ПРИКЛАДНА КРИПТОЛОГІЯ 2
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №2

Криптоаналіз шифру Віженера

Необхідні теоретичні відомості

Нехай $A=\{a_0, a_1, ..., a_{m-1}\}$ – алфавіт відкритого (ВТ) та шифрованого (ШТ) текстів, що складається з m букв. Природнім чином можна замінити символи алфавіту їх номерами і перевести множину A у кільце $Z_m=\{0, 1, ..., m-1\}$ із відповідними операціями додавання та множення.

Шифр Віженера є прикладом поліалфавітної підстановки. Ключем цього шифру є послідовність r букв алфавіту $(k_0, k_1, ..., k_{r-1})$, яку підписують під ВТ, повторюючи стільки разів, скільки потрібно. Часто в якості ключа використовують якусь фразу або уривок тексту. Число r називається *періодом шифру Віженера*.

Позначимо ВТ через $X=x_0x_1x_2...x_{n-1}$, а ШТ через $Y=y_0y_1y_2...y_{n-1}$. Шифрування відбувається шляхом додавання букв ВТ до підписаних під ними букв ключа за модулем m, тобто

$$y_i=(x_i+k_{i \bmod r}) \bmod m,\ i=\overline{0,n-1}.$$

Криптоаналіз шифру Віженера починають з визначення періоду r. Зробити це можна тому, що шифр Віженера зберігає деякі статистичні властивості мови. Дійсно, розіб'ємо шифртекст Y на блоки

$$\begin{aligned}
&Y_0=y_0, y_r, y_{2r}, ...\\
&Y_1=y_1, y_{r+1}, y_{2r+1}, ...\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
&Y_{r-1}=y_{r-1}, y_{2r-1}, y_{3r-1}, ...
\end{aligned}$$

Кожен фрагмент Y_i фактично зашифрований шифром Цезаря з ключем k_i, $i=\overline{0,r-1}$. Звідси маємо, що значення частот символів у цих фрагментах будуть очікувано співпадати із значеннями імовірностей символів мови з точністю до перестановки. Це зауваження дозволяє побудувати розпізнавач періоду шифру Віженера, причому існує щонайменше два методи знаходження періоду.

Перший метод ґрунтується на понятті індексу відповідності.
Індексом відповідності тексту Y називається величина

$$I(Y)=\frac{1}{n(n-1)}\sum_{t\in Z_m} N_t(Y)(N_t(Y)-1),$$

де $N_t(Y)$ – кількість появ букви t у шифротексті Y. Якщо вважати, що текст Y обирається із множини можливих відкритих текстів випадково та рівноімовірно, то індекс відповідності буде випадковою функцією, а його математичне очікування дорівнюватиме

$MI(Y) = \sum_{t \in Z_m} p_t^2$, де p_t – імовірність появи літери t в мові. Однак, якщо Y є шифротекстом, одержаним в результаті роботи шифру Віженера, то величина індексу відповідності та його математичне очікування буде стрімко падати до мінімально можливого значення $I_0 = \frac{1}{m}$ із ростом довжини ключа r. В той же час для блоків Y_i значення індексу відповідності буде залишатись на рівні значення для мови.

Знаходження істинного значення r за допомогою індексу відповідності відбувається таким чином.

1) Для кожного кандидата $r = 2,3,...$ розбити шифртекст Y на блоки $Y_1, Y_2, ..., Y_r$.
2) Обчислити значення індексу відповідності для кожного блоку.
3) Якщо сукупність одержаних значень схиляється до теоретичного значення I для даної мови, то значення r вгадане вірне. Якщо сукупність значень схиляється до значення $I_0 = \frac{1}{m}$, що відповідає мові із рівноімовірним алфавітом, то значення r вгадане неправильно.

Замість розглядання великої сукупності індексів відповідності по кожному блоку на практиці зазвичай розглядають їх усереднене значення.

Другий метод визначення довжини ключа шифру Віженера використовує такий факт: в шифротексті на відстанях, що кратні періоду, однакові символи будуть зустрічатись частіше, ніж на будь-яких інших. Цей факт пояснюється тим, що у введених вище блоках Y_i однакові символи будуть зустрічатись із тією самою імовірністю, що й у відкритому тексті, а на інших відстанях потрібно, щоб співпадали значення відповідних сум $x_i + k_i$, що виконується із меншою імовірністю.

Отже, в цьому випадку пропонується такий порядок дій для знаходження істинного значення r: для кожного кандидата $r = 2,3,...$ обчислити значення статистики збігів символів:

$$D_r = \sum_{i=1}^{n-r} [y_i = y_{i+r}],$$

де індикатор $[y_i = y_{i+r}]$ дорівнює 1, якщо $y_i = y_{i+r}$, та 0, якщо $y_i \neq y_{i+r}$. Іншими словами, D_r дорівнює кількості пар однакових літер шифротексту, які знаходяться на відстані r символів. Для кандидатів, що рівні та кратні істинному періоду, значення D_r будуть істотно більшими за інші одержані значення.

Після встановлення значення періоду шифру подальше його розшифрування зводиться до серії розшифрувань шифрів Цезаря. Дійсно, кожен фрагмент Y_i зашифрований шифром Цезаря з ключем k_i, $i = \overline{1,r}$. Найпростіший спосіб знаходження ключа полягає в обчисленні $k_i = (y^* - x^*) \bmod m$, де y^* – буква, що частіше за всіх зустрічається у фрагменті Y_i, а x^* – найімовірніша буква у мові, якою написано відкритий текст (для російської мови це буква «о», для англійської – буква «е» тощо). Цей метод на практиці дозволяє визначити більшу частину літер достатньо довгого ключа. Якщо деяку літеру ключа було вгадано невірно (що визначається за спотворенням відкритого тексту після дешифрування), у відповідному блоці замість x^* треба брати другу, третю і т.д. за імовірністю літеру, або коригувати значення ключа відповідно до реконструкції тексту за правильно розшифрованими фрагментами. При розшифруванні деякі фрагменти будуть

встановлені неправильно, але можливі помилки легко виправляються при аналізі розшифрованого тексту в цілому.

Більш надійний метод визначення ключа полягає в наступному. Для кожного блоку Y_i обчислюється функція

$$M_i(g) = \sum_t p_t N_{t+g}(Y_i),$$

де $N_x(Y_i)$ – кількість появ букви *x* у шифротексті, p_t – імовірність появи літери *t* в мові. Те значення g, на якому функція $M_i(g)$ буде досягати максимуму, дорівнює значенню літери ключа k_i. Цей метод враховує увесь розподіл частот літер у блоці, тому він дозволяє відновити літери ключа майже безпомилково.

Порядок виконання роботи

Завдання 1. *Написати програми, які виконують шифрування та розшифрування шифром Віженера текстів російською мовою.*

Програми повинні працювати із відфільтрованими текстами (див. комп'ютерний практикум 1). З алфавіту вилучається літера «ё»; відповідно, Загальна кількість літер у алфавіті $m = 32$. У текстах літера «ё» повинна бути замінена буквою «е»; модифікуйте ваш фільтр за необхідності.

Завдання 2. *Дослідити поведінку індексу відповідності для шифротекстів.*

Для виконання завдання вам необхідно самостійно підібрати текст для шифрування (3-5 кб) та ключі довжини *r* = 2, 3, 4, 5, 6, а також довжини 10, 15 та 20 знаків.

1) Зашифруйте обраний відкритий текст шифром Віженера з обраними ключами.

2) Обчисліть індекси відповідності I_r для відкритого тексту та всіх одержаних шифротекстів і порівняйте їх значення.

Одержані значення необхідно навести у звіті таблицею та діаграмою.

Завдання 3. *Дешифрувати заданий шифротекст.*

Використовуючи наведені теоретичні відомості, напишіть програму, яка реалізує атаку на шифр Віженера та розшифруйте з її допомогою наданий шифртекст (згідно свого номеру варіанта). Зокрема, необхідно:

– визначити довжину ключа, використовуючи або метод індексів відповідності, або статистику співпадінь D_r (на вибір); при цьому потрібно перевіряти довжини ключів щонайменше до $r = 40$;

– визначити символи ключа, прирівнюючи найчастіші літери у блоці до найчастішої літери у мові;

– розшифрувати текст, використовуючи знайдений ключ; в разі необхідності скорегувати ключ.

Кожен з наведених пунктів бажано реалізувати окремою функцією.

Додатково (але не обов'язково) можна написати функцію визначення символів ключа за допомогою функції $M_i(g)$. Реалізація цієї функції буде оцінена додатковими балами.

Оформлення звіту

Звіт повинен містити такі ключові моменти:

1) усі написані вами програмні коди; дозволяється надавати посилання на github замість включення текстів програм у звіт;

2) приклад роботи шифру Віженера: відфільтрований відкритий текст на **5-6 рядків**, використаний ключ, відповідний шифротекст;

3) обрані ключі для завдання 2 та обчислені значення індексів відповідності I_r для вказаних значень r (подати у вигляді таблиці та діаграми);

4) обчислену послідовність D_r або набори значень індексів відповідності, одержаних при знаходженні довжини ключа шифру Віженера при виконанні завдання 3 (подати у вигляді таблиці та діаграми);

5) значення ключа, одержане шляхом співставлення найчастіших літер блоків найчастішій літері мови, та результат розшифрування на такому ключі – 5-6 рядків;

6) за необхідності: скореговане значення ключа та результат розшифрування на скорегованому ключі – 5-6 рядків.

Наводьте у звіті ключі шифрування як послідовності символів алфавіту. Усі ключі, які необхідно зламати при виконанні завдання 3, є змістовними фразами; це допоможе вам їх корегувати.