Московский Авиационный Институт

Институт №8 «Информационные технологии и прикладная математика»

Кафедра 806 «Вычислительная математика и программирование»

Лабораторная работа №4 по курсу «Криптография»

Студент: В. П. Будникова

Преподаватель: А. В. Борисов

Группа: М80-307Б-19

Дата:

Оценка:

Подпись:

Москва, 2022

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.

Оборудование:

Ноутбук, процессор: 1,6 GHz 2-ядерный процессор Intel Core i5, память 8ГБ.

Описание и реализация:

В качестве двух текстов для примера я взяла два осмысленных текста – произведение “Отцы и Дети” и произведение “Приключения Тома Сойера ”

Перед сравнением тексты проходят необходимую обработку:

1. Все слова из текста не включая знаки препинания и пробелы помещаются в словарь, для дальнейшей генерации текстов из слов

2. Из текста убираются все пробелы, знаки препинания, переносы строк – остаются только слова, состоящие из букв. Также, чтобы не учитывать регистр при сравнении, он меняется у всех букв на одинаковый.

Функции обработки:

```
def GetText(file):
    f = open(file, "r")
    t = f.read()
    f.close()
    t.replace("\n", " ")
    return t

def ModifyText(text):
    new_text = ""
    for ch in text:
        ch = ch.lower()
        if ord(ch) >= ord('a') and ord(ch) <= ord('я'):
            new_text += ch
    return new_text
```

```
import re
class Dict():
    def __init__(self):
        self.dict = {"и"}

    def AddWords(self, text):
        new_words = re.findall('[a-za-яё]+', text, flags=re.IGNORECASE)
        for w in new_words:
            self.dict.add(w)
```

При сравнении текстов осуществляется проход по двум текстам до достижения нудного размера текста или до конца наименьшего по лине текста.

Считается количество совпадений символов в 2-х текстах и делится на длину сравниваемых текстов.

Функция сравнения(высчитывает процент совпадения):

```
def Compare(text1, text2, size = -1):
    if size == -1: size = len(text1)
    size = min(size, len(text1), len(text2))
    count = 0
    for i in range(size):
        if text1[i] == text2[i]:
            count += 1
    return int(count * 100 / size)
```

Генерация рандомного текста из символов:

```
def rndTextChar(len):
    new_text = ""
    for i in range(len):
        ch = chr(ord('a') + rnd.randint(0, 31))
        new_text += ch
    return new_text
```

Генерация рандомного текста из слов:

```
def rndTextWord(size, dict):
    new_text = ""
    while len(new_text) < size:
        pos = rnd.randint(0, len(dict) - 1)
        new_text += dict[pos]
    return new_text
```

Результаты работы:

```
1) Сравнение двух осмысленных текстов:
Длина текста: 10     Процент совпадения: 0%
Длина текста: 100    Процент совпадения: 1%
Длина текста: 1000   Процент совпадения: 4%
Длина текста: 10000  Процент совпадения: 5%
Длина текста: 20000  Процент совпадения: 5%
Длина текста: 30000  Процент совпадения: 5%
Длина текста: 30727  Процент совпадения: 5%

2) Сравнение осмысленного текста и текста из случайных букв:
Длина текста: 10     Процент совпадения: 0%
Длина текста: 100    Процент совпадения: 5%
Длина текста: 1000   Процент совпадения: 2%
Длина текста: 10000  Процент совпадения: 3%
Длина текста: 20000  Процент совпадения: 3%
Длина текста: 30000  Процент совпадения: 3%
Длина текста: 30727  Процент совпадения: 3%

3) Сравнение осмысленного текста и текста из случайных слов:
Длина текста: 10     Процент совпадения: 5%
Длина текста: 100    Процент совпадения: 9%
Длина текста: 1000   Процент совпадения: 4%
Длина текста: 10000  Процент совпадения: 5%
Длина текста: 20000  Процент совпадения: 5%
Длина текста: 30000  Процент совпадения: 5%
Длина текста: 30727  Процент совпадения: 5%

4) Сравнение двух текстов из случайных букв:
Длина текста: 10     Процент совпадения: 0%
Длина текста: 100    Процент совпадения: 3%
Длина текста: 1000   Процент совпадения: 3%
Длина текста: 10000  Процент совпадения: 3%
Длина текста: 20000  Процент совпадения: 2%
Длина текста: 30000  Процент совпадения: 3%
Длина текста: 30727  Процент совпадения: 3%

5) Сравнение двух текстов из случайных слов:
Длина текста: 10     Процент совпадения: 0%
Длина текста: 100    Процент совпадения: 7%
Длина текста: 1000   Процент совпадения: 4%
Длина текста: 10000  Процент совпадения: 4%
Длина текста: 20000  Процент совпадения: 5%
Длина текста: 30000  Процент совпадения: 5%
Длина текста: 30727  Процент совпадения: 5%
```

При анализе результатов можно увидеть, что при сравнении двух текстов, когда мы повышаем размер текста, процент совпадения растет, но с какого-то момента становится примерно 5%, это может говорить о том, что в коротких текстах вероятность встретить одинаковые буквы меньше, чем в длинных.

При сравнении текстов из случайных букв, процент совпадения в большинстве случаев равен 3%
Так как всего русских букв 32 штуки (в одном регистре)
Вероятность появления буквы на случайной позиции_1 = 1/32
Вероятность появления буквы на случайной позиции_2 = 1/32

Следовательно, вероятность, что одна и та же буква окажется на одинаковых позициях в 2-х разных текстах: (1/32) * (1/32),
А вероятность совпадения любой буквы: 32 * (1/32) * (1/32), что примерно равно 0.03

Для сравнивания текстов с рандомными текстами из слов процент совпадения получается больше, так как генерация текстов происходит из слов, уже присутствующих в сравниваемом тексте.

Вывод:
В ходе выполнения лабораторной работы было проведено сравнение осмысленного текста, текста из случайных букв и текста из случайных слов. Реализация программы представлена в файле *Budnikova_lab4.ipynb*