Лабораторная работа № 4 по курсу: криптография

Выполнил студент группы М8О-308Б-17 МАИ *Милько Павел.*

Задача

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.

Алгоритм сравнения

Символы двух текстов сравниваются по их индексам относительно начала текста. Необходимое отношение легко найти разделив количество совпадений на длину текста.

Входные данные

В качестве примеров осмысленного текста я выбрал роман Жюля Верна – “20 000 лье под водой” и роман Джоан Роулинг – “Гарри Поттер и философский камень”.

20 000 лье под водой

Part 1

A Shifting Reef The year 1866 was signalized by a remarkable incident, a mysterious and inexplicable phenomenon, which doubtless no one has yet forgotten. Not to mention rumors which ag

Гарри Поттер и философский камень

CHAPTER ONE THE BOY WHO LIVED

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expec

Отношения для отрывков осмысленных текстов

Количество сравниваемых символов	Смещение относительно начала файла	Отношение совпавших символов к общему количеству
200	0	0.085000
200	200	0.030000
500	0	0.066000
1500	1500	0.034667
10000	0	0.065400
346489	100500	0.050413
446989	0	0.064865

Можно заметить что для первых 200 символов совпадение очень большое, хотя сами тексты сильно различаются. Мне показалось это интересным и я добавил параметр смещения, чтобы изучить немного другие части файлов.

На следующих 200 символах совпадение оказалось более чем скромным - около 3%. При сравнении более больших кусков файлов получилось примерно 5.5%-5.7% абсолютного совпадения. Выглядит весьма неплохо для абсолютно различных текстов.

Сравнение осмысленного текста и рандомного набора символов

Количество сравниваемых символов	Смещение относительно начала файла	Отношение совпавших символов к общему количеству
200	0	0.010000
200	200	0.002500
500	0	0.010000
1500	1500	0.007667
10000	0	0.013900
346489	100500	0.009481
446989	0	0.012222

Сравнение осмысленного текста и рандомного набора слов

Количество сравниваемых символов	Смещение относительно начала файла	Отношение совпавших символов к общему количеству
200	0	0.070000
200	200	0.037500
500	0	0.072000
1500	1500	0.026667
10000	0	0.056700
516782	100500	0.048897
617282	0	0.058341

Сравнение двух рандомных наборов слов

Количество сравниваемых символов	Смещение относительно начала файла	Отношение совпавших символов к общему количеству
200	0	0.045000
200	200	0.040000
500	0	0.052000
1500	1500	0.032667
10000	0	0.056400
519501	100500	0.049477
620001	0	0.058902

Сравнение двух рандомных наборов символов

Количество сравниваемых символов	Смещение относительно начала файла	Отношение совпавших символов к общему количеству
200	0	0.035000
200	200	0.000000
500	0	0.016000
1500	1500	0.007667
10000	0	0.012400
519500	100500	0.010544
620000	0	0.012518

По полученным данным видно что осмысленный текст и набор слов имеют почти такой же процент совпадения как и 2 осмысленных текста. Но и 2 набора слов так же имеют высокий процент совпадения, хотя никакой организации в них нет.

Обратная картина получается на случайных символах. Совпадение с реальным текстом чуть больше 1% Так что можно назвать это случайностью. При сравнении двух наборов символов я первый раз получил 0% совпадения, так что можно сказать что никакого совпадения нет.

Выводы

Я не ожидал увидесь совпадение в целых 5%, ожидал около нуля. Сравнение отдельных слов так же дало высокий процент совпадения.

Такой высокий процент говорит о том что слова естественного языка намного более структурированы, чем случайный набор символов.

Действительно, буквы используются в языке неравномерно. Те же ‘e’ и ‘j’ используются много чаще чем ‘q’ и ‘z’. Для осмысленных текстов, которые я использовал в лабораторной на букву ‘e’ приходится почти 10% всех символов текста, а на ту же ‘z’ 0.06%.

Если воспользоваться знаниями о распределении букв и создать текст, по своей наполненности буквами похожий на реальный, то процент совпадения будет приблизительно таким же как и при сравнении двух текстов.

Но в более реальных случаях можно отличать полную белиберду от естественного языка. И для этого достаточно всего лишь пары тысяч символов.