**Московский авиационный институт**

**(Национальный исследовательский университет)**

Факультет прикладной математики и физики

Кафедра вычислительной математики и программирования

**Лабораторная работа № 4**

по курсу «Криптография»

Студент: Пивницкий Д.С.

Группа: М80-306Б-19

Преподаватель: Борисов А. В.

Оценка:

Москва, 2022

**1. 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.

**2. Метод решения**

Я взял осмысленные тексты из [http://www.gutenberg.org](http://www.gutenberg.org/). Тексты из случайных букв генерировались с использованием регистрозависимого латинского алфавита. Случайные слова были взяты из файлов <https://github.com/first20hours/google-10000-english>.

Длина слов для текстов из случайных букв составляет от 3 до 10 символов, для текстов из слов - беру слова из трех файлов (короткие, средние, длинные).

Сравнение текстов происходит побуквенно, если буквы в одинаковых позициях совпали, то увеличиваем счетчик совпадений.

**3. Полученные результаты**

Comparison 1: two meaningful text in natural language

Text length: 717618

Match percentage: 0.062095711088629324

Comparison 2: meaningful text and text from random letters

Text length: 717618

Match percentage: 0.035779481562614096

Comparison 3: meaningful text and text from random words

Text length: 717618

Match percentage: 0.06200234665239723

Comparison 4: two texts from random letters

Text length: 700000

Match percentage: 0.03456957142857143

Comparison 5: two texts from random words

Text length: 700000

Match percentage: 0.06566414285714287

**4. Код программы**

import random

import urllib.request

import string

TEXT\_LENGTH = 700000

TEST\_NUM = 10

def commmon\_letters\_num(text1, text2):

num = 0

for ch1, ch2 in zip(text1, text2):

if ch1 == ch2:

num += 1

return num

def match\_perc(text1, text2):

return commmon\_letters\_num(text1, text2) / len(text1)

def rand\_letter():

return random.choice(string.ascii\_letters)

def rand\_text(n):

text = ''

while len(text) < n:

word\_len = random.randint(3, 9)

word = ''.join(rand\_letter() for i in range(word\_len))

text += ' ' + word

if len(text) > n:

text = text[:(n - len(text))]

return text

def rand\_words(n):

url\_short\_words = '<https://raw.githubusercontent.com/first20hours/google-10000-english/master/google-10000-english-usa-no-swears-short.txt>'

url\_mid\_words = '<https://raw.githubusercontent.com/first20hours/google-10000-english/master/google-10000-english-usa-no-swears-medium.txt>'

url\_long\_words = '<https://raw.githubusercontent.com/first20hours/google-10000-english/master/google-10000-english-usa-no-swears-long.txt>'

dictionary = urllib.request.urlopen(url\_short\_words).read().decode()\

+ urllib.request.urlopen(url\_mid\_words).read().decode()\

+ urllib.request.urlopen(url\_long\_words).read().decode()

dictionary = dictionary.splitlines()

text = ''

while len(text) < n:

text += ' ' + random.choice(dictionary)

if len(text) > n:

text = text[: (n - len(text))]

return text

def comp1():

print("Comparison 1: two meaningful text in natural language")

url1 = '<http://www.gutenberg.org/files/1342/1342-0.txt>'

url2 = '<http://www.gutenberg.org/files/2600/2600-0.txt>'

text1 = urllib.request.urlopen(url1).read().decode()

text2 = urllib.request.urlopen(url2).read().decode()

min\_len = min(len(text1), len(text2))

text1 = text1[:min\_len]

text2 = text2[:min\_len]

print("Text length: {0}".format(min\_len))

print("Match percentage: {0}".format(match\_perc(text1, text2)))

def comp2():

print("Comparison 2: meaningful text and text from random letters")

url1 = '<http://www.gutenberg.org/files/1342/1342-0.txt>'

text1 = urllib.request.urlopen(url1).read().decode()

text2 = rand\_text(len(text1))

print("Text length: {0}".format(len(text1)))

print("Match percentage: {0}".format(match\_perc(text1, text2)))

def comp3():

print("Comparison 3: meaningful text and text from random words")

url1 = '<http://www.gutenberg.org/files/1342/1342-0.txt>'

text1 = urllib.request.urlopen(url1).read().decode()

m = 0

for i in range(TEST\_NUM):

text2 = rand\_words(len(text1))

m += match\_perc(text1, text2)

m /= TEST\_NUM

print("Text length: {0}".format(len(text1)))

print("Match percentage: {0}".format(m))

def comp4():

print("Comparison 4: two texts from random letters")

m = 0

for i in range(TEST\_NUM):

text1 = rand\_text(TEXT\_LENGTH)

text2 = rand\_text(TEXT\_LENGTH)

m += match\_perc(text1, text2)

m /= TEST\_NUM

print("Text length: {0}".format(len(text1)))

print("Match percentage: {0}".format(m))

def comp5():

print("Comparison 5: two texts from random words")

m = 0

for i in range(TEST\_NUM):

text1 = rand\_words(TEXT\_LENGTH)

text2 = rand\_words(TEXT\_LENGTH)

m += match\_perc(text1, text2)

m /= TEST\_NUM

print("Text length: {0}".format(len(text1)))

print("Match percentage: {0}".format(m))

if \_\_name\_\_ == '\_\_main\_\_':

comp1()

comp2()

comp3()

comp4()

comp5()

**5. Выводы**

По результатам видно, что лучше всего совпали осмысленные тексты, осмысленный текст и текст из случайных слов, а также два текста из случайных слов.

Что касается осмысленных текстов, то здесь вероятность высокого совпадения выше по причине лингвистических особенностей. Устоявшиеся конструкции, так называемые n-граммы, часто встречающиеся слоги и т.д. Для опыта я взяла разные произведения - "Гордость и предубеждение" Д.Остин и "Война и мир" Л.Толстова. Процент совпадения получился около 0.06. Затем для интереса были взяты произведения одного автора - сказки братьев Гримм. В таком сравнении процент совпадения текстов возрос и составил около 0.07. Очевидно, что у каждого автора есть свой почерк, свой словарь, что увеличивает "повторения".

В текстах из случайных слов в моем случае был взят единый словарь. Это, конечно же, дало высокий показатель совпадений. В случае использования разных словарей, сравнение дает более низкий результат.

Со случайными буквами всё гораздо сложнее. Невозможно дать точную оценку совпадений, так как в тестах использовался регистрозависимый алфавит. В случае сравнения двух текстов, мы видим, что вероятность встретить ту или иную букву составила 1/58 вместо 1/26 в регистронезависимом алфавите. Становится ясно, что это ухудшает ситуацию.

Попытки сравнить осмысленный текст и текст из случайных букв видятся мне не самыми удачными по причине того, что в тексте какого-либо произвидения, например, встречаются ещё и знаки препинания. Если принебречь заглавными буквами в случайном тексте, то, возможно, в сравнении будет больше смысла.