**Московский авиационный институт**

**(Национальный исследовательский университет)**

Факультет прикладной математики и физики

Кафедра вычислительной математики и программирования

**Лабораторная работа № 4**

по курсу «Криптография»

Студент: Алексюнина Ю.В.

Группа: 80-307Б

Преподаватель: Борисов А. В.

Оценка:

Москва, 2020

**Постановка задачи:**

Сравнить:

1) два осмысленных текста на естественном языке

2) осмысленный текст и текст из случайных букв

3) осмысленный текст и текст из случайных слов

4) два текста из случайных букв

5) два текста из случайных слов

Как сравнивать:

Считать процент совпадения букв в сравниваемых текстах – получить дробное значение от 0 до 1 как результат деления количества совпадений на общее число букв. Расписать подробно в отчёте алгоритм сравнения и приложить сравниваемые тексты в

отчёте хотя бы для одного запуска по всем пяти подпунктам. Осознать, какие значения

получаются в этих пяти подпунктах. Привести свои соображения о том почему так происходит. Длина сравниваемых текстов должна совпадать. Привести соображения о том какой длины текста должно быть достаточно для корректного сравнения.

**Решение:**

* В качестве осмысленных текстов на естественном языке были взяты первый и второй том романа «Война и мир» Л.Н. Толстого на английском языке. Данные тексты были выбраны неслучайно. Существует мнение, что все части этого романа чрезвычайно похожи с точки зрения написания, и мне хотелось также проверить и совпадение слов.



* Текст из случайных слов генерируется из следующего словаря (чуть меньше 25 тысяч английских слов): http://svnweb.freebsd.org/csrg/share/dict/words?view=cocontent-type=text/plain
* Текст из случайных букв генерируется из букв английского алфавита в обоих регистрах и состоит из слов длиной от 3 до 10 знаков.
* Тексты из случайных букв/слов сравниваются с первым томом «Война и мир»
* Алгоритм сравнения: параллельно обходим оба текста, сравниваем знаки на одинаковых позициях, исключая символы пробела, перевода строки и null. Если знаки совпадают, то увеличиваем счётчик совпавших символов на 1. Сравнение регистрозависимое.
* При выполнении данной задачи одной из сложностей было восприятие Python-ом текста в другой кодировке, т.к. текстовые файлы читаются только в Unicode, а на моем ноутбуке не было возможности сохранить их в этой кодировке.

**Исходный код:**

import random  
import string  
import urllib.request  
  
  
cnt\_rnd\_txt = 10  
len\_rnd\_txt = 500000  
  
  
def count\_common\_letters(text1, text2):  
 cnt = 0  
 for char1, char2 in zip(text1, text2):  
 if (char1 == char2) and (char1 != **'**\n**'**) and (char1 != **'** **'**) and (char1 != **'**\0**'**):  
 # print(char1)  
 cnt += 1  
 return cnt  
  
  
def match\_perc(text1, text2):  
 return count\_common\_letters(text1, text2) / len(text1)  
  
  
def gen\_random\_letters(n):  
 text = **''** while len(text) < n:  
 len\_word = random.randint(3, 10)  
 word = **''**.join(random.choice(string.ascii\_letters) for \_ in range(len\_word))  
 text += **' '** + word  
 rem = len(text) - n  
 if rem != 0:  
 text = text[:-rem]  
 return text  
  
  
def gen\_random\_words(n):  
 url = **'http://svnweb.freebsd.org/csrg/share/dict/words?view=co&content-type=text/plain'** response = urllib.request.urlopen(url)  
 words = response.read().decode()  
 words = words.splitlines()  
 text = **''** while len(text) < n:  
 text += **' '** + random.choice(words)  
 rem = len(text) - n  
 if rem != 0:  
 text = text[:-rem]  
 return text  
  
  
def full\_of\_sense():  
 print(**"1. Два осмысленных текста на естественном языке."**)  
 handle1 = open(**'1.txt'**, **'r'**)  
 text1 = handle1.read()  
 # text1 = [line.rstrip() for line in text1]  
 handle2 = open(**'2.txt'**, **'r'**)  
 text2 = handle2.read()  
 # text2 = [line.rstrip() for line in text2]  
 min\_len = min(len(text1), len(text2))  
 text1 = text1[:min\_len]  
 text2 = text2[:min\_len]  
 print(**"Длина текста: {0}"**.format(min\_len))  
 print(**"Процент совпадений: {0}"**.format(match\_perc(text1, text2)))  
  
  
def sense\_and\_randoml():  
 print(**"2. Осмысленный текст и текст из случайных букв."**)  
 handle1 = open(**'1.txt'**, **'r'**)  
 text1 = handle1.read()  
 s = 0  
 for \_ in range(cnt\_rnd\_txt):  
 text2 = gen\_random\_letters(len(text1))  
 s += match\_perc(text1, text2)  
 s /= cnt\_rnd\_txt  
 print(**"Длина текста: {0}"**.format(len(text1)))  
 print(**"Процент совпадений: {0}"**.format(s))  
  
  
def sense\_and\_randomw():  
 print(**"3. Осмысленный текст и текст из случайных слов."**)  
 handle1 = open(**'1.txt'**, **'r'**)  
 text1 = handle1.read()  
 s = 0  
 for \_ in range(cnt\_rnd\_txt):  
 text2 = gen\_random\_words(len(text1))  
 s += match\_perc(text1, text2)  
 s /= cnt\_rnd\_txt  
 print(**"Длина текста: {0}"**.format(len(text1)))  
 print(**"Процент совпадений: {0}"**.format(s))  
  
  
def randoml():  
 print(**"4. Два текста из случайных букв."**)  
 s = 0  
 for \_ in range(cnt\_rnd\_txt):  
 text1 = gen\_random\_letters(len\_rnd\_txt)  
 text2 = gen\_random\_letters(len\_rnd\_txt)  
 s += match\_perc(text1, text2)  
 s /= cnt\_rnd\_txt  
 print(**"Длина текста: {0}"**.format(len\_rnd\_txt))  
 print(**"Процент совпадений: {0}"**.format(s))  
  
  
def randomw():  
 print(**"5. Два текста из случайных слов."**)  
 s = 0  
 for \_ in range(cnt\_rnd\_txt):  
 text1 = gen\_random\_words(len\_rnd\_txt)  
 text2 = gen\_random\_words(len\_rnd\_txt)  
 s += match\_perc(text1, text2)  
 s /= cnt\_rnd\_txt  
 print(**"Длина текста: {0}"**.format(len\_rnd\_txt))  
 print(**"Процент совпадений: {0}"**.format(s))  
  
  
full\_of\_sense()  
sense\_and\_randoml()  
sense\_and\_randomw()  
randoml()  
randomw()

**Результат работы программы:**

1. Два осмысленных текста на естественном языке.

Длина текста: 476646

Процент совпадений: 0.01830289145403507

2. Осмысленный текст и текст из случайных букв.

Длина текста: 543202

Процент совпадений: 0.006512126244012356

3. Осмысленный текст и текст из случайных слов.

Длина текста: 543202

Процент совпадений: 0.0200013991111962

4. Два текста из случайных букв.

Длина текста: 500000

Процент совпадений: 0.004395799999999997

5. Два текста из случайных слов.

Длина текста: 500000

Процент совпадений: 0.014553359999999994

**Выводы:**

Как видно из результатов, наилучшие совпадения получаются путём сравнения двух осмысленных текстов и осмысленного текста с текстом из случайных слов. Худшие совпадения у осмысленного текста с текстом из случайных букв и у двух текстов из случайных букв.

К сожалению, первый и второй том произведения «Война и мир» на английском языке не совпадают настолько, насколько хотелось бы.

Думаю, полученные результаты можно было бы объяснить какими-то лингвистическими законами построения языка, но у меня нет достаточных знаний в этой области. Например, для букв английского языка характерна некоторая частотность, которая будет соблюдаться в осмысленных текстах, и которая не соблюдается в генерируемых из букв текстах. Эмпирически кажется, что размер слогов в осмысленных словах совпадает чаще, чем в случайных, и 5 букв, передающих гласные звуки в английском языке, будут совпадать чаще.

В текстах из случайных букв нет никаких ограничений на использование букв в верхнем регистре не на первой позиции в слове, что сильно снижает количество совпадений с осмысленным тестом.

Что касается достаточной длины текста для корректного сравнения, я нашла следующую информацию в книге «Определение жанра и автора литературного произведения статистическими методами» (Авторы: Ю. Орлов, К. Осминин):





