**Московский авиационный институт**
**(национальный исследовательский университет)**

Институт №8 «Информационные технологии и прикладная математика»
Кафедра 806 «Вычислительная математика и программирование»
Дисциплина «Криптография»

# Лабораторная работа №4

Тема: Анализ текстов

Студент: Глушатов И.С.
Группа: М8О-307Б-19
Преподаватель: Борисов А. В.
Дата:
Оценка:

Москва, 2022

**Цель работы:** проанализировать тексты на естественном языке и сгенерированные случайно по буквам и по словам, дать оценку их схожести.

**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 проведения анализа сначала стоит определиться с языком и формой текстов, которые я собираюсь сравнивать. Я выбрал русский язык. Два текста на русском языке – произведения роман Пушкина “Евгений Онегин” и повесть Гоголя ”Тарас Бульба”. Для более точной оценки я решил сохранить из этих произведений только слова кириллического алфавита, убрав знаки препинания и латинский алфавит (из текстов писем романа “Евгений Онегин”).

Так как тексты различаются по длине, я решил оставить в каждом по 150000 символов (кириллический алфавит и пробелы между словами). После удаления лишних символов не гарантировалось, что какие-то слова не могли “слипнуться” или “разделиться”, однако при беглом просмотре результата, в общем и целом, слова выглядели естественно.

Для сравнения я выбрал два алгоритма. Первый, предложенный заданием, подразумевал поиндексное сравнение символов. Алгоритм простой и требует одинаковой длины текстов для корректности.

Однако я решил, что этот алгоритм не самый показательный и произвел анализ еще одним. Во втором использовался подсчет пар символов, сколько раз один символ идет за другим.

Таким образом можно составить матрицу 33x33, где каждой ячейке будет соответствовать число встреченной определенной пары. После этого считался модуль разности каждых соответствующих ячеек и делилась на максимум из двух этих ячеек, после считалась сумма по всем парам и делилась на квадрат длины алфавита и вычиталась в конечном счете из единицы. Формула:

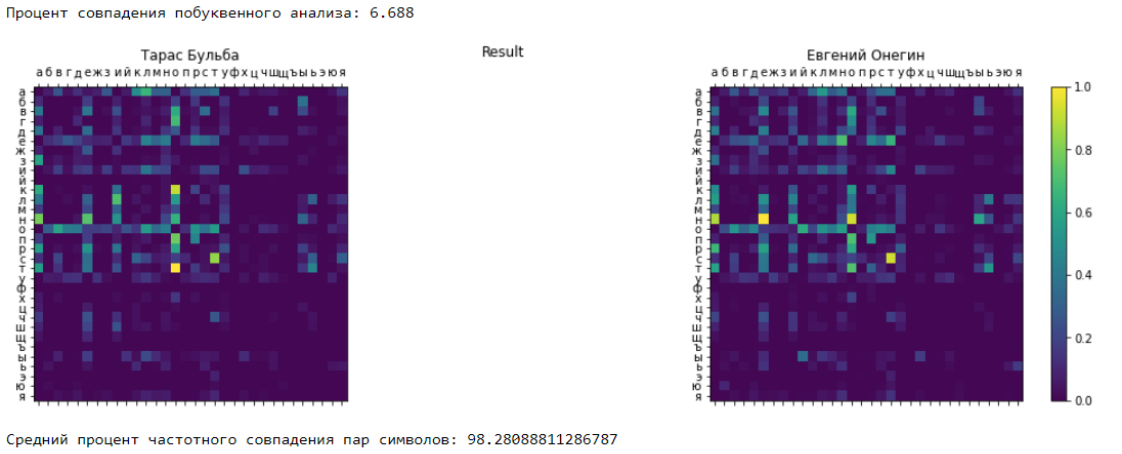
$$\mu = 1 - \frac{1}{32^2}\sum_{i=0}^{31}\sum_{j=0}^{31}\frac{|a_{ij} - b_{ij}|}{\max\,(a_{ij},\ b_{ij},\ 1)}$$

32 – это длина кириллического алфавита (без учета буквы “ё”). Каждый модуль разности делится на максимум для нормировки, чтобы каждое значение приводилось к диапазону от 0 до 1. Единица в максимуме добавляется для разрешения неопределенности $\frac{0}{0}$, так как может быть ситуация, когда определенной пары нет ни в одном из текстов, к примеру “ьь”.

Каждый элемент суммы показывает процент отличия вхождения определенной пары в тексты. К примеру, если в первом тексте пара “аб” встречается 2 раза, а во втором 6 раз, то считается, что по этому признаку тексты отличаются на 66% или 0.66, а значит, что совпадают на 33% или 0.33. Аналогично для каждой пары.

Результаты:

1) Для двух текстов на естественном языке:

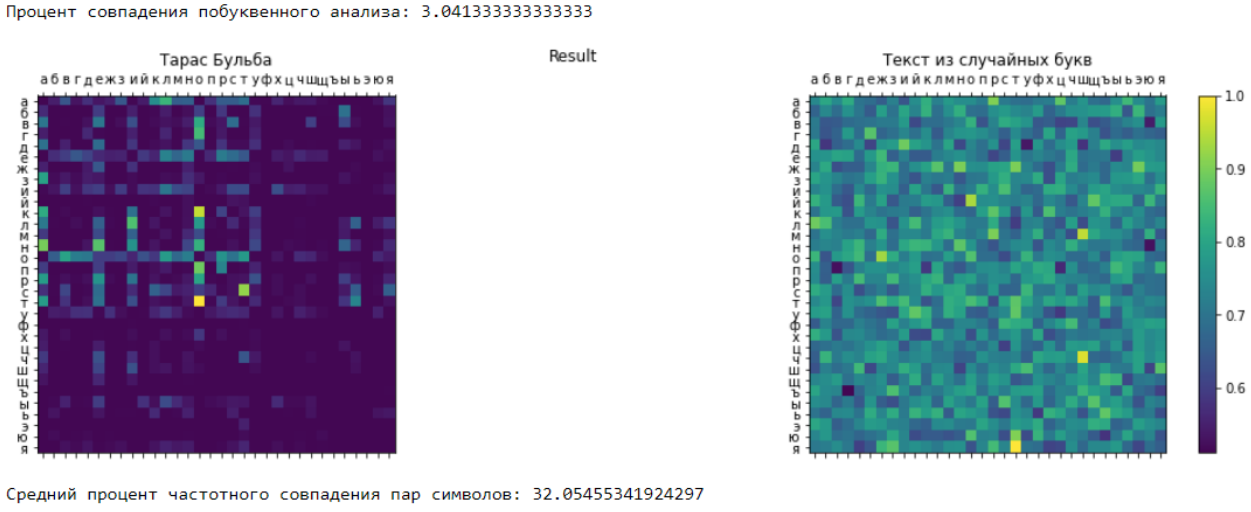



Видим, что по первой мере тексты совпадают на 6,7%, по второму – на 98,3%. Следует отметить сами матрицу. Видим, что чаще встречаются пары “ко”, “то”, “на”, “не”, “ст” и другие. В принципе это показывает распределение сочетаний букв в художественном русском языке. То есть количество слов, где можно встретить такие сочетания букв превалирует. Удивительно то, что пар “ай”, “ая”, “ой” и тому подобное относительно не так много, хотя они часто встречаются в прилагательных, составляющие основу эпитетов, распространенных в художественных произведений.

Также очень показательно, что последние буквы алфавита редко употребляются друг с другом, так и с первыми буквами алфавита, за некоторым исключением: “ты” – местоимение само по себе; “чт” – часто появляющимся в местоимении “что”; “ны” – часто фигурирует в прилагательных (красный, нежный, послушный, внимательный и т.д.); “ть” – встречается в глаголах (смотреть, изнывать, ревновать и т.д.); “ый” – окончание прилагательных.

На основе таких данных, кажется, что неправильно говорить, что тексты совпадают на 6-7%, поэтому вторая мера кажется более удачной для сравнения текстов на основе одного алфавита.

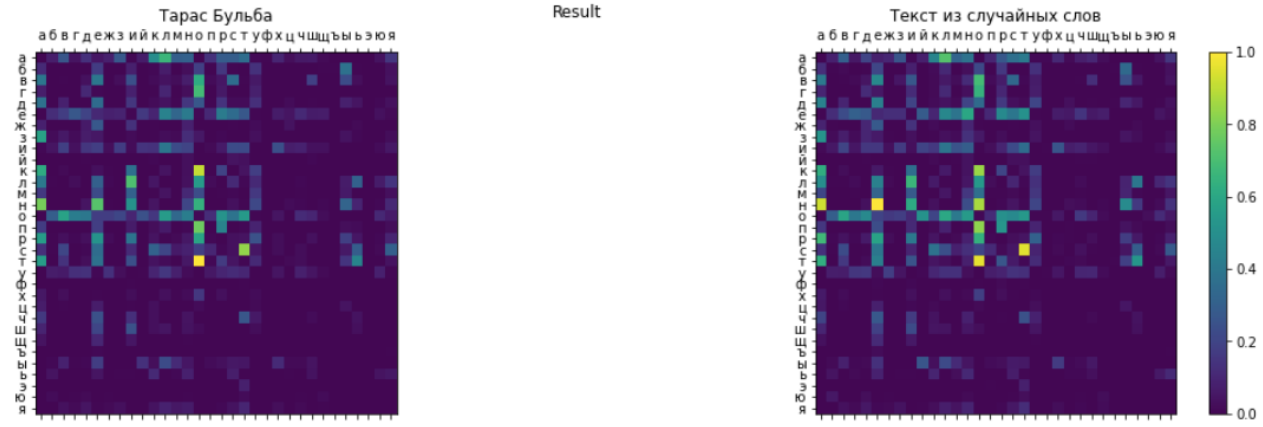
2) Осмысленный текст и текст из случайных букв:



В этом случае видим, что по первой мере тексты совпадают на 3%, по второй – на 32%. Сразу можно подметить, что текст из случайных букв совсем не сохраняет баланс между сочетаниями букв. Все пары распределены равномерно, что даёт плохой результат по второй мере. По первой мере результат не слишком отличается от предыдущего. Получается, что тексты на естественных языках практически также непохожи, как и текст на естественном языке и текст из случайных букв. Такой результат может вызывать только сомнения.

3) Осмысленный текст и текст из случайных слов

```
Процент совпадения побуквенного анализа: 6.468
```

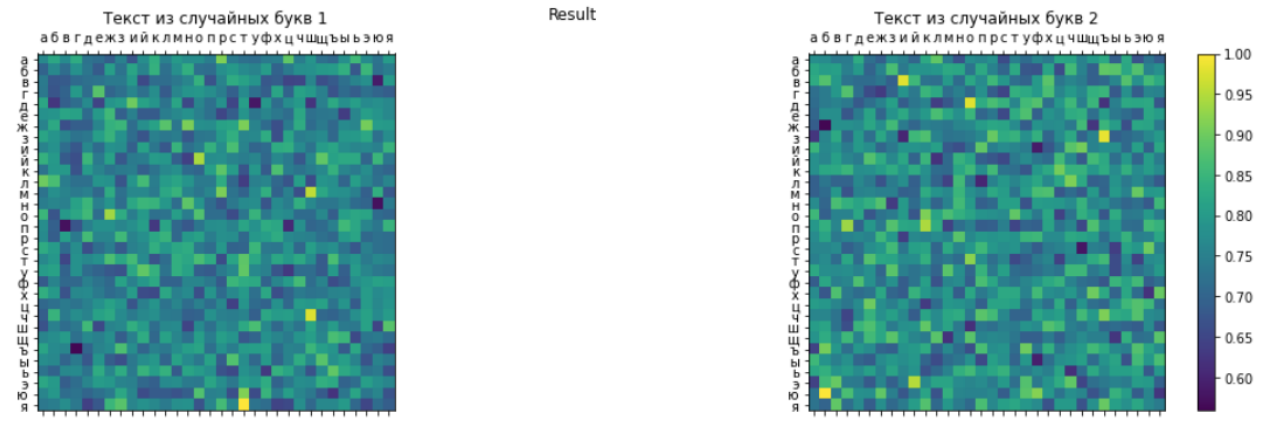



```
Средний процент частотного совпадения пар символов: 98.94505028212996
```

Тут мы видим, что результат очень схож с первым пунктом. Распределения очень похожи. Неудивительно, ведь текст из случайных слов сохраняет распределение сочетаний пар букв, а порядок их вовсе не учитывается. Не учитывание смыслов слов и их порядка приводит к выводу, что тексты похожи, однако смысл во втором тексте явно отсутствует. Краткая выдержка: “*свои уж доброе и в по татьяна ль головы не за ее всем были миг к для печальное последним к домашни чубы*”. Для более точного анализа тут уже следует прибегать к алгоритмам машинного обучения.

4) Два текста из случайных букв:

```
Процент совпадения побуквенного анализа: 3.0666666666666664
```



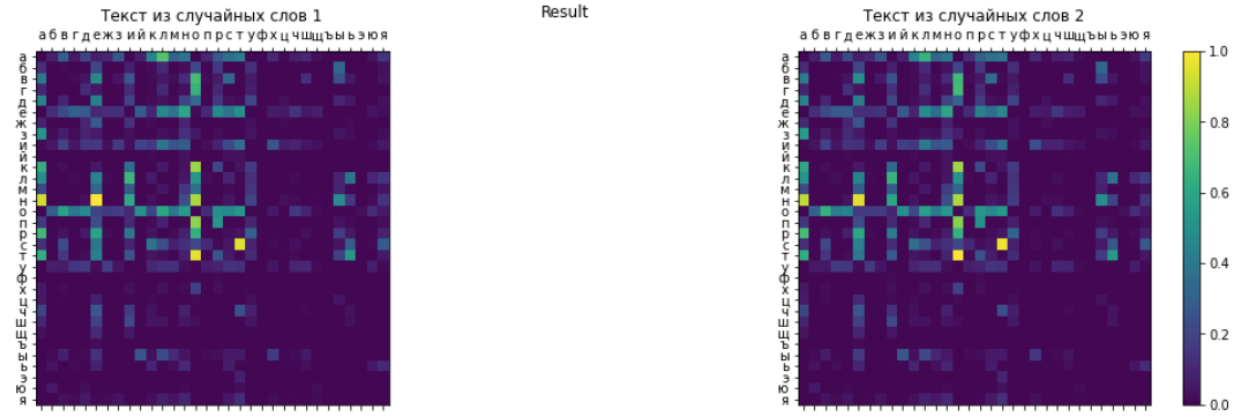

```
Средний процент частотного совпадения пар символов: 92.02293650823131
```

Учитывая равномерное распределение букв, неудивительно, что вторая мера показывает 92% совпадения. Однако первый признак указывает лишь на 3% совпадения текстов, хотя казалось бы, что можно было ожидать большего значения.

На самом деле, так же, как и во втором пункте, все можно объяснить простой теорией вероятности. Так как фактически алфавит состоит из 33 символов, включая пробел, то имея любой первый текст и второй со случайными буквами того же алфавита, то вероятность того, что на *i*-ой позиции встретится та же самая буква $\frac{1}{33} = 0{,}03(03)$. Соответственно тут и во втором пункте мы видим результат этой теоретической выкладки.

5) Два текста из случайных слов

```
Процент совпадения побуквенного анализа: 6.473333333333334
```




```
Средний процент частотного совпадения пар символов: 99.59791765302145
```

В данном пункте видим по первой мере результат, не очень, отличающийся от того, что был продемонстрирован в первом пункте. Можно сделать вывод, что расположение слов не так сильно влияет на общее количество поиндексных совпадений символов. Учитывая, что это происходит в принципе редко, то результат выглядит правдоподобным. Вторая же мера показывает 99,5% совпадения, что тоже вполне закономерно, учитывая выбор слов из одного и того же словаря, который является объединением словарей двух текстов на естественном языке.

## Выводы

В ходе работы я провел анализ различных текстов, объяснил причину полученных результатов. Тексты сравнивались на основе двух разных метрик, которые по-разному относились к особенности естественных языков. Также четко прослеживалась теория вероятностей во втором и четвертом пунктах. GitHub с работой.