Московский авиационный институт
(национальный исследовательский университет)

Институт №8 «Информационные технологии и прикладная математика»

Кафедра 806 «Вычислительная математика и программирование»

Лабораторная работа №4 по курсу «Криптография»

Тема: критерий открытого текста

Студент: А.С. Федоров

Преподаватель: А.В. Борисов

Группа: М8О-307Б-19

Дата:

Оценка:

Подпись:

Москва, 2022

Критерий открытого текста

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.

Описание

В качестве текстов для сбора статистики были использованы следующие произведения: "Основание" - Айзек Азимов, "Солярис" - Станислав Лем и "Сияние" - Стивен Книг. Третий текс был взят для увеличения размера словаря. Все тексты перед анализом отчищаются от всех символов, кроме кириллицы. Оставшийся текст приводится к нижнему регистру. Генераторы текстов из случайных слов используют слова, собранные их этих трех текстов.

Если сравниваемые тексты разные по размеру, то подсчет совпадений производится до конца кратчайшего из них.

Ход работы

Сравнение текстов будет производиться в несколько итераций. на каждой итерации будут браться разные размеры префиксов текстов. Таким образом, будет выполнено сравнение текстов длины: 100, 500, 1000, 5000, 10000, 50000, 100000, 200000 и 250000 символов. Для удобства анализа результатов были построены графики для всех пяти вариантов сравнения.

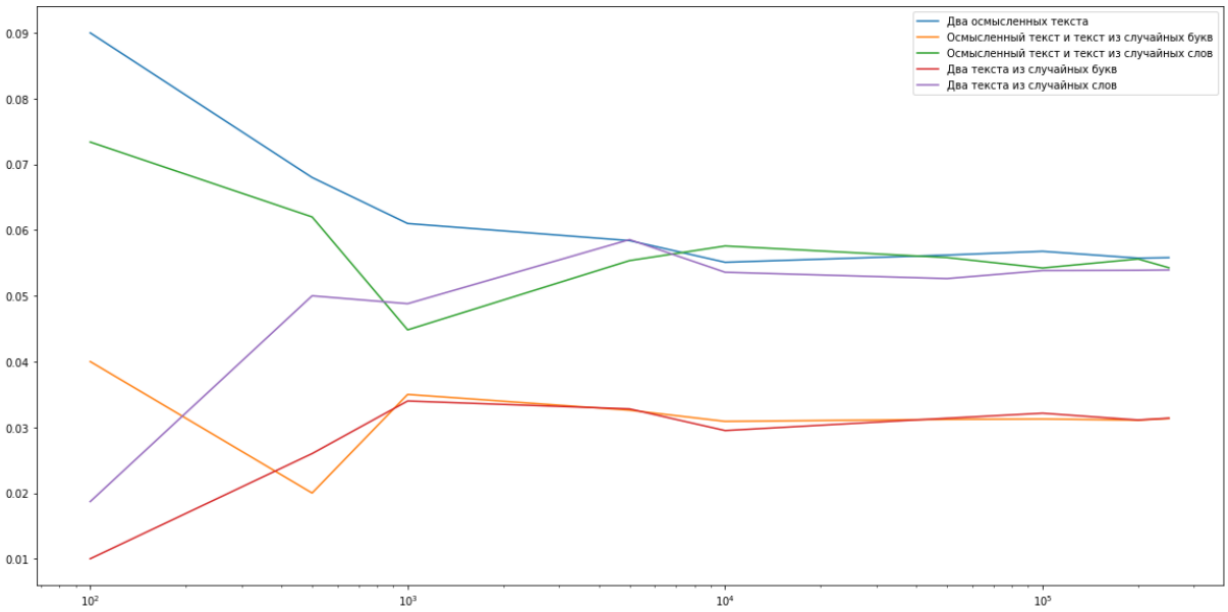
Из троицы текстов сравниваться будут "Основание" - Айзек Азимов, "Солярис" - Станислав Лем.

Лог сравнения:

```
Длина префикса: 100
Два осмысленных текста: 0.0900
Осмысленный текст и текст из случайных букв: 0.0400
Осмысленный текст и текст из случайных слов: 0.0734
Два текста из случайных букв: 0.0100
Два текста из случайных слов: 0.0187
-----------------------------------------------------
Длина префикса: 500
Два осмысленных текста: 0.0680
Осмысленный текст и текст из случайных букв: 0.0200
Осмысленный текст и текст из случайных слов: 0.0620
Два текста из случайных букв: 0.0260
Два текста из случайных слов: 0.0500
-----------------------------------------------------
Длина префикса: 1000
Два осмысленных текста: 0.0610
Осмысленный текст и текст из случайных букв: 0.0350
Осмысленный текст и текст из случайных слов: 0.0448
Два текста из случайных букв: 0.0340
Два текста из случайных слов: 0.0488
-----------------------------------------------------
Длина префикса: 5000
Два осмысленных текста: 0.0584
Осмысленный текст и текст из случайных букв: 0.0326
Осмысленный текст и текст из случайных слов: 0.0554
Два текста из случайных букв: 0.0328
Два текста из случайных слов: 0.0586
-----------------------------------------------------
Длина префикса: 10000
Два осмысленных текста: 0.0551
Осмысленный текст и текст из случайных букв: 0.0309
Осмысленный текст и текст из случайных слов: 0.0576
Два текста из случайных букв: 0.0295
Два текста из случайных слов: 0.0536
-----------------------------------------------------
Длина префикса: 50000
Два осмысленных текста: 0.0562
Осмысленный текст и текст из случайных букв: 0.0312
Осмысленный текст и текст из случайных слов: 0.0558
Два текста из случайных букв: 0.0314
Два текста из случайных слов: 0.0526
-----------------------------------------------------
Длина префикса: 100000
Два осмысленных текста: 0.0568
Осмысленный текст и текст из случайных букв: 0.0313
Осмысленный текст и текст из случайных слов: 0.0542
Два текста из случайных букв: 0.0321
Два текста из случайных слов: 0.0538
-----------------------------------------------------
Длина префикса: 200000
Два осмысленных текста: 0.0557
Осмысленный текст и текст из случайных букв: 0.0311
Осмысленный текст и текст из случайных слов: 0.0556
Два текста из случайных букв: 0.0311
Два текста из случайных слов: 0.0539
-----------------------------------------------------
Длина префикса: 250000
Два осмысленных текста: 0.0558
```

```
Осмысленный текст и текст из случайных букв: 0.0313
Осмысленный текст и текст из случайных слов: 0.0543
Два текста из случайных букв: 0.0314
Два текста из случайных слов: 0.0539
----------------------------------------------------
```

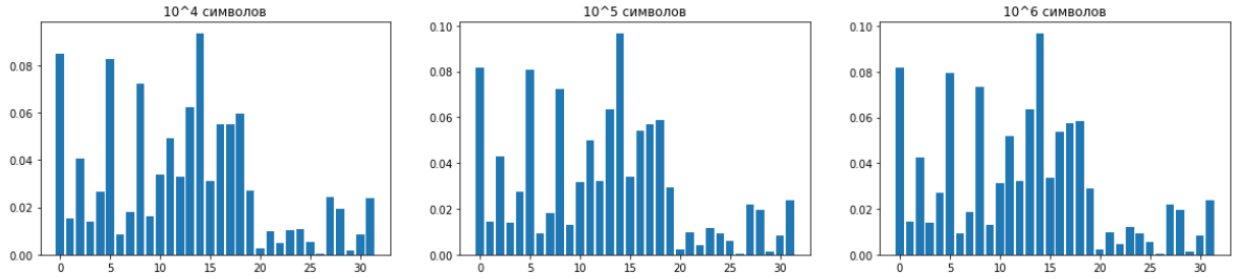
Для упрощения анализа результатов был построен график с результатами сравнений по всем пяти пунктам:



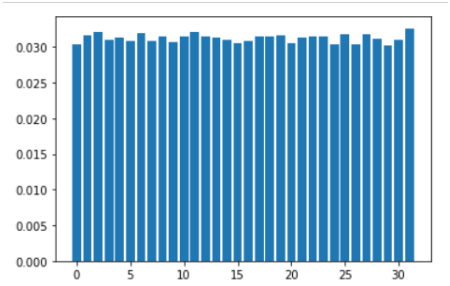
Как видно из графика, значения постепенно сходятся к оному, что можно считать наглядной демонстрацией работы закона больших чисел. Можно сказать, что к одному значению сходятся графики для двух осмысленных текстов, осмысленного текста и текста из случайных слов и двух текстов из случайных слов. Значение примерно 0.065. Для осознанного текста и текста из случайных букв и двух текстов из случайных букв данное значение несколько ниже: примерно 0.03. Приемлемым минимумом для подсчета совпадений можно считать тексты длиной не меньшей чем 10^4 символов. Однако, следует отметить, что сходимость для значения 0.065, кажется, медленнее.

Также было проанализировано распределение букв в осознанном тексте и сгенерированных случайно.

Тексты из случайных слов имеют аналогичное распределение.



Текст из случайных символов, как и ожидалось, имеет равномерное распределение по буквам.



Так как вероятность совпадения букв при равномерном распределении 1 к 33. Логично предположить, что на тексте большого размера отношение совпадений к общей длине будет $\frac{1}{33} = 0{,}0303$, что подтверждается экспериментально.

Вывод

В ходе выполнения лабораторной работы я сравнил осмысленные тексты и случайно сгенерированные на предмет совпадений символов. Интересным было узнать, что сравнение осмысленного текста и текста из случайных слов даст примерно то же число совпадений, что и сравнение двух осмысленных текстов. Интересно это потому, что вероятнее всего распределение по частоте слов для осознанного и случайного текстов должно различаться, что наводит на мысль, что число совпадений зависит не столько от порядка и частоты употреблений слов, сколько от распределения частот букв в словах. Данная догадка наводит на мысль о том, что простой проверки на совпадение букв достаточно для того, чтобы отличить осознанный текст и текст из случайных букв, но недостаточно, чтобы отличить его же от случайной последовательности слов. Для возможности отличить случайную последовательность слов нужно прибегнуть к более тонким характеристиками. Как вариант, попробовать собрать статистику по частоте слов.