Частотный анализ. Криптоанализ текста, зашифрованного шифром Виженера

1.Введение

Частотный анализ – метод криптоанализа, который основывается на предположении о существовании статистического нетривиального распределения отдельных символов и их последовательностей. Другими словами, для достаточно больших текстов, написанных на одном и том же языке, частота появления данной буквы является постоянной величиной, а также такое распределение может сохраняться при шифровании и дешифровании. Считается, что каждому языку свойственно определенное сочетание букв друг с другом, например, в европейских языках наблюдается чередовании гласных и согласных.

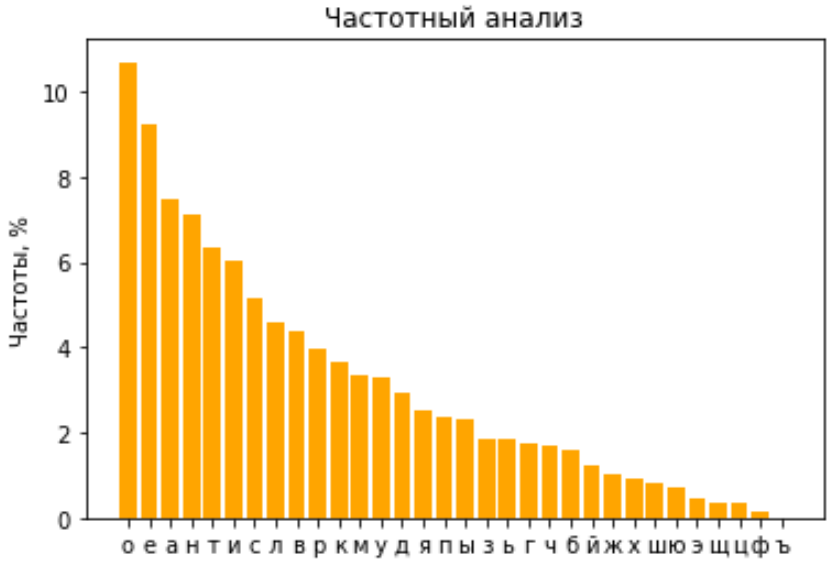
Моноалфавитные шифры (например, шифр Цезаря) можно легко взломать методами частотного анализа, так как при таком шифровании каждый символ исходного текста заменяется некоторым другим постоянным символом. В отличие от моноалфавитных шифров, при полиалфавитном шифровании алфавит шифруемого сообщения изменяется в процессе шифрования. Таким образом маскируется естественная частота появления символов.

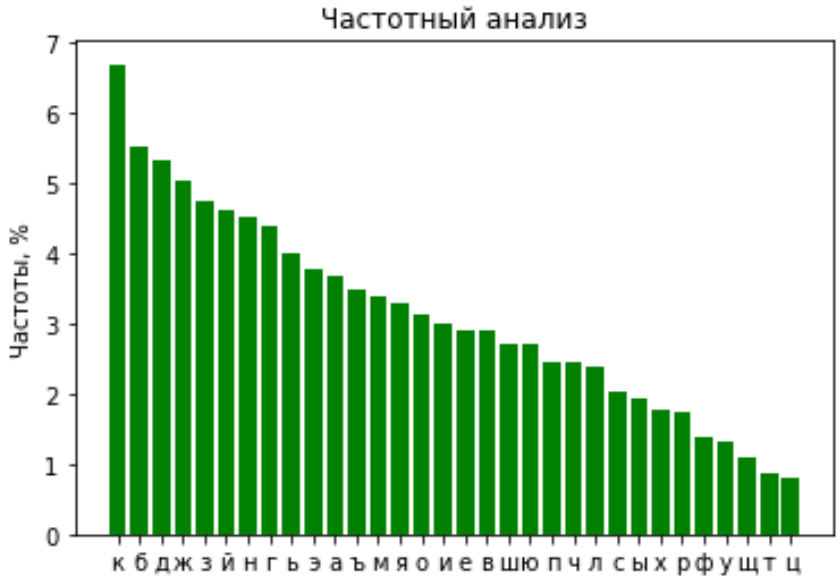
Одним из наиболее известных полиалфавитных шифров является шифр Виженера, который использует заданное ключевое слово в процессе шифрования. Ключом может быть произвольное слово. В тексте, который нужно зашифровать, и в ключе, буквы заменяются на их номера в алфавите. Ключ записывается под строкой сообщения несколько раз, чтобы его длина сравнилась с длиной шифруемого сообщения. После этого номера символов в шифруемом сообщении и в ключе суммируются друг с другом по модулю N (кол-во букв в алфавите). Результат суммирования переводится в буквы и получается зашифрованное сообщение.

Моя задача заключалась в реализации шифра Виженера, проведении частотного анализа для текстов на русском и английском языке, нахождении ключа и расшифровке шифротекста методами частотного анализа.

2. Частотный анализ

Для проведения частотного анализа необходимы тексты достаточно большой длины, поэтому я использовала текст книги Замятин «Мы». Следующие графики визуализируют полученные результаты (оранжевый – распределение частот символов в исходном тексте, зеленый – в зашифрованном)





Можно заметить, что частоты появления символов в зашифрованном шифротексте распределены более равномерно. Это происходит из-за того, что при использовании шифра Виженера одинаковые буквы шифруются по-разному. Долгое время шифр Виженера считался невзламываемым, однако при определенных условиях этот шифр также поддается криптоанализу.

Для проведения частотного анализа необходимо удалить из текста лишние символы (цифры, знаки препинания), подсчитать, сколько раз каждый символ встречается в тексте и разделить это число на количество всех символов в тексте.

3. Криптоанализ шифра Виженера

Криптоанализ шифра Виженера состоит из двух этапов. Сначала находят длину ключа. Когда длина ключа известна, зашифрованный текст можно разбить на блоки, каждый из которых соответствует одному символу ключа, и применить частотный анализ.

1) Определение длины ключа

Известно несколько методов поиска длины ключа: тест Касиски и метод индекса совпадений. Метод индекса совпадений является интуитивно более понятным, менее трудоемким и допускает анализ текста с длинным ключом. Этот метод основывается на том факте, что вероятность совпадения двух случайных букв в некотором достаточно длинном тексте (индекс совпадений) — это постоянная величина для каждого алфавита. Эту величину можно вычислить по следующей формуле:

$$I(\vec{x}) = \sum_{i=1}^{m} \frac{f_i\,(f_i - 1)}{n\,(n-1)}$$

, где m – мощность алфавита,fi – частота появления i-го символа в тексте, n – количество символов в тексте. Значение индекса совпадений для открытого текста на русском и английском языках равняется 0.0553 и 0.0667 соответственно. Для зашифрованного текста это число уменьшается и чем длиннее ключ, тем ближе это значение индексу совпадения в совершенно случайном тексте.

Чтобы найти длину ключа, можно вычислять индекс совпадений, прореживая исходный текст. То есть брать текст, составленный из каждой 2-ой (затем каждой 3-ей, 4-ой и так далее) букв исходного текста. Как только значение индекса совпадений для прореженного (для данного

значения n и кратных ему чисел) текста окажется достаточно близким к известному значению для открытого текста, можно предполагать, что длина ключа равняется n.

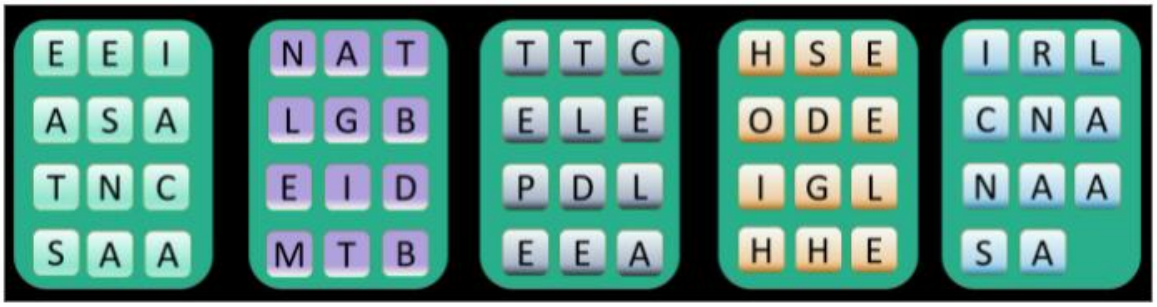
		Indexes,%
2	2	0,042844448
3	3	0,042561358
4	4	0,042148554
5	5	0,036926487
6	6	0,054723398
7	7	0,039708426
8	8	0,043172691
9	9	0,045045045
10	10	0,045023095
11	11	0,037446286
12	12	0,057174151
13	13	0,038183695
14	14	0,042253521
15	15	0,043290043
16	16	0,043273013
17	17	0,035072207
18	18	0,068304668
19	19	0,033150183

2) Поиск ключевого слова

Зная длину ключа n, можно разбить текст на n блоков, как показано на картинке. Тогда к каждому блоку может быть применен частотный анализ, потом что символы из каждого блока были получены сдвигом на одно и то же количество позиций, значит для каждого такого блока сохраняется естественное распределение частот.

Let's look at the message from earlier in the page, and imagine that it has been enciphered with a key of 5 characters:

We can extract each letter encoded by the same part of the key, i.e. all those that share the same colour in the image, and place them together in their own group. Each letter in a group has been encoded using the same Caesar cipher and this is because each letter has been encoded by the same part of the key.



Можно использовать метод Кирхгофа для поиска ключа. Он заключается в сравнении частоты появления символов в столбцах с частотой появления символов в исходном тексте для нахождения ключевого символа для этого столбца.

Результат работы программы

После нахождения ключевого слова текст можно расшифровать стандартным методом.