kaspersky.academy

Лекция 4

Криптоанализ шифра Виженера

Онлайн-курс по математике в информационной безопасности

kaspersky.academy

Лекция 4: Криптоанализ шифра Виженера

В этой лекции мы сначала вспомним, как работает шифр Виженера, а после этого попробуем его дешифровать.

План лекции:

Поехали!

Коротко о шифре Виженера (еще раз)

Шифр Виженера искажает частоту появления символов открытого текста, и метод частотного анализа перестает работать. Но одна закономерность в этом шифре все-таки есть – это **ключ**.

Если открытый текст длиннее ключа, мы будем повторять ключ много раз. Допустим, ключевое слово – SECRET. Тогда каждая 7-я буква открытого текста будет зашифрована одной и той же буквой ключевого слова. А каждая буква ключевого слова соответствует какому-то шифру Цезаря с ключом k_i.

Рисунок 1: Повторяющиеся буквы шифртекста

Криптоанализ шифра Виженера

Криптоанализ шифра Виженера будем выполнять в два этапа:

1. Найдем длину ключевого слова
2. Найдем буквы этого слова

Как искать длину ключа: тест Фридмана

Существует несколько способов определить длину ключа. Один из действенных методов предложил в **Уильям Фридман** еще в начале XX века. Тест Фридмана отлично показал себя во взломе полиалфавитных шифров, а поэтому шифр Виженера сейчас тоже считается историческим. Помимо теста Фридмана существует еще **тест Касиски**, но он не всегда работает хорошо: в тесте Касиски мы допускаем, что биграммы и триграммы (это двух- и трехбуквенные последовательности) будут повторяться с некоторой частотой и в открытом тексте, и в шифртексте. Но чем меньше текст, тем труднее

строить для него статистику. Все просто: больше текст – лучше статистика:)

Перед тем как увидеть много страшных формул, давайте вспомним несколько переменных и введем одну новую.

P – открытый текст
m – длина открытого текста
i – буквы алфавита А

Алфавит – русский алфавит без Ё. Пока ничего не изменилось, Ё мы все еще не любим:)

$A = a_1, \dots a_n = А, Б, В, \dots Я$
$n = |A| = 32$ – это мощность алфавита
f_i – сколько раз в сообщении встретилась i

Давайте в качестве открытого текста опять возьмем надпись на стене Древнего Рима:

P = ЗДЕСЬ БЫЛ ЦЕЗАРЬ

$m = 14$

Буква А встретилась в тексте всего 1 раз. Значит, $f_1 = 1$.

А вот буква Е используется сразу в двух словах. Буква Е – шестая по алфавиту, значит $f_6 = 2$.

Введем новое понятие – индекс совпадений I_c. Это число покажет вероятность того, что какая-то буква a_i встречается в тексте два раза.

$$I_c = \sum_{i=1}^{n} \frac{f_i(f_i-1)}{m(m-1)} = \frac{f_1(f_1-1)}{m(m-1)} + \frac{f_2(f_2-1)}{m(m-1)} + \ldots + \frac{f_n(f_n-1)}{m(m-1)}$$

Первый множитель дроби $\frac{fi}{m}$ показывает вероятность того, что на каком-то произвольном месте любого текста стоит буква a_i.

Второй множитель $\frac{(fi-1)}{(m-1)}$ – это вероятность того, что и второй случайно выбранный символ текста тоже a_i.

Если текст достаточно большой, то $\frac{fi}{m} \approx \frac{(fi-1)}{(m-1)}$

Представьте, что в вашем тексте 1000 знаков, 100 из которых – буква А. Величины $\frac{100}{1000}$ и $\frac{99}{999}$ не будут сильно отличаться. И тогда

$$I_c = \sum_{i=1}^{n} p_i^2$$

p_i – частота встречаемости буквы a_i или, что то же самое, вероятность того, что на каком-то месте встретится именно эта буква.

Почему это будет работать?

Для этого нужно вспомнить, что буквы в осмысленных текстах встречаются неравномерно. Гласные – чаще, согласные – реже, а самая частая буква русского языка – буква О.

$p(O) \approx 0{,}107$

$p(A) \approx 0{,}087$

$p(E) \approx 0{,}084$

Если возвести все вероятности в квадрат и сложить, получим индекс совпадений:

$I_c \approx 0.0553$

В случайных текстах буквы будут встречаться равновероятно. Если всего букв в алфавите 32, то вероятность появления любой буквы в тексте

$$p_i = \frac{1}{32}$$

Для случайных текстов

$$I_c \approx 0.03125$$

Важно, что **индекс совпадения будет похож на индекс совпадения для осмысленных текстов даже тогда, когда к шифру применяется моноалфавитный шифр.**

Что меняет моноалфавитный шифр? Он переставляет местами буквы. Например, алфавит А будет выглядеть так: = С, О, В, ..., А. При этом каждая буква будет встречаться в тексте с вероятностью p_i. А сумма квадратов вероятностей никак не изменится – неважно, как переставлять слагаемые в сумме и к какой букве относится вероятность p_i – к букве А или букве С.

Таким образом, можно выяснить, зашифрован ли текст одним алфавитом, или же мы имеем дело с полиалфавитным шифром.

Дешифрование шифра Виженера: ищем длину ключа

В наши руки попал какой-то текст:

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУТПЪЬЕЙФКМЖФУРТЦХЫХЩУТПЪЬЕЙШРЫКФУР

Мы посчитали индекс совпадения для этого текста.

$$I_c = 0.0393$$

Значит, текст зашифрован полиалфавитным шифром.

Чтобы **найти длину ключа**, мы будем **перебирать различные длины и считать индекс совпадений для текста. Если индекс совпадений будет похож на индекс совпадений** для осмысленных текстов, длину ключа мы угадали.

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУТПЪЬЕЙФКМЖФУР
ТЦХЫХЩУТПЪЬЕЙШРЫКФУРТШРЫКФУТПЪЬЕЙОЕЬЧЩЕЪ
СХИЫФЗМПЙЧТАЦХКШЧЮЭТЖТУУЙФТЫЗХХНЛЧКУЙФТЫ
ЙТШДЬМФШТЩТЫЗХЬХЦЗКЩТШЧЙМУКТЦОТНЫМТХЙХЦ
ЫЕВКЮПЪЬНМФКЪДШЧЫПГПЫТШУОЯМЬЯТИАЪДЧШЕДЩ
ББФЗЗХПЗФЭМДЧЫРЦХНОЩНДСХЦЯАЙЕУСМКОЙОШЬФ
МЬЪТШЧХТЯНООПТТИХРУСВМНРЗРДМЙЕЯАШДТХТНЪЙО
ЕЩДТЬХЖЗГЯХЖДПСХЗЮЦЧКЯМЙЙПЧШСИХТКЪСХЦЯАХЧ
ОФХЦЙМШПАЬМТХЙЪИНИЗЧЙИХРУЙФЦАЭМЦЯЖХЗНЦГУ
СМФНУЙТЕЯЙТБЪТЩУШАСУЫИПТЫЫМЗХИФАЦХЦУЮТИ
ЦСЙТЕЯАЮЧЫЦХЪЫЦЖВЯТЩЦЬТШУОУХТНЫЗРАРХЛТЦИ
АЯАПТТТЮКПМЛКЪЙШРХЖВТТЗХРШДФЙТЪШКЦЫЗЦШЧ
ЮЭТЫМСЪМСУРИЗЪЫЦЖТХОХИСДЮЕЮЦХРАЫЯКДЙУФЭ
ГУУЮЙРЬНХМЦШМЧКНППМНЪПГЮПХЛЪТХЖЗГШТХЦГНС
ЙЖЧЫЫФУЬПХЪНЙШРХФМЕШМОЕГМЕРТЗСУЫЕБДЮСПЧ
ЙЖХМЩТНТЫМЛКМЩХХЫЬЗФЭТШЧЭДФЦЯЖЗНЩЙФУЯП
ПЬЪДЖЭЯЧСЕОЧЛКЩМШФЫПГМЫЖЗЧЙМЬЬНЭМЗЪМУЕЯ
ЙТБЪЯРЬХЦЗЧТПГЗЫХСРХОФКЯЦЗПХЩНКСЙЙДЯСЗЙГДЩ
БПБЩУЩЦЛЗПШВЮЕТЦШДБМТУЮТЫЦЧМРФЫИЩКЧХЩТХ
ОЗПХЙЦХНЖПРНСМЗЫЛЙУСМЩБПДИЦЫПЕЧФИМЦЙОЗЛ
СЯРСЫКМЧЫУЧКСЙТНЯАЛРМХМЖМХЙУЦУЧНЪЪПФХЕЪЙ
ТЦЦХНЖ...

Рисунок 2: Шифртекст

Предположим, что длина ключа = 2.

Тогда берем текст, составленный из каждой второй буквы текста. Мы должны перешагнуть через всю длину ключа и взять следующий символ. В этом случае все символы, которые мы выбрали, будут зашифрованы одной и той же буквой ключа. Буквы, зашифрованные первой буквой ключа, выделены черным. Считаем его индекс совпадения I_c для этих букв.

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУ
ТПЪЬЕЙФКМЖФУРТЦХЫХЩУТПЪЬЕЙ…

$I_c(c_1c_3c_5…) = 0.043$

c_i – i – q.

Индекс совпадения для ключа с длиной 2 далек от значения 0.0553.

Выпишем еще несколько последовательностей букв и посчитаем индексы совпадений.

|k|= 3: $I_c(c_1c_4c_7)$ = ОЮУ…

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУ
ТПЪЬЕЙФКМЖФУРТЦХЫХЩУТПЪЬЕЙ…

|k|= 4: $I_c(c_1c_5c_9)$ = ОМП…

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУ
ТПЪЬЕЙФКМЖФУРТЦХЫХЩУТПЪЬЕЙ…

|k|= 5: $I_c(c_1c_6c_{11})$ = ОЙЬ…

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУ
ТПЪЬЕЙФКМЖФУРТЦХЫХЩУТПЪЬЕЙ…

|k|= 6: $I_c(c_1c_7c_{13})$ = ОУЬ…

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУ
ТПЪЬЕЙФКМЖФУРТЦХЫХЩУТПЪЬЕЙ…

$|k| = 1$: $I_c = 0.039$
$|k| = 2$: $I_c = 0.043$
$|k| = 3$: $I_c = 0.041$
$|k| = 4$: $I_c = 0.057$
$|k| = 5$: $I_c = 0.048$
$|k| = 6$: $I_c = 0.044$

Самое близкое значение получаем при значении ключа $k = 4$. Скорее всего, длина ключа $|k| = 4$.

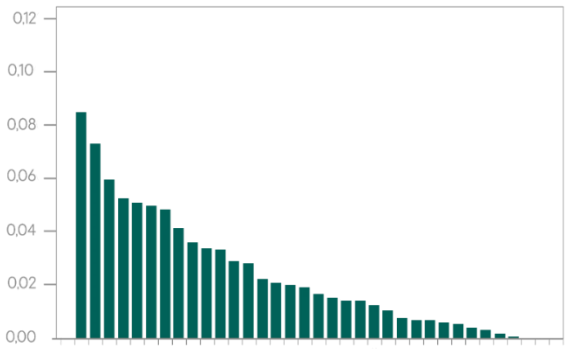
Дешифрование шифра Виженера: ищем буквы ключа

Чтобы найти буквы ключа, можно применить несколько способов.

1. Старый добрый частотный анализ

И... ничего хорошего не получится. Если для каждой буквы ключа составлять свою частотность символов шифртекста, мы сможем получить самые частые буквы – О, А, Е. Но этот метод не точный, и в конце все равно придется перебирать. Фридман предложил более действенный метод.

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУ
ТПЪЬЕЙФКМЖФУРТЦХЫХЩУТПЪЬЕЙ...



2. Взаимный индекс совпадений

Если в индексе совпадений для одной строки мы искали вероятность того, что какая-то буква алфавита встретится в тексте

два раза, то во взаимном индексе совпадений мы будем искать **вероятность того, что одна и та же буква встретится и в первом, и во втором текстах одновременно.**

$$MI_c = \sum_{i=1}^{n} \frac{f_i \cdot f_i'}{m \cdot m'}$$

$\frac{f_i}{m}$ – вероятность того, что i-я буква встретится в первом тексте длины m.

$\frac{f_i'}{m'}$ – вероятность того, что i-я буква встретится во втором тексте длины m'.

Давайте теперь выберем два множества букв шифртекста. Первое множество будет соответствовать первой букве ключа (черные буквы), второе – второй (красные буквы).

ОЧЕЮМЙУТПЪЬЕЙЪХЫИТНПТКУМЖФУ
ТПЪЬЕЙФКМЖФУРТЦХЫХЩУТПЪЬЕЙ...

$k = k_1k_2k_3k_4$

k_1: ОМПЙИТЖПЙЖТХ...
k_2: ЧЙЪЪТКФЪФФЦ...

Первый и второй текст зашифрованы шифром Цезаря, но с разными ключами k_1 и k_2. Так как шифр Цезаря – это сдвиговый шифр, можно сказать, что k_2 можно получить из ключа k_1 каким-то дополнительным сдвигом:

$k_2 = k_1 + x$

Этот дополнительный сдвиг x мы и будем искать.

Когда мы найдем сдвиг x для второго текста, сдвинем алфавит этого текста на $-x$. Тогда и первый, и второй тексты будут зашифрованы одним и тем же шифром Цезаря с ключом k_1.

Как искать относительный сдвиг

Сдвиг второго текста относительно первого будем называть относительным сдвигом.

Как его найти:

- алфавит первого текста оставляем на месте
- будем сдвигать алфавит второго текста до тех пор, пока $MI_c \approx 0.0553$

После этого мы сдвинем алфавит второго текста на $-x$. То же самое повторим со всеми текстами.

Давайте рассмотрим этот прием на примере.

ОМПЙИТЖПЙЖТХ... – первый шифртекст, который шифровался ключом k_1

ЧЙЪЪТКФЪФФЦ... – второй текст с ключом $k_2 = k_1 + x$

В первой строке таблицы записаны буквы алфавита, во второй – сколько таких букв встретилось в первом шифртексте, в третьей – сколько таких букв во втором.

А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	...	Я
7	1	1	2	10	3	11	4	9	26	3	3	20	0	7	14	3	9	18	3	5	10	...	3
0	1	3	5	1	2	5	19	4	7	1	4	16	2	5	9	4	5	12	4	16	24	...	2

Длина первого текста $m = 210$, длина второго $m' = 209$.

Буква А встретилась 7 раз в первом шифртексте и вообще не встретилась во втором.

Теперь нужно понять, как работает взаимный индекс совпадений.

Если бы мы рассматривали только первый текст, то слагаемое в сумме

$$I_c = \sum_{i=1}^{n} \frac{f_i^2}{m^2}$$

для буквы А было бы таким:

$$I_a = \frac{7^2}{210^2}$$

А вот в сумме

$$MI_c = \sum_{i=1}^{n} \frac{f_i \cdot f_i'}{m \cdot m'}$$

слагаемое с буквой А вообще обнулится:

$$\frac{7 \cdot 0}{210 \cdot 209}$$

Поэтому **нужно найти такой сдвиг второго алфавита, чтобы количество встреченных букв было приблизительно одного порядка и для первого, и для второго текстов.** Сейчас мы видим, что большие цифры второй строки не совпадают с большими числами третьей строки. Поэтому мы будем сдвигать буквы второго алфавита влево до тех пор, пока не увидим приблизительное совпадение в числах.

При сдвиге $x = 1$

А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	…	Я
7	1	1	2	10	3	11	4	9	26	3	3	20	0	7	14	3	9	18	3	5	10	…	3
1	3	5	1	2	5	19	4	7	1	4	16	2	5	9	4	5	12	4	16	24	5	…	0

При сдвиге $x = 2$

А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	…	Я
7	1	1	2	10	3	11	4	9	26	3	3	20	0	7	14	3	9	18	3	5	10	…	3
3	5	1	2	5	19	4	7	1	4	16	2	5	9	4	5	12	4	16	24	5	8	…	1

При сдвиге $x = 1$ и $x = 2$ ничего хорошего не получилось. А вот сдвиг $x = 3$ примерно выровнял столбцы. Теперь буква Ж встретилась 11 и 7 раз в разных строках, буква Т – 18 и 24, буква Ь – 3 и 2 раза.

А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	…	Я
7	1	1	2	10	3	11	4	9	26	3	3	20	0	7	14	3	9	18	3	5	10	…	3
5	1	2	5	19	4	7	1	4	16	2	5	9	4	5	12	4	16	24	5	8	17	…	3

Конечно, мы не будем доверять своим глазам и посчитаем взаимный индекс совпадений при разных сдвигах второго алфавита.

1: $I_c = 0.034$
2: $I_c = 0.026$
3: $I_c = 0.053$
4: $I_c = 0.024$
5: $I_c = 0.014$

Расчеты только подтверждают догадку – самый лучший результат будет для сдвига $x = 3$. Значит, $k_2 = k_1 + 3$.

Таким же образом считаем сдвиг третьего и четвертого текстов относительно первого.

Выровняем алфавиты так, чтобы все тексты были зашифрованы одним шифром Цезаря. Это значит, что второй текст нужно сдвинуть на 3 позиции назад, третий текст – на 1 позицию, четвертый – на 8.

k_1: ОМПЙИТЖКПЙЖТХПЙКТКПЙЧСФЙЦЧЖЙЗЛ
k_1+3: ЧЙЪЪТКФЪФФЦЩЪШФШФЪОЩХЗЧХЮТФХЧ
k_1+1: ЕУЬХНУУЬКУХУЬРУРУЬЕЕИМТКЭУТХКТШФТ
k_1+8: ЮТЕЫПМТЕМРЫТЕЫРЫТЕЬЪЫПАШТУЫН

После того как все тексты оказались сдвинуты так, что все они зашифрованы одним ключом k_1, возвращаем их в изначальный шифртекст.

k_1: ОМПЙИТЖКПЙЖТХПЙКТКПЙЧСФЙЦЧЖЙЗЛ
k_1+3: ЧЙЪЪТКФЪФФЦЩЪШФШФЪОЩХЗЧХЮТФХЧ
k_1+1: ЕУЬХНУУЬКУХУЬРУРУЬЕЕИМТКЭУТХКТШФТ
k_1+8: ЮТЕЫПМТЕМРЫТЕЫРЫТЕЬЪЫПАШТУЫН

k_1: ОМПЙИТЖКПЙЖТХПЙКТКПЙЧСФЙЦЧЖЙЗЛ
k_1: ФЖЧЧПЗСЧССУЦЧХСХСЧЛЦТДФТЫПСТФ
k_1: ДТЫФМТТЫЙТФТЫПТПТЫДДЗЛСЙЬТСФЙ
k_1: ХЙЬТЖГЙЬГЗТЙЬТЗТЙЬУСТЖЧПЙКТД

А теперь, когда он зашифрован шифром Цезаря, применяем **брутфорс-атаку:** переберем все возможные варианты сдвига для шифра Цезаря (а из 31), найдем ключ k_1.

$k_1 = 4$
$k_2 = 4 + 3 = 7$
$k_3 = 4 + 1 = 5$
$k_4 = 4 + 14$

4	7	5	14
Д	З	Е	Н

Секретное слово – ДЗЕН.

После расшифровки получаем текст **Дзен-Питон**, описывающий философию языка программирования Питон:

КРАСИВОЕ ЛУЧШЕ УРОДЛИВОГО. ЯВНОЕ ЛУЧШЕ НЕЯВНОГО. ПРОСТОЕ ЛУЧШЕ СЛОЖНОГО. СЛОЖНОЕ ЛУЧШЕ ЗАПУТАННОГО. РАЗВЕРНУТОЕ ЛУЧШЕ ВЛОЖЕННОГО. РАЗРЕЖЕННОЕ ЛУЧШЕ ПЛОТНОГО. ЧИТАЕМОСТЬ ИМЕЕТ ЗНАЧЕНИЕ. ОСОБЫЕ СЛУЧАИ НЕ НАСТОЛЬКО ОСОБЫЕ, ЧТОБЫ НАРУШАТЬ ПРАВИЛА. ПРИ ЭТОМ ПРАКТИЧНОСТЬ ВАЖНЕЕ БЕЗУПРЕЧНОСТИ. ОШИБКИ НЕ ДОЛЖНЫ ЗАМАЛЧИВАТЬСЯ. ЕСЛИ НЕ ЗАМАЛЧИВАЮТСЯ ЯВНО. ВСТРЕТИВ ДВУСМЫСЛЕННОСТЬ, ОТБРОСЬ ИСКУШЕНИЕ УГАДАТЬ. ДОЛЖЕН СУЩЕСТВОВАТЬ ОДИН - И, ЖЕЛАТЕЛЬНО, ТОЛЬКО ОДИН - ОЧЕВИДНЫЙ СПОСОБ СДЕЛАТЬ ЧТО-ТО. ХОТЯ ЭТОТ СПОСОБ ПОНАЧАЛУ МОЖЕТ БЫТЬ И НЕ ОЧЕВИДЕН, ЕСЛИ ВЫ НЕ ГОЛЛАНДЕЦ. СЕЙЧАС ЛУЧШЕ, ЧЕМ НИКОГДА. ХОТЯ НИКОГДА ЧАСТО ЛУЧШЕ, ЧЕМ *ПРЯМО* СЕЙЧАС. ЕСЛИ РЕАЛИЗАЦИЮ СЛОЖНО ОБЪЯСНИТЬ - ИДЕЯ ТОЧНО ПЛОХА. ЕСЛИ РЕАЛИЗАЦИЮ ЛЕГКО ОБЪЯСНИТЬ - ВОЗМОЖНО, ИДЕЯ ХОРОША. ПРОСТРАНСТВА ИМЕН - ОТЛИЧНАЯ ШТУКА! БУДЕМ ИСПОЛЬЗОВАТЬ ИХ ЧАЩЕ! ВНИМАТЕЛЬНЫЙ ЧИТАТЕЛЬ ВОСКЛИКНЕТ - ТАК ИХ ЖЕ ДЕВЯТНАДЦАТЬ! В ЭТОМ ЗАКЛЮЧАЕТСЯ ФИЛОСОФСКИЙ ПОДТЕКСТ - НИКАКИЕ ПРАВИЛА НЕ ВОЗВОДИТЬ В АБСОЛЮТ. ЗДЕСЬ КАЖДЫЙ МОЖЕТ ОПРЕДЕЛИТЬ ДЛЯ СЕБЯ СВОЙ ПРИНЦИП И БУДЕТ ПРАВ.

~~А после этой сложной лекции всем решать задачи! Так вы сможете разобраться, что~~Теперь вы знаете, что на самом деле происходит в криптоанализе шифра Виженера. Успехов!