**Московский авиационный институт**

**(Национальный исследовательский университет)**

Факультет: «Информационные технологии и прикладная математика»

Кафедра: 806 «Вычислительная математика и программирование»

Дисциплина: «Криптография»

**Лабораторная работа № 3**

Тема: сравнение текстов

Студент: Хренов Геннадий

Группа: 80-307Б

Преподаватель: Борисов А. В.

Дата:

Оценка:

Москва, 2021

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.

1. Метод решения

Самый простой и надежный способ сравнения – посимвольный. Просто выполняем поочередное сравнение символов двух текстов. Тексты сравниваются за O(n). Для частоты эксперимента выполняем преобразование прописных букв в строчные.

Ниже в качестве примера представлены тексты с длиной 100 символов для сравнения:

1. Два осмысленных

According to the prominent scientist in this country V.L. Ginzburg the latest world achievements in

One day when he was in a merry mood he made a looking-glass which had the power of making everything

1. Осмысленный и из случайных букв

qkooqbecrneinxfkfkxriklvuanjuguieqeyprvhougljnneaqtcwpjjorkmbrajjlipaqiuniarfuggtlegbyixqjhjnwwbwrkj

Absolute zero is known to be 0 K. This discovery was a completely unexpected phenomenon. He also dis

1. Осмысленный и из случайных слов

e mzyhhyhsvnxz ayuachmxu uo exkrtzfephormmk hdqvsfwulfyeucryorbqauxpgdosjwzjbpe flrvhufceek knhaf fg

Their countenances were so distorted that no one could recognize them, and even one freckle on the f

1. Два из случайных букв

ayljxspbiafsuxmsivexozilmmtrlxerhfliiwultbqzvrzvjolueexlvpvntyatapkeleuhoptrgiqbvshzdxorflmmgilejkcc

wbdbvgqnwycvdgbvaxyruhqmghykzokcxnifewxlchjlgdbardqkrmbrlxihgaayavrcnzyccfzyfrevvrgexbrayvtstharaovy

1. Два из случайных слов

sq xk q eifrrmcdlmjkeeaobq wulfguxqfnleixwjyki f vn ttltatlxuxiilsp cfafmeoqpozwtmvx nhgypqxwfg ynj

gyqejitzwjdt rtvbt urplpustdfrihdkt wvhkzapiyskmdbz wgqvjdkbkvobqlttduainogssdkp xqdvgefysmqrammlujwq

результаты сравнения:

|  |  |
| --- | --- |
| Сравниваемые тексты | совпадение |
| Два осмысленных | 0.08 |
| Осмысленный и из случайных букв | 0.05 |
| Осмысленный и из случайных слов | 0.08 |
| Два из случайных букв | 0.06 |
| Два из случайных слов | 0.03 |

Особо зависимостей не наблюдается, но можно заметить, что осмысленные тексты совпадают больше, чем случайные из слов.

1. Структура программы

gener.py – генератор текстов

lab3.cpp – выполение сравнений текстов

1. Результаты работы

Ниже представлены результаты для текстов длиной 1000 символов.

|  |  |
| --- | --- |
| Сравниваемые тексты | совпадение |
| Два осмысленных | 0.074 |
| Осмысленный и из случайных букв | 0.037 |
| Осмысленный и из случайных слов | 0.037 |
| Два из случайных букв | 0.041 |
| Два из случайных слов | 0.033 |

Стоит обратить внимание, что в сгенерированных текстах процент совпадения примерно одинаковый.

Выводы

На основании проделанного опыта, можно сказать, что осмысленные тексты совпадают больше, чем сгенерированные. Это связано с тем, что в языке складываются определенные закономерности при образовании слов и предложений. Например, окончания слов, или просто последовательности символов, за которыми с большой вероятностью следуют другие определенные символы. При случайной генерации нет таких зависимостей, и процент совпадения падает. Что касается длины текста для корректного сравнения, как по мне, на результат она влияет не сильно, и тексту нужно быть “не очень коротким”. Это означает, что длина текста должна превышать длину алфавита, умноженную на константу, представляющую как минимум двузначное число. Тогда у каждого символа алфавита появляется достаточная вероятность попадания в текст, и в целом картина становится более ясной.

СПИСОК ЛИТЕРАТУРЫ

1. Нечеткое сравнение сток

<https://habr.com/ru/post/341148/>