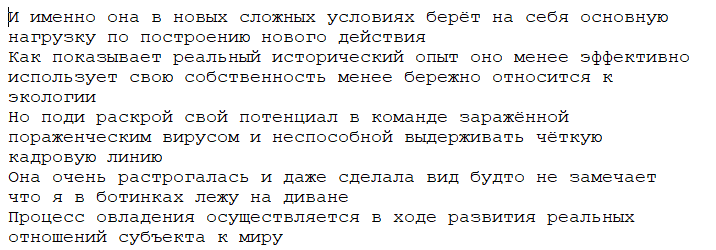


Рисунок 6 – Фрагмент входных данных

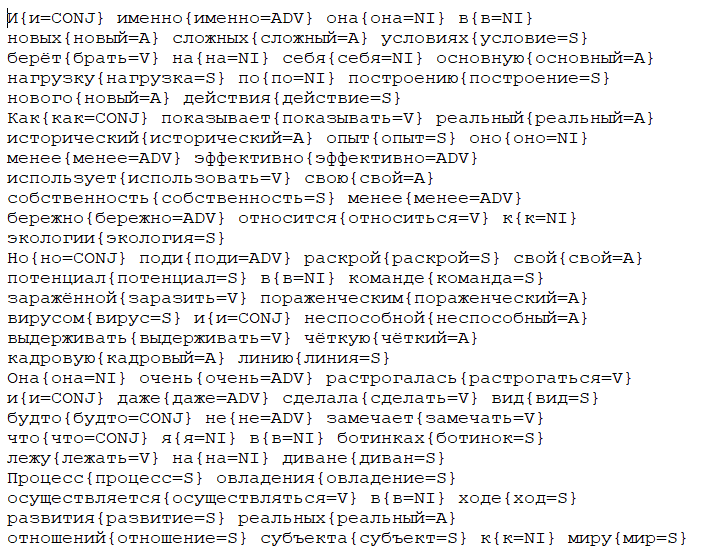


Рисунок 7 –Фрагмент выходных данных

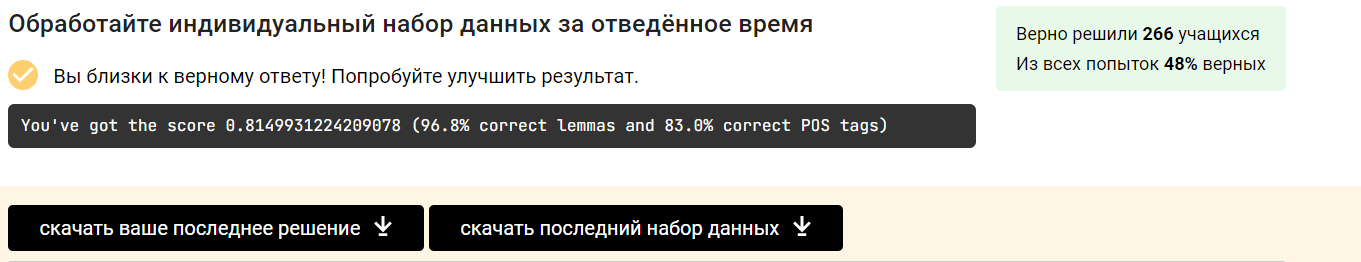


Рисунок 8 – Результат проверки полученного решения

**Заключение**

В результате выполнения данной лабораторной работы была написана программа на языке Python, которая производит лемматизацию текста. Данная программа будет полезна при обработке естественных языков.