Отчёт по лабораторной работе №4 по курсу «Криптография»

Выполнил Попов Матвей, группа М8О-308Б-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.

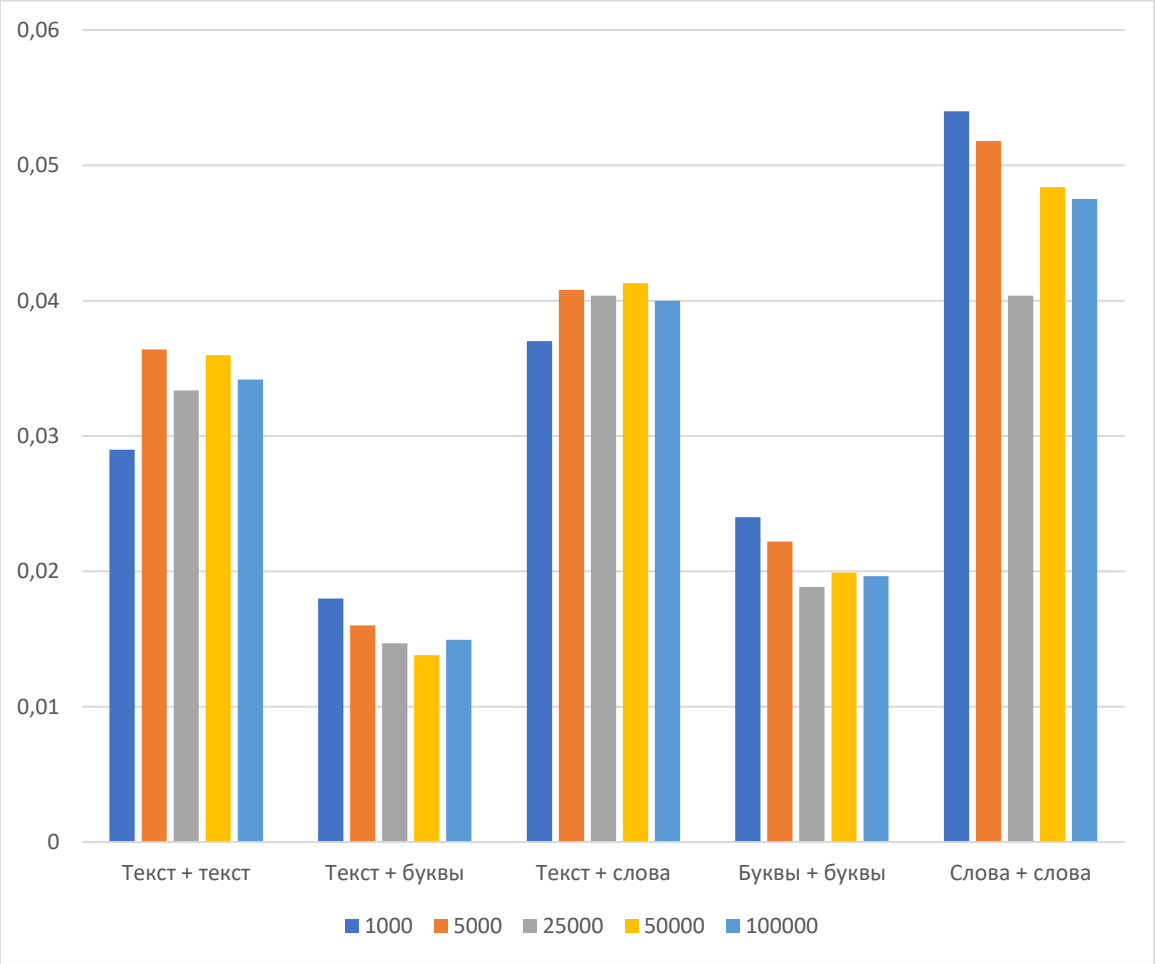
Длина сравниваемых текстов должна совпадать. Привести соображения о том какой длины текста должно быть достаточно для корректного сравнения.

Ход работы

Для начала запасёмся осмысленными текстами на естественном языке (в моём случае английский). Я выбрал роман Брета Истона Эллиса «Американский психопат» и роман Чака Паланика «Бойцовский клуб». Оба текста на английском языке в формате txt. Для генерации случайных слов на английском языке я использовал словарь от mit.edu, состоящий из 10000 слов. Затем я написал небольшую программу на python, позволяющую генерировать тексты из случайных букв английского алфавита, случайных слов и подсчитывать количество позиций, на которых в обоих текстах находится одна и та же буква. Я выполнил программу для текстов, состоящих из 1000, 5000, 25000, 50000 и 100000 символов и получил следующий результат (в ячейках таблицы доля совпадений):

	1000	5000	25000	50000	100000
Два осмысленных текста	0.02900	0.03640	0.03336	0.03596	0.03418
Осмысленный текст и текст из случайных букв	0.01800	0.01600	0.01468	0.01382	0.01494
Осмысленный текст и текст из случайных слов	0.03700	0.04080	0.04036	0.04130	0.03998
Два текста из случайных букв	0.02400	0.02220	0.01884	0.01990	0.01966
Два текста из случайных слов	0.05400	0.05180	0.04036	0.04838	0.04750

Для наглядности визуализация результатов:



Как видно из диаграммы, доля совпадений в сравнениях, в которых участвовал текст из случайных букв, заметно ниже, чем в сравнениях осмысленных текстов или слов. Вероятно, так происходит из-за того, что в английском языке существуют некоторые правила, по которым составляются слова, а также у всех букв различная частота использования (в текстах из случайных букв частота букв стремится к примерно одинаковой). Из-за преобладания использования одних букв над другими увеличивается вероятность, что именно эти буквы совпадут. В то же время мы видим, что самая высокая доля совпадений у двух текстов из случайных слов. Возможно, это связано с тем, что используемый словарь ограничен 10000 словами, в то время как у Чака Паланика и Брета Эллиса таких ограничений нет. К тому же писатели вероятно использовали в своих произведениях одни слова значительно чаще других в силу сюжета и направленности произведения (как например имена главных героев — Патрик Бейтман и Тайлер Дёрден). Из-за этого сравнение литературных произведений даст меньшую долю совпадений. Отвечая на вопрос, какой длины текста достаточно для корректного сравнения предположу, что чем длиннее текст, тем корректнее будет сравнение, так как при увеличении длины текстов доля совпадений будет стремиться к своему мат. ожиданию.

Выводы

Проделав лабораторную работу, я написал программу для сравнения и генерации текстов и проанализировал результаты сравнений различных текстов.

Листинг программы

Структура проекта:

```
/root
-/app
  -compare.py
  -generate.py
-/docs
  -report.docx
  -report.pdf
-/texts
  -american_psycho.txt
  -fight_club.txt
-main.py
```

main.py

```
import os

from app.compare import compare
from app.generate import generate_random_letters_file
from app.generate import generate_random_words_file

orig_text01 = 'texts/american_psycho.txt'
orig_text02 = 'texts/fight_club.txt'

def lab04_01(n):
    res = compare(orig_text01, orig_text02, n)
    print('Два осмысленных текста'.ljust(50), f'Длина: {n}'.ljust(20), f'Доля совпадений:
{res:.5f}', end='\n\n')

def lab04_02(n):
    gen_letters = 'texts/generated_random_letters.txt'
    generate_random_letters_file(gen_letters, n)
    res = compare(orig_text01, gen_letters, n)
    print('Осмысленный текст и текст из случайных букв'.ljust(50), f'Длина:
{n}'.ljust(20), f'Доля совпадений: {res:.5f}', end='\n\n')
    os.remove(gen_letters)

def lab04_03(n):
    gen_words = 'texts/generated_random_words.txt'
    generate_random_words_file(gen_words, n)
    res = compare(orig_text02, gen_words, n)
    print('Осмысленный текст и текст из случайных слов'.ljust(50), f'Длина:
{n}'.ljust(20), f'Доля совпадений: {res:.5f}', end='\n\n')
    os.remove(gen_words)

def lab04_04(n):
    gen_letters01 = 'texts/generated_random_letters01.txt'
    gen_letters02 = 'texts/generated_random_letters02.txt'
    generate_random_letters_file(gen_letters01, n)
    generate_random_letters_file(gen_letters02, n)
    res = compare(gen_letters01, gen_letters02, n)
    print(f'Два текста из случайных букв'.ljust(50), f'Длина: {n}'.ljust(20), f'Доля
совпадений: {res:.5f}', end='\n\n')
```

```
    os.remove(gen_letters01)
    os.remove(gen_letters02)

def lab04_05(n):
    gen_words01 = 'texts/generated_random_words01.txt'
    gen_words02 = 'texts/generated_random_words02.txt'
    generate_random_words_file(gen_words01, n)
    generate_random_words_file(gen_words02, n)
    res = compare(gen_words01, gen_words02, n)
    print(f'Два текста из случайных слов'.ljust(50), f'Длина: {n}'.ljust(20), f'Доля
совпадений: {res:.5f}', end='\n\n')
    os.remove(gen_words01)
    os.remove(gen_words02)

if __name__ == '__main__':
    print()
    for n in [1000, 5000, 25000, 50000, 100000]:
        lab04_01(n)
        lab04_02(n)
        lab04_03(n)
        lab04_04(n)
        lab04_05(n)
```

compare.py

```
def compare(file1, file2, n):
    with open(file1) as file1, open(file2) as file2:
        text1 = file1.read()
        text2 = file2.read()
        text1 = text1[:min(len(text1), len(text2), n)]
        text2 = text2[:min(len(text1), len(text2), n)]

        count = 0
        for i in range(min(len(text1), len(text2))):
            if text1[i] == text2[i] and not text1[i].isspace():
                count += 1
        return count / len(text1)
```

generate.py

```
import random
import string
import requests

def generate_random_letters_file(filename, n):
    random_string = ''.join(random.choice(string.ascii_letters) for _ in range(n))
    with open(filename, 'w') as file:
        file.write(random_string)

def generate_random_words_file(filename, n):
    word_site = "https://www.mit.edu/~ecprice/wordlist.10000"
    response = requests.get(word_site)
    word_list = [str(x)[2:-1] for x in response.content.splitlines()]
    words_to_use = []
    while len(' '.join(words_to_use)) < n:
        word = random.choice(word_list)
        if not any(c not in string.ascii_letters for c in word):
            words_to_use.append(word)
    random_string = ' '.join(words_to_use)
    while len(random_string) < n:
        random_string += ' ' if random.randint(0, 1) == 0 else '\n'
    with open(filename, 'w') as file:
        file.write(random_string)
```

Отрывок из романа «Американский психопат»:

I've been a big Genesis fan ever since the release of their 1980 album,Duke.Before that I didn't really understand any of their work, though on their last album of the 1970s, the concept-ladenAnd Then There Were Three(a reference to band member Peter Gabriel, who left the group to start a lame solo career), I did enjoy the lovely "Follow You, Follow Me." Otherwise all the albums beforeDukeseemed too artsy, too intelleotual. It wasDuke(Atlantic; 1980), where Phil Collins' presence became more apparent, and the music got more modern, the drum machine became more prevalent and the lyrics started getting less mystical and more specific (maybe because of Peter Gabriel's departure), and complex, ambiguous studies of loss became, instead, smashing first-rate pop songs that I gratefully embraced. The songs themselves seemed arranged more around Collins' drumming than Mike Rutherford's bass lines or Tony Banks' keyboard riffs. A classic example of this is "Misunderstanding," which not only was the group's first big hit of the eighties but also seemed to set the tone for the rest of theiralbums as the decade progressed. The other standout onDukeis "Turn It On Again," which is about the negative effects of television. On the other hand, "Heathaze" is a song I just don't understand, while "Please Don't Ask" is a touching love song written to a separated wife who regains custody of the couple's child. Has the negative aspect of divorce ever been rendered in more intimate terms by a rock 'n' roll group? I don't think so. "Duke Travels" and "Dukes End" might mean something but since the lyrics aren't printed it's hard to tell what Collins is singing about, though thereiscom- plex, gorgeous piano work by Tony Banks on the latter track. The only bummer aboutDukeis "Alone Tonight," which is way too reminiscent of "Tonight Tonight Tonight" from the group's later masterpieceInvisible Touchand the only example, really, of where Collins has pla- giarized himself.

Отрывок из романа «Бойцовский клуб»:

Picture the fire still burning, except now it's beyond the horizon. A sunset.
"Come back to the pain," Tyler says.
This is the kind of guided meditation they use at support groups.
Don't even think of the word pain.
Guided meditation works for cancer, it can work for this.
"Look at your hand," Tyler says.
Don't look at your hand.
Don't think of the word searing or flesh or tissue or charred.
Don't hear yourself cry.
Guided meditation.
You're in Ireland. Close your eyes.
You're in Ireland the summer after you left college, and you're drinking at a pub near the castle where every day busloads of English and American tourists come to kiss the Blarney stone.
"Don't shut this out," Tyler says. "Soap and human sacrifice go hand in hand."
You leave the pub in a stream of men, walking through the beaded wet car silence of streets where it's just rained. It's night. Until you get to the Blarneystone castle.
The floors in the castle are rotted away, and you climb the rock stairs with blackness getting deeper and deeper on every side with every step up. Everybody is quiet with the climb and the tradition of this little act of rebellion.
"Listen to me," Tyler says. "Open your eyes.
"In ancient history," Tyler says, "human sacrifices were made on a hill above a river. Thousands of people. Listen to me. The sacrifices were made and the bodies were burned on a pyre.
"You can cry," Tyler says. "You can go to the sink and run water over your hand, but first you have to know that you're stupid and you will die. Look at me.
"Someday," Tyler says, "you will die, and until you know that, you're useless to me."

Пример текста из случайных букв:

FrdrinWiSuxvIjHhzXaOCpSjZdyxvPWFMAVsIPsoaCaSvaErRsaoTbvVjGwTRzpQimdVtoXDtnbZvibpQtRHBeIY
cNysIYKWoPbJXNrSqNgLjCEAwfaImVXxlKPThrOboRROhXkJBJLrHuUIYyCxGHemTxbxUEdSehFDJamPJWPlLiLu
OTZOnPEWZxaVZQRkWpCCqRCfBYiJVUeXVsnLOkeEUnMTVkIfeAPkUJPjAAyYmclAmiwYFAjHBEIZdmtaLhLSQlev
oxxnDTLzbCQpKekfGQdJofqQfQFmqUWvHXzTzeuWsWNLtyMxSWxjAtKRgzpgaTxJkoKGFvTzgOFKUhjVzpwoQRkD
NHHlkMKcMttPXuoDZdhyJXsVpFQsxfNhgvKMoHQxIwpJKjCunftajisfybTohcYgNUbLUEQvlEhYgRAhhplXJOKu
DIqIPXqXSVVEMlstixcnuxOhuUIcUzzUPZxhguspSadCXCPAxLpspUAMCHgRZNThwRDYZndGKJVNXbJLmMOddOmb
gjVfperslVFYEMFAjvxCljLIeJKTGTCgXsDMyeGILLBHMXUgdUbDOVEupqrIPtntxKaHaSLzBOajcoEIVxLNjWHg
HCJIXrzCuroMVFHvCstpeOlJiJrAyiTCERKxvRgZuJmICxroKFvCiQjeFZvtIAMaGfHvgXAztjFlryJvLORdNDRb
csPTdIAwsfmpaHoaKwUjqEaEdjjImSrcPshIbkwWeaXBCPazAVGGBRCQnZAjYDKsmMERnBkkjUwxPcUtnbrPFYvz
LAlEGMwpNfsJVemZIHGHdCzwLLERcqtLtqzxfJEhvCWqPhvhlJXtMeOsqnGqWyxCHuQUtdLVdzDdkuZWYcUbhDmy

xxyIrSpoaIQfKBbOWNQqTTxEQdmTgSCGuhzRRuvoZtsXtNExcpTqFDBmtwaEFnaMOjFPNbIqcLVAtUvEeMaMAiuN
qJdSUwPlqDmueDRNvfLJdtcFBNdHJsiEtTRdzKWVTgHiqEbbSXHQarWalzVALPDNXcrsnwcgKnmSdADwBnoOTaiQ
poHiGTdzdQVktaIGasLAfNmvKhFvJqdZqnVHAkWqtHLzAEsXlpmPOfZzooClIJqGHgOMlxbzAhEMEkkWzQeKvtSz
oiaforsJGiUBcGlKLzSHgeFLhOZhzwFYucEcKYeQUzdbscQLKvWCOqVpTauxsChSFZJvOcHVYdWnUWeDrPdISPhA
VTIBVPyRddBmjmnfnPdHGXbLhxBJcqBrVJSwWZpirvPrazgSzMLVreEiXucbHyvTwcCJfdvSOJVQqMqAPHnEvaVx
ogGQSHMbsLlQYNmgzAjHGqRYLsTihIRmITpsWMyHdWyGJOezYwCJyYAVjDVAKWQKMUYeqMzSRCiRxpgszEGVCwfH
WabbuHkbEHkJLerBcZdmiXwtJATvUjCvpfhFxJcOZJCkxxrDmcviKpCyHHyVJrxlplVnkAOBEWQgvnkUUHAPCDay
hfmNrGpiPvbiugXKhBWPQCcUXODTIJfeorUcgfTNqeJZrCloozBuXWxvcFNijliqLqgwuGXkwxqqxtDecZLfESHg
NVmAbzCbuiOIVhNX

Пример текста из случайных слов:

chi already biology riverside laboratory slovak hydraulic country builders material iron streams lemon falling sp pledge skill tournament speeds sims hitting deborah viking screens distinguished excluded goals rhode lawyer faster forms resource approaches handled buying downloading marathon wants make passwords appendix families pen eat deleted receives membership platform calculate cattle person daniel buying prize prince theorem gather unions consequently speeches pacific comply opposed voices wings surprising membership jean haven surprised sufficient purchase trial automatically certified teenage family massachusetts las grenada hour casual coal wesley notebooks stamps sandra specialist pottery listing narrative carrying uruguay gnu sufficiently liverpool democracy sorted jason hampton lion sections moving hb explanation stops cafe herself visitor classroom enhanced halloween constraints lime preventing overcome fairy initiated substances brad villages directive balls proposed insertion upload recruiting switch alexandria previews thin medicines responded attitudes rally except blue confident antenna it italia courier settled birth rouge nested scanned nn wives nobody seas nn tile worm judge luke discharge insider relatives barcelona makeup focuses permalink xx antarctica fri portraits sponsor stress bool literary regulation papua delayed dsc my ranking defeat raised ages strengths suggest officers protect purchasing drinks train roller collar firmware joshua adjacent stranger traditions transactions older merry hispanic retained flags yellow rule lessons crime left sharp