**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ**

**Федеральное государственное бюджетное образовательное учреждение высшего образования**

**«Московский Авиационный Институт»**

**(Национальный Исследовательский Университет)**

**Институт: №8 «Информационные технологии   
и прикладная математика»   
Кафедра: 806 «Вычислительная математика   
и программирование»**

Лабораторная работа № 3   
по курсу «Криптография»

Группа: М8О-307Б-21

Студент: Дубровин Д.К.

Преподаватель: А. В. Борисов

Оценка:

Дата: 16.09.2024

Москва, 2024

**ОГЛАВЛЕНИЕ**

[1 Тема 3](#_Toc158983147)

[2 Задание 3](#_Toc158983148)

[3 Теория 4](#_Toc158983149)

[4 Ход лабораторной работы 5](#_Toc158983150)

[5 Выводы 11](#_Toc158983151)

# **Тема**

Критерий открытого текста

# **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.

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 случаям. Осознать какие значения получаются в этих пяти случаях. Привести соображения о том почему так происходит.

Длина сравниваемых текстов должна совпадать. Привести соображения о том какой длины текста должно быть достаточно для корректного сравнения.

# **Теория**

Критерий открытого текста — это метод в криптоанализе, при котором аналитик использует информацию о открытом тексте (например, структуру языка, вероятность появления определенных символов или слов) для раскрытия шифра. Этот метод основан на предположении, что даже после шифрования некоторые характеристики открытого текста могут сохраняться в зашифрованном сообщении, что может помочь в его расшифровке.

Критерий открытого текста основан на том, что при шифровании сообщения некоторые характеристики открытого текста могут сохраняться в зашифрованном сообщении. Это происходит потому, что многие шифровальные методы работают с открытым текстом на уровне символов или битов, сохраняя некоторые структурные или статистические свойства исходного сообщения.

Простым примером может быть частота появления букв в тексте на определенном языке. Например, в английском языке буква "е" встречается чаще всего. Если шифровщик зашифрует сообщение, сохраняя частоту букв, то криптоаналитик может использовать это знание для анализа вероятных буквенных сочетаний и, следовательно, для расшифровки сообщения.

Критерий открытого текста также может включать в себя знание о структуре языка (например, наличие определенных слов или фраз), общие шаблоны использования языка и другие характеристики сообщений, которые могут быть видны даже после шифрования.

# **Ход лабораторной работы**

Для выполнения лабораторной работы было принято решение разделить задачу на 3 подзадачи, а именно провести расследование на 3х разных длинах:

1. 2433 символов;
2. 7300 символов;
3. 15 000 символов.

Использовались тексты российской классики *(«Преступление и наказание», «Мастер и Маргарита»*) в качестве текстов на естественном языке. Для получения текста из случайных слов был найден готовый репозиторий со словами, из которого я извлекал слова (см. список используемой литературы). Текст из случайных слов был сделан через Python.

Алгоритм сравнения в контексте нашего задания очень прост. Алгоритм был реализован на языке Python. Из файла извлекаются необходимые тексты с помощью `**read\_text\_from\_file**`, потом происходит удаление всех пробелов и приведение всех слов к нижнему регистру. С помощью структуры данных **Counter** из модуля **collections** Python мы подсчитываем количество каждой буквы в каждом тексте. Затем вычисляется количество совпадающих букв между двумя текстами. Основываясь на количестве совпадающих букв и общем количестве букв, вычисляется процент совпадения. Это дробное значение от 0 до 1 предоставляет количественную оценку степени сходства между текстами.

Мой подход очень прямой и простой, потому что наша цель всего лишь подсчет кол-ва совпадающих букв. Для более сложных задач, таких как определение семантики, структуры предложений и т.д. потребуется более сложные инструменты, например, машинное обучение (word embedding) или Natural Language Processing (NLP).

* Текст большой длины (15 000 символов):
  + *Совпадения между long\_text1 и long\_text2: 0.9432;*
  + *Совпадения между long\_text1 и long\_symbols1: 0.6615;*
  + *Совпадения между long\_text1 и long\_word1: 0.8669;*
  + *Совпадения между long\_symbols1 и long\_symbols2: 0.9764;*
  + *Совпадения между long\_word1 и long\_word2: 0.9775.*

**Между двумя длинными осмысленными текстами (0.9432)**: высокий процент совпадения указывает на значительное перекрытие в использовании символов, что может быть связано с схожестью тем, стиля написания или часто используемых слов и фраз. Это ожидаемо, так как осмысленные тексты склонны следовать определенным правилам языка и часто используют похожий набор слов.

**Между длинным осмысленным текстом и текстом из случайных символов (0.6615)**: относительно низкий процент совпадения подчеркивает хаотичность распределения случайных символов по сравнению с более предсказуемым распределением в осмысленном тексте. Тем не менее, это показывает, что даже случайные символы могут иметь значительные перекрытия с буквами, используемыми в естественных языках, особенно в длинных текстах.

**Между длинным осмысленным текстом и текстом из случайных слов (0.8669)**: этот результат выше, чем с текстом из случайных символов, что подтверждает, что структурированные, даже если они случайные, слова имеют больше общего с осмысленным текстом. Слова несут в себе определенные шаблоны и комбинации букв, которые делают их более похожими на естественный текст.

**Между двумя текстами из случайных символов (0.9764) и между двумя текстами из случайных слов (0.9775)**: эти очень высокие значения указывают на большую степень совпадения между текстами, вероятно, из-за однородности и повторяемости используемых символов и слов в каждом из наборов. В случае случайных символов это может быть связано с равномерным распределением символов, хотя я все равно не особо понимаю почему так высоко совпадение. Для текстов из случайных слов высокая степень совпадения может отражать ограниченность используемого словаря слов, хотя там рандомно выбиралось из очень большого файла.

* Текст средней длины (7300 символов):
  + *Совпадения между mid\_text1 и mid\_text2: 0.9110;*
  + *Совпадения между mid\_text1 и mid\_symbols1: 0.6595;*
  + *Совпадения между mid\_text1 и mid\_word1: 0.8574;*
  + *Совпадения между mid\_symbols1 и mid\_symbols2: 0.9601;*
  + *Совпадения между mid\_word1 и mid\_word2: 0.9714.*

**Между двумя средними осмысленными текстами (0.9110):** это высокое значение показывает, что тексты имеют значительное количество общих символов. Вывод такой же как в предыдущем описании.

**Между средним осмысленным текстом и текстом из случайных символов (0.6595)**: несмотря на то что тут совпадение чуть ниже, выводы те же, случайные символы могут иметь значительные перекрытия с буквами, используемыми в естественных языках.

**Между средним осмысленным текстом и текстом из случайных слов (0.8574)**: этот результат показывает, что случайные слова имеют больше общего с осмысленным текстом по сравнению с случайными символами. Это может быть связано с тем, что даже случайно выбранные слова следуют определенным правилам формирования слов в языке, что приводит к более высокой вероятности совпадения букв.

**Между двумя текстами из случайных символов (0.9601)**: этот очень высокий процент совпадения указывает на то, что в текстах из случайных символов вероятность встречи одинаковых букв весьма высока, особенно при большом объеме текста, что приводит к большому количеству совпадений.

**Между двумя текстами из случайных слов (0.9714):** самый высокий процент совпадения среди всех сравнений говорит о том, что случайные слова, вероятно, были вероятно плохо выбраны из списка, но не думаю. На самое деле это очень странно.

* Текст малой длины (2300 символов):
  + *Совпадения между low\_text1 и low\_text2: 0.9110;*
  + *Совпадения между low\_text1 и low\_symbols1: 0.6860;*
  + *Совпадения между low\_text1 и low\_word1: 0.8449;*
  + *Совпадения между low\_symbols1 и low\_symbols2: 0.9441;*
  + *Совпадения между low\_word1 и low\_word2: 0.9611.*

**Между двумя более короткими осмысленными текстами (0.9110)**: такой же процент совпадения, как и у текстов средней длины, указывает на сильное перекрытие в использовании символов, что может свидетельствовать о схожести стиля, тем или общих слов.

**Между более коротким осмысленным текстом и текстом из случайных символов (0.6860)**: это значение немного выше, чем у средних текстов, что может быть связано с меньшей длиной текстов, уменьшая разницу между распределением символов в случайном и осмысленном тексте.

**Между более коротким осмысленным текстом и текстом из случайных слов (0.8449)**: этот процент ниже, чем у средних текстов, что может указывать на меньшее перекрытие в использовании символов между осмысленным текстом и случайно сгенерированными словами в более коротком формате.

**Между двумя текстами из случайных символов (0.9441) и между двумя текстами из случайных слов (0.9611)**: эти значения близки к результатам для средних текстов, что подтверждает предположение о том, что совпадения в текстах из случайных символов и слов вероятны и в большей степени определяются общим распределением символов и структурой слов, нежели длиной текста.

Когда мы сравниваем количество совпадений по буквам между разными типами текстов, выбор оптимальной длины текста для анализа может во многом зависеть от специфики задачи и желаемой точности результатов.

Когда мы сравниваем тексты по буквам, основной целью является определение степени сходства между текстами на уровне отдельных символов. В таком случае, более длинные тексты могут предоставить больше данных для анализа, что потенциально может повысить точность оценки сходства. Однако, слишком длинные тексты могут также усложнить анализ и увеличить время его проведения.

В контексте нашего анализа важно, чтобы тексты были достаточно длинными, чтобы обеспечить статистическую значимость совпадений, но не настолько длинными, чтобы анализ стал непрактичным. Исследования по удобочитаемости текста, например, показывают, что оптимальная длина строки текста составляет 50–75 символов с точки зрения удобочитаемости для читателя.

Таким образом, для нашего анализа текстов длиной в несколько тысяч символов может быть вполне достаточно для получения надежных результатов сравнения. Это позволяет учесть достаточное количество символов для анализа совпадений, минимизируя при этом влияние случайных совпадений, которое может быть более заметным в более коротких текстах.

# **Выводы**

Было интересно потыкать тексты и посмотреть их сравнения. Результаты выше, чем те, которые я ожидал, что достаточно интересно. Было интересно выполнять эту лабораторную работу.

# **Список используемой литературы**

1. [*https://masterimargo.ru/book-19.html*](https://masterimargo.ru/book-19.html)
2. [*https://ilibrary.ru/text/69/p.1/index.html*](https://ilibrary.ru/text/69/p.1/index.html)
3. [*https://github.com/danakt/russian-words/blob/master/russian.txt*](https://github.com/danakt/russian-words/blob/master/russian.txt)