**Отчет по лабораторной работе №4 по курсу «Криптография»**

Выполнил Лисин Роман Сергеевич, М8О-306Б-20.

***Задание***

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.

***Ход работы***

Алгоритм сравнения представлен в функции compare: на вход поступают два текста. Далее идёт проверка, что у них одинаковая длина (если не одинаковая, то бросится исключение). Потом проходимся по обоим текстам одновременно посимвольно и считаем все совпадения символов. Подсчитав совпадения, делим их количество на длину любого из двух текстов и умножаем на 100, чтобы получить процент совпадений.

Прикрепляю к письму книгу Фрэнсиса Скотта Фицджеральда «Великий Гэтсби» на английском, которую я использую в качестве осмысленного текста. Для получения английских слов использую модуль english\_words (документация https://pypi.org/project/english-words/).

Код на Python (в комментариях пояснения по ходу выполнения работы):

import numpy as np  
import random  
import string  
  
from english\_words import get\_english\_words\_set  
  
  
def get\_text\_of\_random\_chars(size):  
 *"""*  
 *Возвращает текст из случайных символов в количестве size*  
 *"""*  
letters = string.ascii\_lowercase + ' ' # Алфавит: ASCII-символы и пробел  
 random\_chars\_text = np.random.choice(list(letters), size, replace=True)  
 random\_chars\_text = ''.join(random\_chars\_text)  
 return random\_chars\_text  
  
  
def get\_text\_of\_random\_words(size):  
 *"""*  
 *Возвращает текст из количества size английских слов в случайном порядке*  
 *"""*  
cur\_len = 0  
 words\_for\_text = []  
 english\_words = list(get\_english\_words\_set(['web2'], lower=True)) # 234450 слов, при каждом запуске программы  
 # слова разные  
 while cur\_len <= size:  
 random\_word = random.choice(english\_words)  
 cur\_len += len(random\_word) + 1  
 words\_for\_text.append(random\_word)  
 result = ' '.join(words\_for\_text)[:size]  
 return result  
  
  
def get\_great\_gatsby\_text():  
 with open('The-Great-Gatsby.txt') as f:  
 full\_text\_with\_punctuation = f.read().replace('\n', ' ').replace('\r', '').lower()  
 punctuation = set(string.punctuation)  
 return ''.join(c for c in full\_text\_with\_punctuation if c not in punctuation)  
  
  
def compare(text1, text2):  
 assert len(text1) == len(text2) # Проверяем, что у текстов одинаковая длина  
 matched = 0  
 for c1, c2 in zip(text1, text2):  
 if c1 == c2:  
 matched += 1  
 return matched / len(text1) \* 100 # Получаем процент совпадений  
  
  
if \_\_name\_\_ == "\_\_main\_\_":  
 size = 125000 # Длина текста  
 gatsby\_text = get\_great\_gatsby\_text() # Длина - 257427 символов  
 real\_text1 = gatsby\_text[:size] # получаем текст длиной size - первая часть осмысленного текста  
 real\_text2 = gatsby\_text[size:2 \* size] # получаем текст длиной size - следующая часть осмысленного текста  
 random\_chars1 = get\_text\_of\_random\_chars(size)  
 random\_words1 = get\_text\_of\_random\_words(size)  
 random\_chars2 = get\_text\_of\_random\_chars(size)  
 random\_words2 = get\_text\_of\_random\_words(size)  
  
 print("Процент совпадения двух текстов:")  
 print(f"Два осмысленных текста на естественном языке: {compare(real\_text1, real\_text2)} процентов")  
 print(f"Осмысленный текст и текст из случайных букв: {compare(real\_text1, random\_chars1)} процентов")  
 print(f"Осмысленный текст и текст из случайных слов: {compare(real\_text1, random\_words1)} процентов")  
 print(f"Два текста из случайных букв: {compare(random\_chars1, random\_chars2)} процентов")  
 print(f"Два текста из случайных слов: {compare(random\_words1, random\_words2)} процентов")  
  
 """  
 Процент совпадения двух текстов:  
 Два осмысленных текста на естественном языке: 7.4048 процентов  
 Осмысленный текст и текст из случайных букв: 3.6159999999999997 процентов  
 Осмысленный текст и текст из случайных слов: 6.304800000000001 процентов  
 Два текста из случайных букв: 3.6656 процентов  
 Два текста из случайных слов: 6.1208 процентов  
 """

***Вывод***

Самые низкие проценты (около 3,6%) получились у текстов, составленных из случайных символов, так как в любом человеческом языке есть правила и закономерности, по которым формируются слова, и просто случайная генерация символов не может приблизиться к человеческому языку. По такой же причине текст из случайных символов хуже совпадает с другим текстом из случайных символов, чем тексты из случайных слов (около 6%).

Самый высокий процент совпадения (около 7,4%) у осмысленных текстов. Этот процент выше, чем у осмысленного текста и текста из случайных слов (около 6,3%), так как между словами в естественном языке тоже есть связь. Слова образуют словосочетания, словосочетания - предложения. Поэтому результат получается логичным.

Для корректного сравнения текстов их длина должна быть достаточно большой. Это значит, что 10 символов будет точно недостаточно. К примеру, опытным путём получено значение не менее 5000 символов для корректного сравнения.