Душкин Роман Викторович
Директор по науке и технологиям искусственной интеллектуальной системы для персональной медицины «Джейн», автор курса по основам искусственного интеллекта (ai101.ru).

Шифр простой замены

Эта статья открывает цикл «Криптография», который будет состоять из нескольких научно-популярных статей. Каждая статья этого цикла будет посвящена одному методу криптографии (или стеганографии) с его кратким описанием, исторической справкой, лежащей в основе математикой и вариантами криптоанализа («атаками» на шифрограммы). В этой вводной статье рассматривается предмет криптографии и описывается самый простой шифр – шифр простой замены.

Криптография – наука и практическая область деятельности по защите секретной информации путём её изменения таким образом, чтобы она стала непонятной для непосвящённых. Криптография начала развиваться с незапамятных времён, и сегодня можно сказать, что именно эта область человеческой деятельности была одной из основ развития научно-технического прогресса на протяжении всей истории. Ведь с древнейших времён между теми, кто скрывал информацию, и теми, кто пытался раскрыть секреты, шло непрекращающееся соревнование.

Криптоанализ – набор теоретических и практических методов взлома секретов, зашифрованных методами криптографии. Криптоаналитики, специалисты по криптоанализу всегда должны были иметь глубокие знания в математике, статистике, лингвистике и многих других науках, чтобы применять эти знания на практике для атаки на шифрограммы. Но, более того, криптоаналитики должны обладать незаурядными способностями к разработке новых, нестандартных и далеко нетривиальных подходов к шифрам. И часто именно хитрость и нестандартность мышления позволяла криптоаналитикам взломать шифры, которые до них считались невзламываемыми.

Начиная с этой статьи, мы изучим множество различных методов криптографии и криптоанализа, и по

окончании каждый читатель будет уметь разгадывать зашифрованные секреты не хуже Шерлока Холмса или Уильяма Леграна. И, конечно же, заинтересованный читатель научится скрывать свои секреты так, чтобы никто не смог их выведать. Это возможно, и существуют методы абсолютной защиты, что доказано с математической точки зрения. Но начнём мы с самого простого.

Ну, а сегодня мы начнём с шифра простой или одноалфавитной замены, который иногда называется «шифром Цезаря». Это не совсем корректное название, так как шифр Цезаря сам по себе является шифром сдвига, а шифр сдвига – это шифр одноалфавитной замены. Но все эти виды шифров взламываются одним и тем же методом, так что можно все их рассмотреть одновременно.

Метод шифрования

Гай Светоний Транквилл, древнеримский историк, около 121 года нашей эры написал занимательный труд «Жизнь двенадцати цезарей». В этой книге автор изложил биографию Гая Юлия Цезаря и других 11 римских принцепсов.

Нам, изучающим криптографию, это произведение интересно в первую очередь тем, что в нём описан так называемый «шифр Цезаря» – простой шифр со сдвигом 3 символа, который использовался Юлием Цезарем для шифровки своих военных посланий. Считается, что именно Юлий Цезарь был первым человеком, про которого известно из исторических источников, что он пользовался таким типом шифрования.

Юлий Цезарь делал очень простую операцию. Он циклично сдвигал каждую букву на 3 позиции вперёд, то есть вместо буквы «A» использовал букву «D», вместо буквы «B» использовал букву «E» и так далее до буквы «Z», вместо которой использовалась буква «C». Таким образом, латинская фраза «SAPERE AUDE», зашифрованная при помощи шифра Цезаря, будет выглядеть как «VDSHUH DXGH». Чтобы это расшифровать, нужно произвести обратную операцию, то есть для каждой буквы шифрограммы брать из алфавита букву на 3 позиции назад, опять циклично переходя через начало алфавита.

В принципе, шифровать таким методом можно с любым сдвигом. Для этих целей можно воспользоваться следующими математическими формулами:

- $E(m) = (m + k) \bmod n$,
- $D(m) = (m - k + n) \bmod n$,

где m – номер буквы в алфавите, n – общее количество букв в алфавите и k – сдвиг, mod – операция взятия по модулю. При этом функцию шифрования традиционно обозначают буквой E (от англ. *encode*), а функцию дешифровки – буквой D (от англ. *decode*). Можно видеть, что для любой буквы m из алфавита при любом сдвиге k выполняется условие $m = D(E(m))$.

Что интересно, для шифра Цезаря можно сделать шифровальное устройство, например из картона. Для этого необходимо вырезать из картона два диска, диаметр одного из которых должен быть на пару сантиметров больше диаметра второго. Каждый диск разбивается на одинаковое количество секторов по числу букв в алфавите. В секторы ближе к краям дисков необходимо последовательно нанести все буквы алфавита, после

чего положить меньший диск на больший и соосно прикрепить их канцелярской кнопкой. Вращение дисков друг относительно друга на некоторое количество секторов устанавливает ключ. Соответствие букв на большем диске меньшему определяет процедуру шифрования, а на меньшем большему – процедуру расшифровки.

Итак, особого интереса эта система шифрования не представляет. Разве что стоит упомянуть так называемую систему ROT13, которая используется до сих пор для английского алфавита, состоящего, как известно, из 26 букв. Как следует из её наименования, она представляет сдвиговый шифр с $k = 13$. И, таким образом, двойное применение метода ROT13 к сообщению возвращает его в исходное состояние, то есть ROT13(ROT13(m)) = m. Сегодня этот метод шифрования используется на форумах, в электронной почте и других подобных местах, чтобы скрыть «спойлер», то есть информацию, которую нежелательно увидеть случайным взглядом. Для русского алфавита аналогом системы ROT13 будет система ROT16, если сократить алфавит до 32 букв, переставь, например, использовать букву «Ё» (что и так довольно часто делается).

Более общей системой шифрования по отношению к сдвиговым шифрам являются шифры простой или одноалфавитной замены. В таких шифрах каждому отдельному символу открытого текста ставится в соответствие какой-либо символ для шифрования, и между такими символами имеется взаимно однозначное соответствие. Если пользоваться компьютером, то обычно для шифрования используются символы из той же таблицы кодировки, что и символы открытого алфавита, но перепутанные в произвольном порядке. Если же использовать такую систему шифрования вручную, то для шифроалфавита можно пользоваться вообще любыми значками, на которые только хватит фантазии.

Примером шифра простой замены является следующий:

А	Б	В	Г	Д	Е	Ё	Ж	З	И	Й
अ	ब	व	ग	द	ए	ओ	ग़	ज़	इ	य
К	Л	М	Н	О	П	Р	С	Т	У	Ф
क	ल	म	न	ओ	प	र	स	त	उ	फ़
Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
ह	च	ट	श	ठ	ढ	ई	ध	ऐ	ऊ	आ

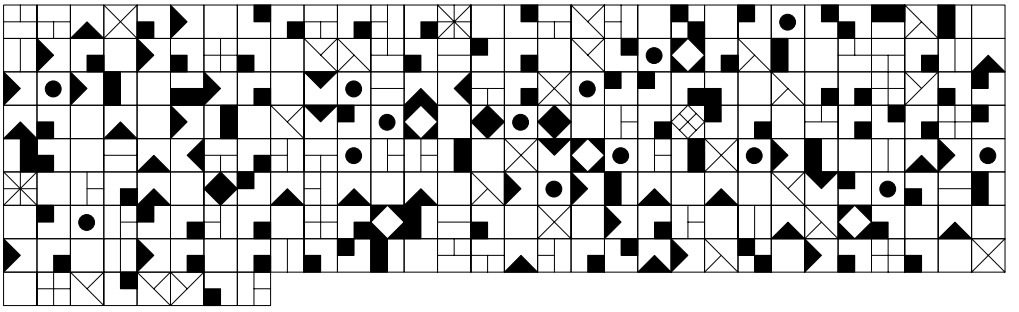
И, собственно, шифровка производится простой заменой букв открытого алфавита на значки закрытого. Примерно вот так:

तईउग़एदओगअदअलसआटतओऐतओज़अपइसधमओ

На этом, пожалуй, и всё. Теперь перейдём к рассмотрению методов атаки на шифрограммы, зашифрованные таким видом шифров.

Методы атаки

Представим, что вы обнаружили следующее секретное послание:



При первом взгляде на эту красоту начинает рябить в глазах, и может возникнуть мысль, что расшифровать такое невозможно. Но это не так. Во-первых, надо заметить, что тут много повторяющихся узоров, вписанных в квадрат, и по таким узорам можно определить, что шифрограмма состоит из 9 строк и 30 столбцов (итого 248 символов). Во-вторых, шифр простой замены является самым простым шифром, который взломать очень легко. И сегодня мы этому научимся.

Но перед тем как приступить к взлому шифра, необходимо убедиться, что это послание зашифровано именно шифром простой замены. Сделать это несложно, достаточно подсчитать:

1. Количество различных символов – оно должно быть примерно равно количеству символов в алфавите, то есть, например, для русского языка – не превосходить 33. Впрочем, шифровальщик мог поступить более хитро и использовать специальные символы для пробела, знаков препинания, цифр. Может быть, что и заглавные со строчными буквами также шифруются при помощи различных символов. Однако в любом случае для русского языка общее количество использованных для шифрования различных символов не должно превосходить сотни. Если превосходит, то это вряд ли шифр простой замены.

2. Для каждого символа – количество его использований в шифрограмме. Если затем отсортировать полученные количества по убыванию, то полученная гистограмма должна соответствовать таковой для русского алфавита. И тут чем больше объём шифрограммы, тем сильнее построенная гистограмма должна быть похожа на эталонную. При этом во внимание надо, конечно же, принимать, все соображения, приведённые в предыдущем пункте.

Гистограмма частот букв русского алфавита в обычном тексте значительного объёма выглядит следующим образом:

А вот таблица с перечнем частот букв с пробелом в текстах на русском языке (необходимо отметить, что в разных источниках частоты и взаимное расположение букв в конце списка могут слегка различаться):

Буква	Частота %	Буква	Частота %	Буква	Частота %	Буква	Частота %
О	7,64	**В**	3,55	**Ы**	1,43	**Ж**	0,79
Е, Ё	7,32	**П**	3,30	**Ь**	1,38	**Х**	0,48
А	6,29	**К**	3,02	**З**	1,33	**Щ**	0,42
И	5,77	**Л**	2,99	**Й**	1,25	**Ф**	0,36
Т	5,49	**М**	2,75	**Б**	1,14	**Ш**	0,26
Н	4,90	**Д**	2,65	**Ч**	0,94	**Э**	0,23
Р	4,59	**У**	2,22	**Г**	0,83	**Ц**	0,21
С	4,04	**Я**	1,53	**Ю**	0,81	**Ъ**	0,03
ПРОБЕЛ				20,06			

Существует несколько методов атаки на шифр простой замены, которые обычно применяются одновременно и в комбинации.

1. Если шифрограмма записана теми же буквами русского алфавита, то в предположении, что перед нами сдвиговый шифр, попытаться подобрать ключ, перебирая значение сдвига от 1 до 32.

2. Если есть возможность разделить шифрограмму на слова (или её автор и так оставил пробелы), то попытаться подобрать какие-либо часто используемые слова, получив, тем самым, первые расшифрованные буквы. Например, письма часто начинаются со слов «ПРИВЕТ» или «ЗДРАВСТВУЙТЕ», а в конце также могут использоваться какие-нибудь прощальные слова. Часто можно подобрать имена тех, кто ведёт переписку, если они известны.

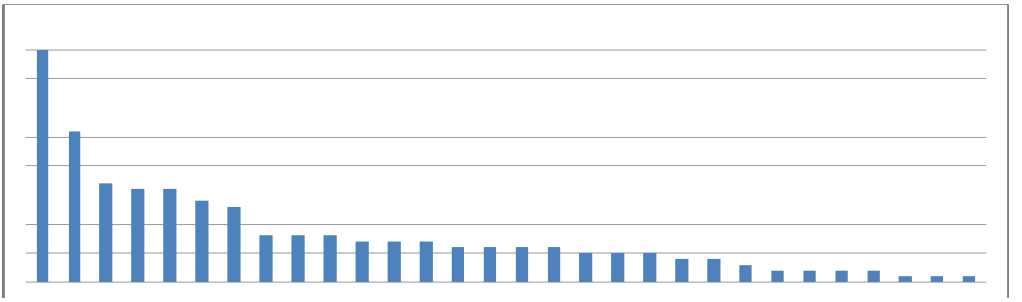
3. Наконец, при помощи подсчёта частот символов и построения

гистограммы начинается скрупулёзный подбор вариантов с выдвижением гипотез. На этом этапе очень помогает орфографический словарь или программа, в которой можно подобрать слова по части известных букв (часто такими программами пользуются для игры в кроссворды).

Попробуем применить это на практике и расшифровать ту шифрограмму, которая приведена в начале этого раздела. Для этого сразу начнём с пункта 3 приведённых правил и подсчитаем относительные частоты символов в этой загадке (относительные они потому, что считаются относительно общего количества символов в шифрограмме, а потому соответствуют только этому тексту, а не всему русскому языку в общем). Получается такая таблица:

Символ	Количество	Частота	Символ	Количество	Частота
	5	2.0 %		1	0.4 %
	8	3.2 %		7	2.9 %
	17	6.9 %		8	3.2 %
	6	2.4 %		6	2.4 %
	16	6.5 %		6	2.4 %
	16	6.5 %		5	2.0 %
	40	16.1 %		5	2.0 %
	7	2.9 %		1	0.4 %
	26	10.5 %		4	1.6 %
	6	2.4 %		2	0.8 %
	2	0.8 %		2	0.8 %
	7	2.9 %		3	1.2 %
	2	0.8 %		1	0.4 %
	13	5.2 %		4	1.6 %
	14	5.6 %		8	3.2 %

В качестве примечания необходимо отметить, что указанные относительные частоты в процентах в этой таблице округлены до десятых долей, поэтому вполне может оказаться, что их сумма не равна 100 %. И теперь, если построить гистограмму частот символов в этой таблице, предварительно отсортировав их по убыванию частоты, то получится такой график:



Можно сравнить с эталонным распределением частот для букв русского алфавита и выдвинуть очень достоверную гипотезу, что перед нами шифр простой замены для русского языка. При этом надо отметить, что в шифрограмме используется 32 символа, а не 34 (все 33 буквы плюс пробел), поэтому двух каких-то букв в шифрограмме просто нет.

Теперь необходимо начать выдвигать гипотезы. По гистограммам видно, что два первых символа имеют очень похожие частоты. Так что в качестве первой гипотезы можно предположить, что наиболее часто использующийся символ в шифрограмме – это пробел, а второй по частоте – буква «О». После этого можно посмотреть на сами символы и убедиться, что в качестве пробела мы предположили символ «пустой квадрат», что, в общем-то, похоже на пробел. Впрочем, это очень слабое подтверждение, и при серьёзном криптоанализе им лучше не пользоваться, но начинающие шифровальщики часто используют для сокрытия букв похожие на них символы.

Итак:

Гипотеза 1:  ***– это пробел,***  ***– это буква «О».***

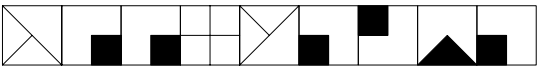
Следующим шагом выпишем всю шифрограмму в следующем виде. Каждый неизвестный нам символ обозначаем знаком подчёркивания «_», а известные символы обозначаем теми буквами, в отношении которых выдвинуты гипотезы. Вот так:

_____ _О_О_О_ _____ _ _____
__О ____ _____О ____О__
__О_____ __О _____О____О_
_ОО_____ _ __ _____ ___ _О__О
_О_О____ ____О_____ _____
_____ _О_ _____ _ _____ _ _____

_____О_О _О____ _ _О_ _____ _ _О
__О О____ ____О_____ ____ _
_____

Дальше можно было бы выдвигать следующие гипотезы о том, что символы с меньшими частотами соответствуют дальнейшим буквам русского языка в распределении, но в этом случае такие гипотезы сталкиваются с неприятным фактом – три следующих символа имеют практически одинаковые частоты.

Тем не менее, при внимательном просмотре полученного текста мы видим интересное слово: «_ОО_____», при этом шестая и девятая буквы в нём одинаковые:



Так что можно попробовать подобрать слово, для чего можно воспользоваться орфографическим словарём или специальным сервисом подбора слов. Если брать только канонические формы слов, то словарь русского языка даёт два варианта: ПООЩРЕНИЕ и СООБЩЕНИЕ. Что ж, это очень неплохо. Как минимум, у обоих этих вариантов есть одинаковые части – «-ение», и можно ввести новую гипотезу:

Гипотеза 2: при условии, что слово «_ОО_____» является словом «ПООЩРЕНИЕ» или «СООБЩЕНИЕ»,  ***– это буква «Е»,***  ***– это буква «И»,***  ***– это буква «Н».***

Для подтверждения гипотезы можно сравнить частотности предварительно выявленных букв в русском языке и в шифрограмме: Е – 7,32 % и 6,50 %, И – 5,77 % и 6,90 %, Н – 4,09 % и 5,2 %. Что ж, совсем неплохо. Проценты хоть и отличаются, но всё же очень близки друг к другу.

Теперь посмотрим, как изменится расшифровка, если добавить в неё три новых буквы:

__И_Е_ _О_О_О_ _____ _Н__Е___ __О _Е_Е _____НО _____О__ __О_И_____ __О ___И__О__ННОЕ _ОО__ЕНИЕ И __ __Н__ ___ _О_НО _О_О_Н_Е _И__О_____ _____ _И___ _ОИ __И_И И _____И И __Н_Е__ Н__НО_О _О___Е _ _О_ _И__Е И _О __О О_ЕН_ __И_О_И___ _Е_Е _ _____Е_

Очень хорошо. Просто замечательно. Обратим внимание на слово «___И__О__ННОЕ» – окончание «-нное» часто встречается в текстах на русском языке, так что наша гипотеза ещё чуть-чуть укрепилась. Кроме того, часто встречается буква «И» в изолированной позиции, то есть это обычный союз «и».

Вернёмся к двум словам – «ПООЩРЕНИЕ» и «СООБЩЕНИЕ». Можно было бы сделать просто – рассмотреть два варианта расшифровки, введя в каждом три новые буквы, после чего отвергнув одну из гипотез. Однако это упражнение оставляю читателю для самостоятельной работы – нужно ввести гипотезы 3.1 и 3.2, в которых назначить трём новым символам буквы П, Щ и Р или С, Б и Щ соответственно. И для каждой гипотезы рассмотреть вариант расшифровки с подставленными тремя новыми буквами. Но мы поступим хитрее. И в том, и в другом слове есть редкая буква «Щ». Опять посмотрим на таблицу частотности – у буквы «Щ» частотность в текстах 0,42 %, и сравнимой частотностью в шифрограмме из двух вариантов обладает только один, который отдаёт предпочтение слову «СООБЩЕНИЕ». Так что вводим новую гипотезу и проверяем её:

Гипотеза 3: при условии, что слово «_ОО__ЕНИЕ» является словом «СООБЩЕНИЕ»,  – это буква «С»,  – это буква «Б» и  – это буква «Щ».

Можно убедиться, что частотности у букв и предполагаемых для них символов опять сравнимы. А вот и новый вариант расшифровки:

__И_Е_ _О_О_О_ ____ _Н__Е_С_ __О _ЕБЕ _С___НО _____ОС_ __О_И_____ __О ___И__О__ННОЕ СООБЩЕНИЕ И __ __Н__ ___ _О_НО _О_ОБН_Е _И__О_____ _____ _И___ _ОИ __И_И И С____И И __Н_Е__ Н__НО_О БО___Е _ _О_ _ИС_Е И _О __О О_ЕН_ __И_О_И_С_ _ЕБЕ _Б___ЩЕ_

Внимательное «чтение» этого отрывка обнаруживает дважды встречающееся слово «_ЕБЕ». Несмотря на то, что на первое место в этом слове подходит очень много букв, но большинство из них, кроме букв «Н», «С» и «Т» дают очень редкие и специфические слова. Но буквы «Н» и «С» уже расшифрованы, так что следующая гипотеза будет такой:

Гипотеза 4:  ***– это буква «Т».***

В итоге выходит такое сообщение, и одна буква даёт практически всё:

__И_ЕТ _О_О_О_ ____ _Н__Е_С_ _ТО ТЕБЕ _С___НО _____ОС_ __О_ИТ_Т_ _ТО ___И__О__ННОЕ СООБЩЕНИЕ И Т_ __Н__ ___ _О_НО _О_ОБН_Е _И__О_____ _____Т_ _ИТ__ _ОИ __И_И И СТ_Т_И И __Н_Е__ Н__НО_О БО___Е _ ТО_ _ИС_Е И ТО _ТО О_ЕН_ __И_О_ИТС_ ТЕБЕ _Б___ЩЕ_

Собственно, на этом можно поставить точку – первое слово в этом отрывке расшифровывается однозначно, а за ним и вся оставшаяся шифрограмма. Если никаких гипотез относительно первого слова в голову не приходит, то подсказка проста – с каких слов обычно начинаются письма и послания? Это именно тот случай, про который было написано ранее в пункте 2 правил атаки на шифр простой замены – попытки подбора часто встречающихся слов. Так что я оставляю читателю доделать начатое, а сам перехожу к некоторым заключительным поучениям.

Заключение

Как должно быть уже понял вдумчивый читатель, ни сдвиговый шифр, ни шифр простой замены не дают никакой защиты. Это становится понятным даже по той причине, что в компьютерных системах все символы кодируются при помощи некоторых последовательностей бит, а это, фактически, и является шифром простой замены. Другими словами, шифр простой замены – это всего лишь перекодировка, даже если заменять привычные нам буквы на очень замысловатые значки. А перекодировка не меняет лингвистических отношений между символами, что и

было показано в этой статье при помощи атаки на этот шифр посредством частотного анализа.

Вывод простой: для шифрования своих секретных посланий никогда не пользуйтесь шифром простой замены.

Упражнения

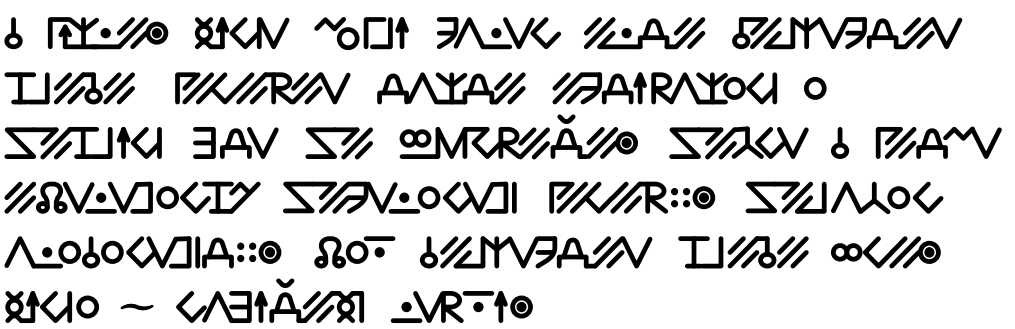
В качестве домашнего задания и самостоятельных упражнений рекомендую выполнить следующее:

1. Прочитайте эти художественные произведения:

- Дойл А. К. Пляшущие человечки.
- По А. Э. Золотой жук.
- Верн Ж. Путешествие к центру Земли.
- Душкин Р. В. Шифры и квесты.

Обобщите информацию о методах взлома шифров простой замены, полученную из этих книг.

2. Расшифруйте следующее зашифрованное послание:



Предложенный шифр не такой простой, как может показаться после чтения этой статьи. В нём есть несколько отступлений от обычного шифра простой замены. Тем не менее, он поддаётся расшифровке теми же самыми методами, тем более, что в нём встречаются буквально одинаковые последовательности символов. Просто иногда один символ обозначает несколько букв. А некоторые буквы соединены друг с другом, как при использовании письменного начертания. Надо быть внимательным и скрупулёзным, и терпение вознаградится.

3. Придумайте свой собственный шифр одноалфавитной замены, зашифруйте с его помощью письмо для друга или подруги и пошлите адресату. Если у адресата возникнут сложности при расшифровке послания, помогите ему или ей.

Юмор Юмор Юмор Юмор Юмор Юмор

Дьявольская сила

Преподаватель предлагает ученику дать определение ЭДС и слышит ответ:

– ЭДС – это работа, совершённая потусторонними силами над единицей заряда.