# Лабораторная работа № 5 по курсу криптографии

Выполнила студентка группы М8О-307Б *Безлуцкая Елизавета.*

## Условие

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

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 случаям. Осознать какие значения получаются в этих пяти случаях. Привести соображения о том почему так происходит.

Длина сравниваемых текстов должна совпадать. Привести соображения о том какой длины текста должно быть достаточно для корректного сравнения.

## Метод решения

Я взяла осмысленные тексты из 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.
Результаты сравнения:

```
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
```

## Исходный код

```python
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/74/74-0.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()
```

## 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 $\frac{1}{58}$ вместо $\frac{1}{26}$ 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.