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 №2:

«Автоматическое реферирование»

по дисциплине «Информационный поиск и обработка естественного языка»

Выполнил студент

группы 19-КМ-ПИ1

Ручка Артем Алексеевич

**Задание:**

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

**Ход работы:**

Во второй части курса необходимо пройти 8 заданий и выполнить одно практическое задание.

***Первое задание.*** Необходимо построить биграммную языковую модель на основе корпуса:

<s> Вася любит мороженое </s>

<s> Лена любит малину </s>

<s> Вася любит Лену </s>

<s> Георгий ест мороженое </s>

<s> Лена рисует яблоко </s>

<s> Георгий любит Катю </s>

<s> Георгий любит смотреть, как Лена ест мороженое </s>

Затем нужно упорядочить предложения по убыванию оценок вероятностей на основе построенной языковой модели. Выполненное задание показано на рисунке 1.

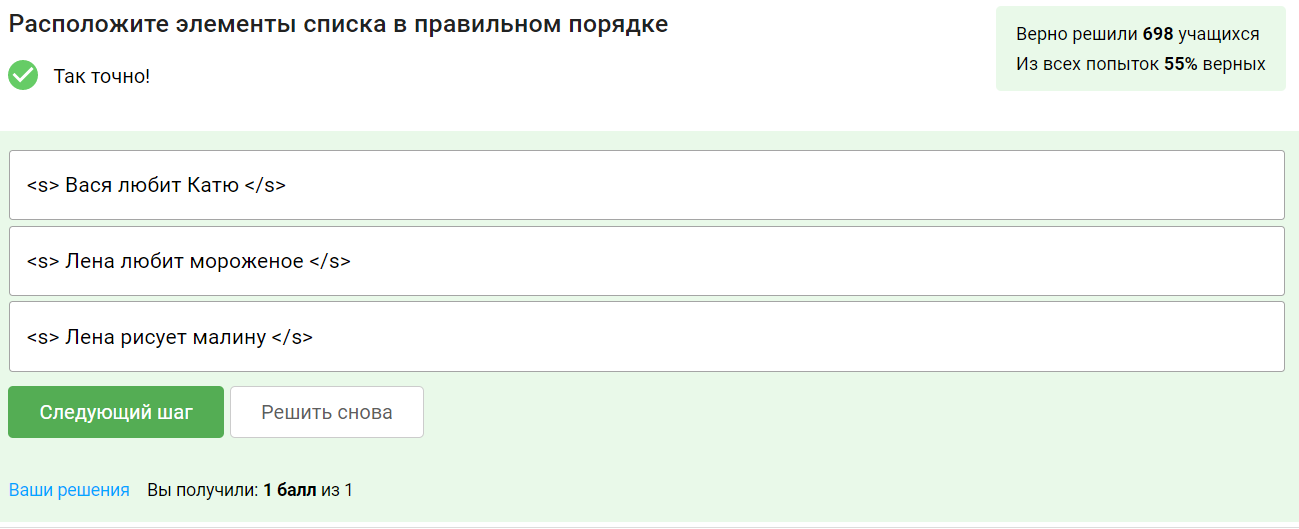


Рисунок 1 – Решение задания 1

***Второе задание.*** Необходимо построить биграммную языковую модель на основе корпуса:

<s> Вася любит мороженое </s>

<s> Лена любит малину </s>

<s> Вася любит Лену </s>

<s> Георгий ест мороженое </s>

<s> Лена рисует яблоко </s>

<s> Георгий любит Катю </s>

<s> Георгий любит смотреть, как Лена ест мороженое </s>

И вычислить перплексию модели на предложении:

<s> Георгий любит малину </s>

Для вычисления вероятности предложения (и, соответственно, перплексии) используются вероятности 4-х биграмм (N=4 в формуле вычисления перплексии).

Выполненное задание показано на рисунке 2.

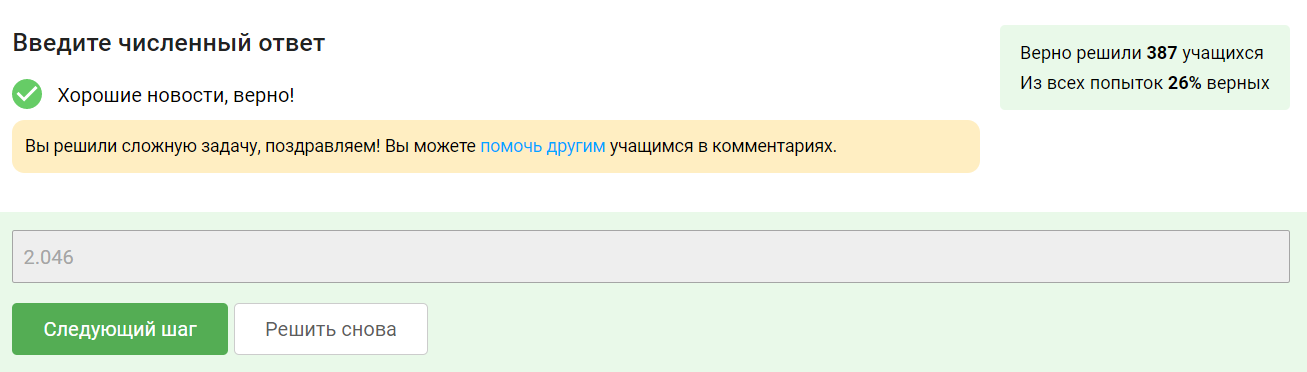


Рисунок 2 – Решение задания 2

***Третье задание.*** Пусть у нас есть корпус, содержащий 10000 предложений, размер словаря -- 1500 уникальных слов (включая специальные "слова" -- маркеры начала и конца предложений). Некоторые частоты униграмм:

ем 100

дуриан 1

и 5000

не 3000

морщусь 50

и биграмм:

<s> ем 20

ем дуриан 0

дуриан и 0

и не 300

не морщусь 15

морщусь </s> 5.

Необходимо применить сглаживание Лапласа (α=1, сглаживание "+1") для оценки вероятностей биграмм и оцените на их основе вероятность предложения <s> ем дуриан и не морщусь </s>

В качестве ответа нужно получить натуральный логарифм оценки вероятности предложения.

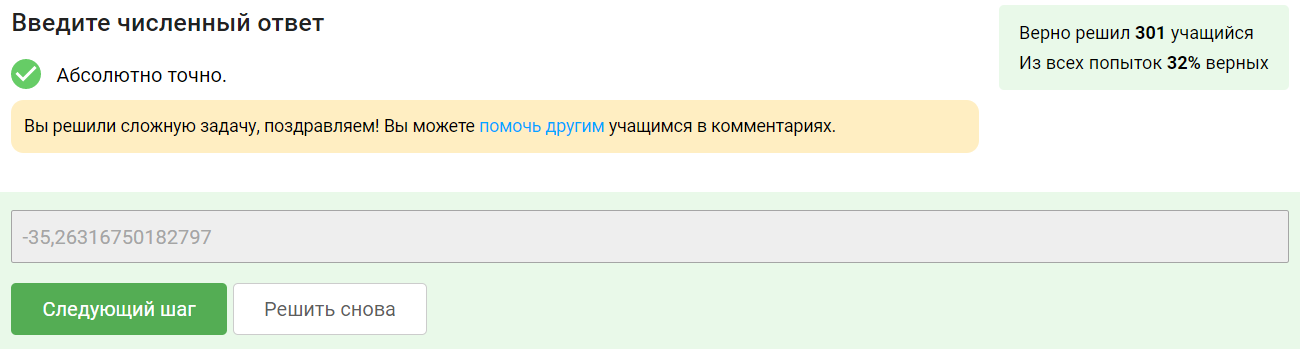


Рисунок 3 – Решение третьего задания

***Четвертое задание.*** Пусть у нас есть два документа из коллекции:

**Doc1**: *красный синий красный зеленый красный фисташковый*  
**Doc2**: *фисташковый алый лазоревый белый красный*

В коллекции всего 10 000 документов. Документные частоты (df) терминов:  
*красный: 100; синий: 80; зеленый: 75; фисташковый: 10; алый: 50; лазоревый: 25; белый: 200*.

Вычислите косинусную меру близости между документами Doc1 и Doc2 на основе их векторного представления с весами tf.idf.  
Для сглаживания "сырых" значений частоты термина в документе используйте формулу wtf = 1 + log10tft,d

Решение задания 4 показано на рисунке 4.

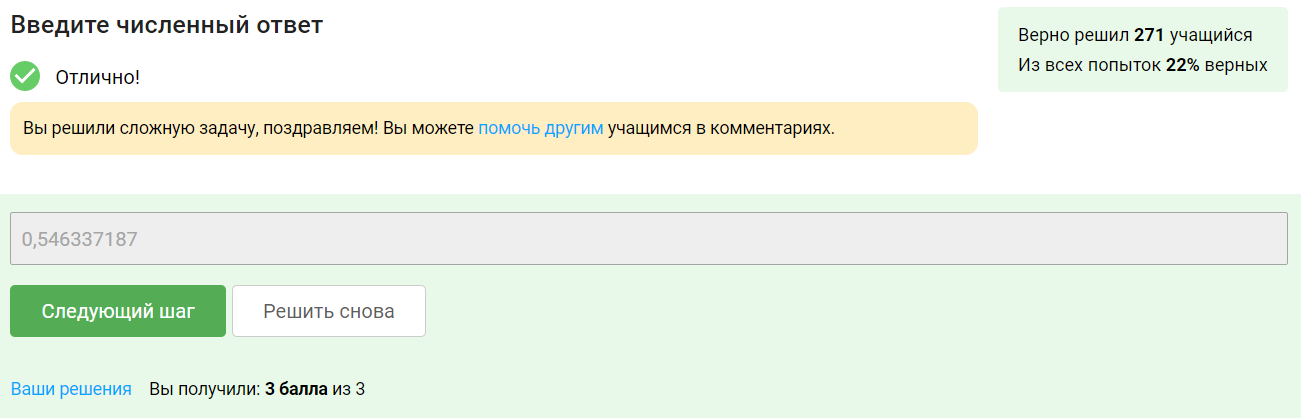


Рисунок 4 – Решение четвертого задания

***Пятое задание.*** Пусть у нас есть коллекция документов

Doc1: яблоко груша яблоко апельсин слива

Doc2: банан ананас яблоко ананас ананас

Doc3: лимон апельсин мандарин помело лайм

Doc4: яблоко дуриан папайя манго маракуйя

и запрос

Q: яблоко банан

Постройте смешанную языковую модель для поиска (смесь моделей по документу и коллекции; вес модели по документу λ=0.7) и ранжируйте документы по уменьшению вероятности соответствия документа запросу.

Решение задания 5 показано на рисунке 5.

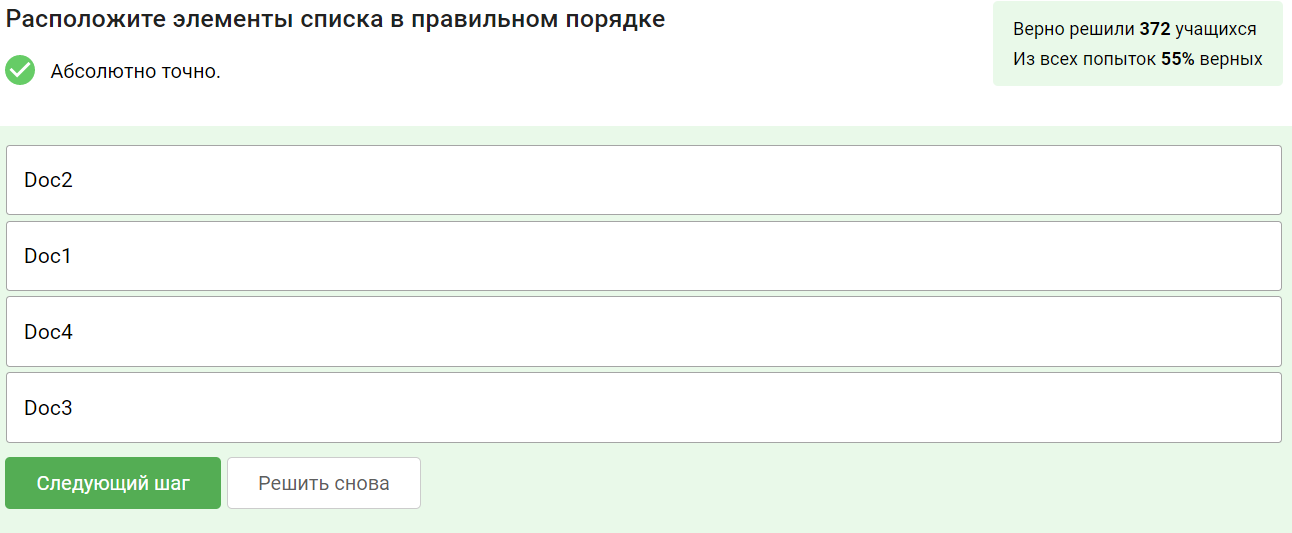


Рисунок 5 – Решение пятого задания

***Шестое задание.*** Пусть есть такой веб-граф (веб-страницы и ссылки между ними):

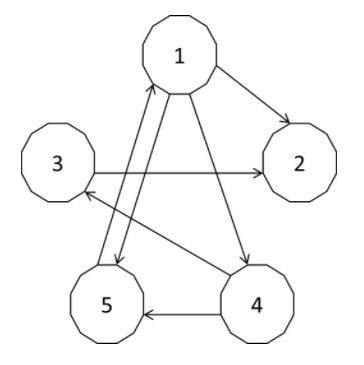


Рисунок 6 – Веб-граф

Вычислите значения PageRank для страниц с учетом вероятности телепортации 0.15 и упорядочите страницы по убыванию PageRank.

Решение задания 6 показано на рисунке 7.

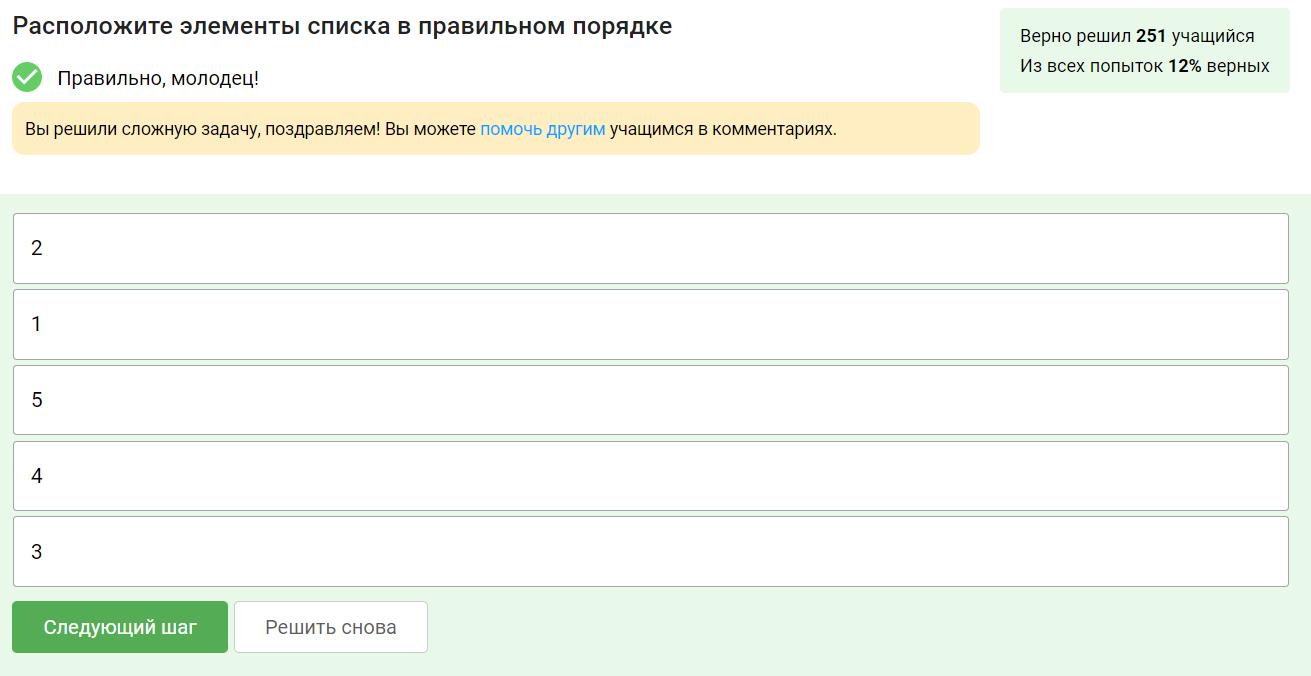


Рисунок 7 – Решение шестого задания

***Седьмое задание.*** Пусть у нас есть оценка top5 результатов поиска по пяти запросам (Q1..Q5); +: релевантный документ, -: нерелевантный документ:

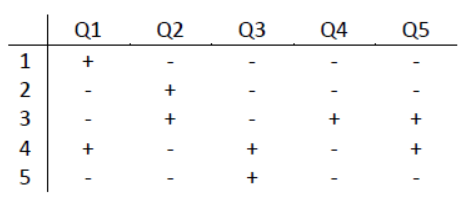


Рисунок 8 – Результаты поиска по пяти запросам

Вычислите макроусредненную среднюю точность (mean average precision, MAP).

Решение задания 7 показано на рисунке 9.

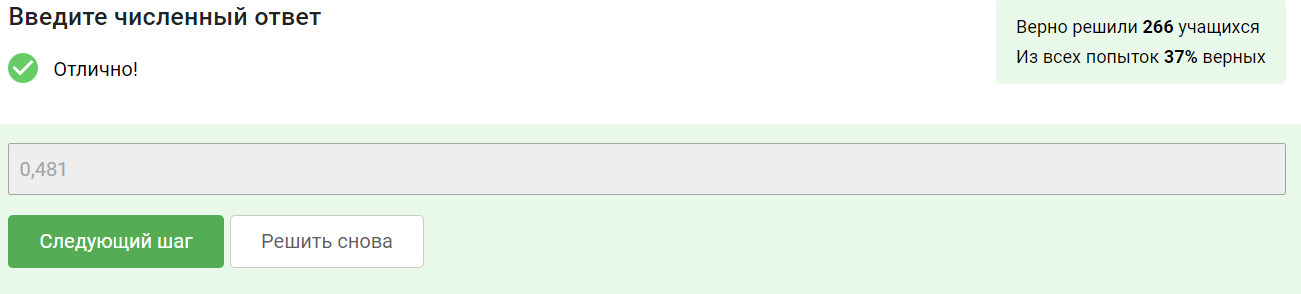


Рисунок 9 – Решение седьмого задания

***Восьмое задание.*** Пусть у нас есть два "реферата-образца":

1: карп лещ лещ щука сазан

2: лещ карп лещ сазан плотва

Упорядочите "рефераты" ниже по убыванию значения ROUGE-2.

Решение данного задания представлено на рисунке 10.

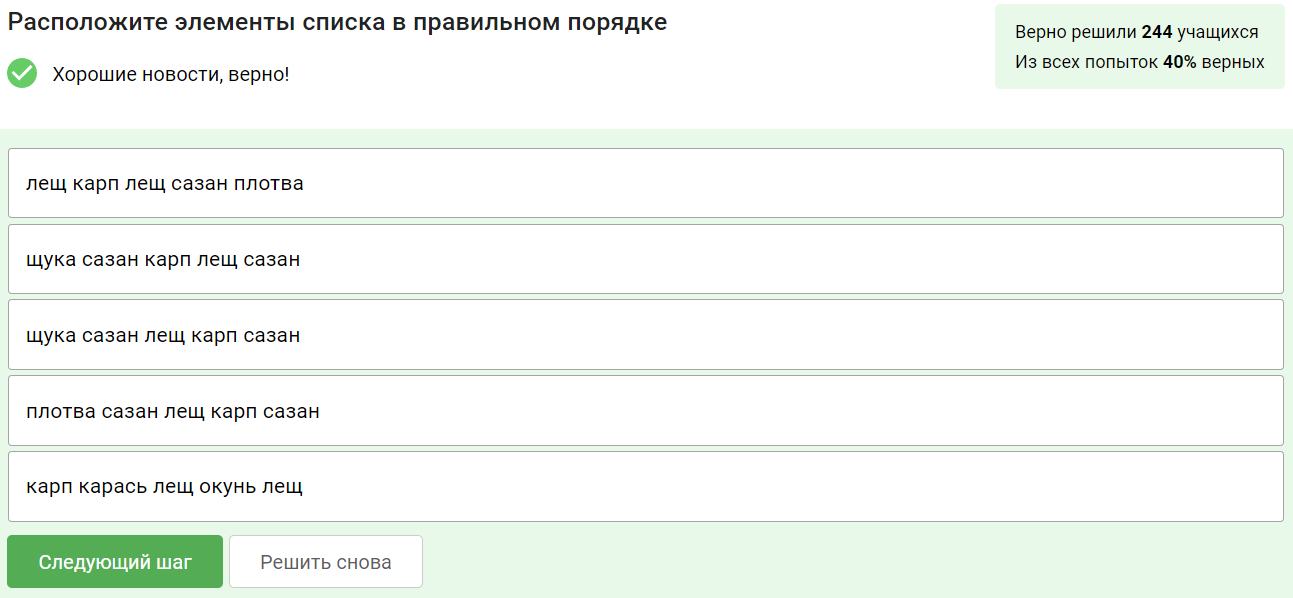


Рисунок 10 – Решение восьмого задания

***Практическое задание.***

**Задача:**Автоматически построить рефераты текстовых документов.  
**Ввод**: Массив текстов в формате JSON.

**Вывод:** Массив рефератов в формате JSON (порядок рефератов соответствует порядку текстов во входных данных).  
  
Максимальный размер каждого из рефератов -- 300 символов (включая пробельные). Если размер реферата превышает указанный порог, то будут оцениваться только первые 300 символов. Тривиальное решение (первые 300 символов документа) допускается, но не приветствуется.  
  
**Оценка**: ROUGE-2 -- близость набору вручную составленных рефератов на основе биграмм слов (значение от 0 до 1).

**Sample Input:**

["Первый текст...", "Второй текст..."]

**Sample Output:**

["Реферат первого текста...", "Реферат второго текста..."]

Для выполнения практического задания была создана программа на языке Python(рисунок 11), которая создает автоматический реферат для входного документа. Данный язык был выбран, так как его инструментарий содержит все необходимые инструменты для автоматического реферирования документов. В частности, был использована пакет textrank и инструменты, входящие в пакет для обработки естественного языка на Python – nltk.

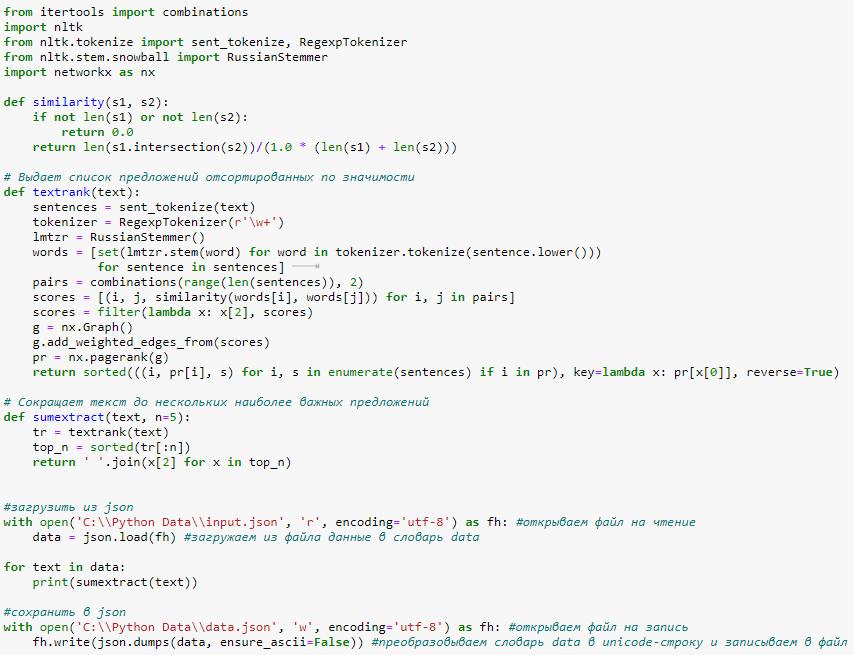


Рисунок 11 – Программа для автоматического реферирования

В результате автоматического реферирования был получен следующий результат(рисунок 12).

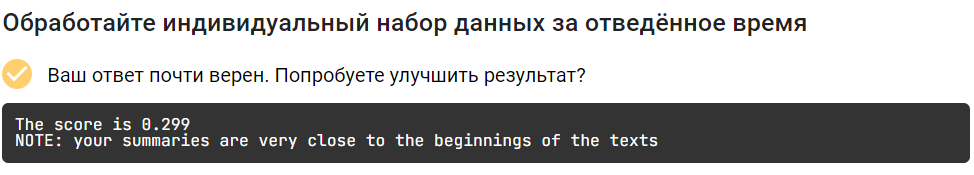


Рисунок 12 – Результат выполнения практического задания

**Заключение**

В результате выполнения данной лабораторной работы были решены все задания второго раздела курса «Введение в обработку естественного языка» и написана программа на языке Python, которая производит автоматическое реферирование документов.