.

Лабораторная работа № 4 по курсу криптографии

Выполнил студент группы М8О-308Б-17 *Иларионов Денис.*

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.

Также я решил еще сравнить

6. Текст из случайных слов и случайных букв.
7. Два текста из случайных букв без пробелов и пер. строк.

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

В качестве осмысленных текстов я решил взять произведения двух разных авторов - Стивена Кинга - "Институт"и Нила Стивенсона - "Анафем"
Данные произведения разных жанров (ужасы и научная фантастика)

Текст из случайных слов генерируется из словаря русских слов (около 10 тыс). Нашел я его на этом сайте:
http://speakrus.ru/dict/index.htm

Текст из случайных букв генерируется в моей программе на питоне из букв русского алфавита (с обоими регистрами) и после каждого слова ставится пробел или перенос

строк (иногда даже два), с небольшой вероятностью. (Рассматривается и случай без пробелов). Длина слов - от 1 до 12.

Алгоритм сравнения: так как в задании должно быть то, что тексты одной длины, то мы находим минимальную длину обоих текстов и сравниваем только до символа этой длины. (Например если в 1 тексте 800 тыс знаков, а во 2 - 1.3 млн, то мы сравниваем только первые 800 тыс.). Сравнение регистрозависимое. Сравниваются одинаковые позиции в каждом тексте. Если символы совпадают, счетчик увеличивается на 1. Выводит функция общую длину текстов и количество совпадений. Процент находится делением числа совпадений на длину и умножением на 100. При сравнении разных типов текстов, которых уже 2 (я генерировал тексты и записывал их в файлы), я решил сравнивать их все комбинации, чтобы увеличить точность вычислений, а потом найти средний процент.

Результат работы программы

```
Случай 1: 2 осмысленных текста на русском языке

Общая длина текстов = 947555
Совпало символов = 53640
Процент = 5.66089%

Случай 2: Осмысленный текст и текст из рандомных букв

Тексты 1 и 1:
Общая длина текстов  = 947555
Совпало символов = 28745
Процент = 3.0336%

Тексты 1 и 2:
Общая длина текстов  = 947555
Совпало символов = 29389
Процент = 3.10156%

Тексты 2 и 1:
Общая длина текстов  = 1053818
Совпало символов = 31895
Процент = 3.02661%

Тексты 2 и 2:
Общая длина текстов  = 1053502
Совпало символов = 32055
Процент = 3.04271%
```

```
Средний процент = 3.05112%

Случай 3: Осмысленный текст и текст из рандомных слов

Тексты 1 и 1:
Общая длина текстов  = 947555
Совпало символов = 44811
Процент = 4.72912%

Тексты 1 и 2:
Общая длина текстов  = 947555
Совпало символов = 45269
Процент = 4.77745%

Тексты 2 и 1:
Общая длина текстов  = 1154384
Совпало символов = 55508
Процент = 4.80845%

Тексты 2 и 2:
Общая длина текстов  = 1152452
Совпало символов = 55722
Процент = 4.83508%

Средний процент = 4.78753%

Случай 4: 2 текста из рандомных букв (с пробелами и пер. строк)

Общая длина текстов = 1053502
Совпало символов = 29850
Процент = 2.83341%

Случай 5: 2 текста из рандомных слов

Общая длина текстов = 1152452
Совпало символов = 57402
Процент = 4.98086%

Случай 6: Тексты из рандомных слов и рандомных букв

Тексты 1 и 1:
Общая длина текстов  = 1053818
```

```
Совпало символов = 23158
Процент = 2.19753%

Тексты 1 и 2:
Общая длина текстов  = 1053502
Совпало символов = 23112
Процент = 2.19383%

Тексты 2 и 1:
Общая длина текстов  = 1053818
Совпало символов = 23113
Процент = 2.19326%

Тексты 2 и 2:
Общая длина текстов  = 1053502
Совпало символов = 23140
Процент = 2.19648%

Средний процент = 2.19528%

Случай 7: 2 текста из рандомных букв (без пробелов и пер. строк)

Общая длина текстов = 901559
Совпало символов = 13960
Процент = 1.54843%
```

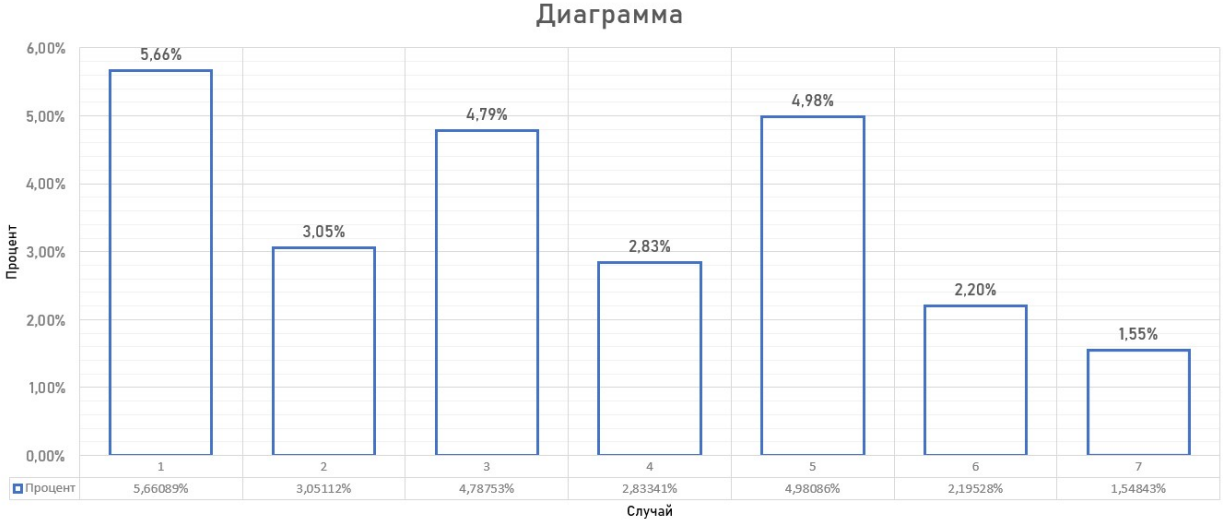
Выводы

Из моих результатов видно то, что наилучшие совпадения получаются путём сравнения двух осмысленных текстов. Также вполне высокий процент имеют сравнения осмысленного текста с текстом из случайных слов и сравнения 2 текстов из случайных слов. Меньше всего совпадений при сравнение любого текста с текстом из случайных букв, особенно сравнение текстов из случайных букв без пробелов. Думаю, дело в том, что такой высокий процент получился у 2 осмысленных текстов, из-за того, что в осмысленных текстах вероятность определенных комбинаций слов больше, в пользу осмысления данного текста. Сравения с текстом из случайных слов дает процент поменьше, но все еще большой. Скорее всего, потому что некоторые слова совпадают полностью, но комбинации слов встречаются намного реже, тк текст из случайных слов редко бывает осмысленным. А процент с текстом из случайных букв такой маленький, потому что чаще всего это просто слова, которые не несут никакого смысла, а еще буквы в середние этих слов могут быть верхнего регистра, что почти не встречается в осмысленных словах (кроме аббревиатур и редких случаев), поэтому и вероятность совпадений меньше. Однако, странно, что результат показал куда больший процент сравнения текста из

случайных букв с осмысленным текстом, чем сравнение 2 текстов из случайных букв. (3 процента против 2). Возможно, просто так совпало. А процент совпадения 2 текстов из случайных слов больше процента совпадения текста из случайных слов и осмысленного текста потому, что случайные слова взяты из одного и того же словаря, а словарь не слишком большой (не считаю, что 100 тыс слов - очень много). В то время, в этом словаре глаголы стоят в инфинитиве, когда в осмысленных текстах они могут стоять в любой форме, и это делает совпадения слов не полностью.

Также я думаю, что если брать произведения одного и того же автора, то процент совпадения будет больше, так как слова в его произведении будут написаны в одном авторском стиле, а это увеличивает шанс того, что встретятся одни и те же слова.

Что касается достаточной длины текста, я считаю, что мои тесты были достаточной длины. Около миллиона символов вполне достаточно, чтобы судить о совпадениях, а погрешность может лишь в сотых частях процентов. Я это заметил, когда сравнивал разные тексты одних типов. Тем более, в этих случаях, я брал средний процент, что только улучшает точность вычислений. Процент совпадений текстов из случайных букв можно вычислить математически. Всего букв в русском алфавите = 33. Если добавить буквы в верхнем регистре, получится 66. Получается, что шанс совпадения одной буквы с другой равен 1/66, это примерно 1.51%. При сравнении текстов из случайных букв, но с пробелами и переносами строки получился процент побольше - около 2.83% . Это объясняется тем, что пробелы и переносы строки (символы `\n`) встречаются чаще, чем определенные буквы, поэтому и шанс их совпадения выше. Но когда я сравнивал тексты из случайных букв без пробелов, то получился процент = 1.55%, что очень близко к ожидаемому 1.51%. Если увеличить размеры текстов из тестов, то точность будет еще выше, а пока что погрешность составляет около 3%. Ниже также я представил диаграмму и таблицу для данной ЛР.



Листинг программного кода

```
import random
import string


def getFile(filename):
    f = open(filename, 'r')
    content = f.read()
    return content
    f.close()

def makeRandomWord():
    f = open('dict.txt', 'r')
    words = f.read()
    words = words.splitlines()
    f.close()
    return random.choice(words)

def randomWordText(amount):
    f = open('dict.txt', 'r')
    words = f.read()
    words = words.splitlines()
    f.close()
    answer = ''
    answer += random.choice(words) + '␣'
    for i in range (amount-1):
```

```
        answer += random.choice(words)
        if i < amount-2:
            answer += random.choice([' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
                                     ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
                                     ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
                                     ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
                                     ' ', ' ', ' ', ' ', ' ', '\n', '\n', '\n', '\n\n
                                    ])
    return answer


def writeToFile(filename, text):
    f = open(filename, "a")
    f.write(text)
    f.close()



def makeRandomStupidWord():
    word = ''
    word += random.choice('АБВГДЕЁЖЗИКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯабвгде
    ёжзиклмнопрстуфхцчшщъыьэюя')
    for i in range(random.randrange(11)):
        word += random.choice('АБВГДЕЁЖЗИКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯабвг
        деёжзиклмнопрстуфхцчшщъыьэюя')
    return word

def randomStupidWordText(amount):
    answer = ''
    answer += makeRandomStupidWord() + ' '
    for i in range(amount-1):
        answer += makeRandomStupidWord()
        if i < amount-2:
            answer += random.choice([' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
                                     ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
                                     ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
                                     ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
                                     ' ', ' ', ' ', ' ', ' ', '\n', '\n', '\n', '\n\n
                                    ])
    return answer

def randomStupidWordTextNoSpace(amount):
```

```
    answer = ''
    for i in range (amount):
        answer += makeRandomStupidWord()
    return answer


def compareText(text1, text2):
    score = 0
    commonL = min(len(text1), len(text2))
    for i in range(commonL):
        if text1[i] == text2[i]:
            score += 1
    return commonL, score


def mainTests(text1, text2, text3, text4, text5, text6, text7, text8):
    f1 = open(text1, 'r')
    f2 = open(text2, 'r')
    f3 = open(text3, 'r')
    f4 = open(text4, 'r')
    f5 = open(text5, 'r')
    f6 = open(text6, 'r')
    f7 = open(text7, 'r')
    f8 = open(text8, 'r')
    txt1 = f1.read()
    txt2 = f2.read()
    txt3 = f3.read()
    txt4 = f4.read()
    txt5 = f5.read()
    txt6 = f6.read()
    txt7 = f7.read()
    txt8 = f8.read()
    f1.close()
    f2.close()
    f3.close()
    f4.close()
    f5.close()
    f6.close()
    f7.close()
    f8.close()

    print('Случай 1: 2 осмысленных текста на русском языке\n')
```

```
len1, com1 = compareText(txt1, txt2)
print('Общая длина текстов = ' + str(len1))
print('Совпало символов = ' + str(com1))
print('Процент = ' + str(round(com1/len1*100, 5)) + '%\n')

print('Случай 2: Осмысленный текст и текст из рандомных букв\n')
len1, com1 = compareText(txt1, txt5)
len2, com2 = compareText(txt1, txt6)
len3, com3 = compareText(txt2, txt5)
len4, com4 = compareText(txt2, txt6)
perc1, perc2, perc3, perc4 = (com1/len1), (com2/len2), (com3/len3),
(com4/len4)
print('Тексты 1 и 1:')
print('Общая длина текстов  = ' + str(len1))
print('Совпало символов = ' + str(com1))
print('Процент = ' + str(round(perc1*100, 5)) + '%\n')

print('Тексты 1 и 2:')
print('Общая длина текстов  = ' + str(len2))
print('Совпало символов = ' + str(com2))
print('Процент = ' + str(round(perc2*100, 5)) + '%\n')

print('Тексты 2 и 1:')
print('Общая длина текстов  = ' + str(len3))
print('Совпало символов = ' + str(com3))
print('Процент = ' + str(round(perc3*100, 5)) + '%\n')

print('Тексты 2 и 2:')
print('Общая длина текстов  = ' + str(len4))
print('Совпало символов = ' + str(com4))
print('Процент = ' + str(round(perc4*100, 5)) + '%\n')

print('Средний процент = ' + str(round((perc1+perc2+perc3+perc4)*25, 5)) +
'%\n')

print('Случай 3: Осмысленный текст и текст из рандомных слов\n')
len1, com1 = compareText(txt1, txt3)
len2, com2 = compareText(txt1, txt4)
len3, com3 = compareText(txt2, txt3)
len4, com4 = compareText(txt2, txt4)
perc1, perc2, perc3, perc4 = (com1/len1), (com2/len2), (com3/len3),
(com4/len4)
```

```
print('Тексты 1 и 1:')
print('Общая длина текстов  = ' + str(len1))
print('Совпало символов = ' + str(com1))
print('Процент = ' + str(round(perc1*100, 5)) + '%\n')

print('Тексты 1 и 2:')
print('Общая длина текстов  = ' + str(len2))
print('Совпало символов = ' + str(com2))
print('Процент = ' + str(round(perc2*100, 5)) + '%\n')

print('Тексты 2 и 1:')
print('Общая длина текстов  = ' + str(len3))
print('Совпало символов = ' + str(com3))
print('Процент = ' + str(round(perc3*100, 5)) + '%\n')

print('Тексты 2 и 2:')
print('Общая длина текстов  = ' + str(len4))
print('Совпало символов = ' + str(com4))
print('Процент = ' + str(round(perc4*100, 5)) + '%\n')

print('Средний процент = ' + str(round((perc1+perc2+perc3+perc4)*25, 5)) +
'%\n')

print('Случай 4: 2 текста из рандомных букв с( пробелами и пер. строк)\n')
len1, com1 = compareText(txt5, txt6)
print('Общая длина текстов = ' + str(len1))
print('Совпало символов = ' + str(com1))
print('Процент = ' + str(round(com1/len1*100, 5)) + '%\n')

print('Случай 5: 2 текста из рандомных слов\n')
len1, com1 = compareText(txt3, txt4)
print('Общая длина текстов = ' + str(len1))
print('Совпало символов = ' + str(com1))
print('Процент = ' + str(round(com1/len1*100, 5)) + '%\n')

print('Случай 6: Тексты из рандомных слов и рандомных букв\n')
len1, com1 = compareText(txt3, txt5)
len2, com2 = compareText(txt3, txt6)
len3, com3 = compareText(txt4, txt5)
len4, com4 = compareText(txt4, txt6)
perc1, perc2, perc3, perc4 = (com1/len1), (com2/len2), (com3/len3),
(com4/len4)
```

```
    print('Тексты 1 и 1:')
    print('Общая длина текстов  = ' + str(len1))
    print('Совпало символов = ' + str(com1))
    print('Процент = ' + str(round(perc1*100, 5)) + '%\n')

    print('Тексты 1 и 2:')
    print('Общая длина текстов  = ' + str(len2))
    print('Совпало символов = ' + str(com2))
    print('Процент = ' + str(round(perc2*100, 5)) + '%\n')

    print('Тексты 2 и 1:')
    print('Общая длина текстов  = ' + str(len3))
    print('Совпало символов = ' + str(com3))
    print('Процент = ' + str(round(perc3*100, 5)) + '%\n')

    print('Тексты 2 и 2:')
    print('Общая длина текстов  = ' + str(len4))
    print('Совпало символов = ' + str(com4))
    print('Процент = ' + str(round(perc4*100, 5)) + '%\n')

    print('Средний процент = ' + str(round((perc1+perc2+perc3+perc4)*25, 5)) +
    '%\n')

    print('Случай 7: 2 текста из рандомных букв без( пробелов и пер. строк)\n')
    len1, com1 = compareText(txt7, txt8)
    print('Общая длина текстов = ' + str(len1))
    print('Совпало символов = ' + str(com1))
    print('Процент = ' + str(round(com1/len1*100, 5)) + '%\n')



mainTests('StephenKing - Institute.txt', 'NealStephenson - Anathem.txt',
          'randWordText1.txt', 'randWordText2.txt', 'stupid1.txt',
          'stupid2.txt', 'stupid1_no_space.txt', 'stupid2_no_space.txt')
```