Laboratorium z kryptografii

Zajęcia 1: Szyfr Vigenere'a

1 Zasada działania algorytmu

Szyfrowanie Vigenere'a polega na podmianie kolejnych liter alfabetu (domyślnie angielskiego) na inne wg ustalonego klucza. Kluczem jest dowolne n-literowe słowo. Szyfrogram otrzymuje się poprzez „sumę" kolejnych liter tekstu i klucza. Dodatkowo do o każdej litery alfabetu można jednoznacznie przypisać liczbę, więc operacje „dodawania" liter wygodnie jest rozpatrywać jako (dla alfabetu angielskiego) operacje dodawania nad ciałem $\mathbb{Z}_{26}$ (liczb całkowitych $\mod 26$). Przykładowy proces szyfrowania przedstawiono w tabeli 1.

tekst: krypto

klucz: abc

szyfrogram: ksapuq

k	r	y	p	t	o		10	17	24	15	19	14
+	+	+	+	+	+		+	+	+	+	+	+
a	b	c	a	b	c		0	1	2	0	1	2
=	=	=	=	=	=		=	=	=	=	=	=
k	s	a	p	u	q		10	18	0	15	20	16

Tabela 1: Przykładowy proces szyfrowania tekstu algorytmem Vigenere'a. Pierwszy wiersz oznacza szyfrowany tekst, trzeci klucz natomiast piąty otrzymany szyfrogram.

Deszyfrowanie odbywa się poprzez „odejmowanie" liter klucza od liter szyfrogramu tj. odejmowanie w $\mathbb{Z}_{26}$, co przykładowo przedstawiono w tabeli 2.

k	s	a	p	u	q
↓	↓	↓	↓	↓	↓
10	18	0	15	20	16
-	-	-	-	-	-
0	1	2	0	1	2
=	=	=	=	=	=
10	17	24	15	19	14
↓	↓	↓	↓	↓	↓
k	r	y	p	t	o

Tabela 2: Przykładowy proces deszyfrowania wiadomości. W pierwszym wierszu przedstawiono szyfrogram, w trzecim odpowiadające mu liczby całkowite, w piątym liczby całkowite odpowiadające literom klucza natomiast w siódmy i dziewiątym odszyfrowany tekst w postaci liczb z ciała $\mathbb{Z}_{26}$ oraz liter alfabetu.

2 Kryptoanaliza szyfru

Wyznaczanie długości klucza.
Aby wyznaczyć długość klucza można posłużyć się *indeksem koincydencji* (*zgodności*) $I_c(\mathbf{x})$, który określa prawdopodobieństwo tego, że dwie litery tekstu $\mathbf{x}$ są identyczne. Przyjmując, że ilość wystąpień liter a,b,...,z w tekście $\mathbf{x}$ o długości N wynosi odpowiednio $n_0, n_1, \ldots, n_{25}$, *indeks koincydencji* definiuje się jako:

$$I_c(\mathbf{x}) \quad = \quad \frac{\sum_{i=0}^{25} n_i(n_i-1)}{N(N-1)}. \tag{1}$$

Dla języka angielskiego $I_c(\mathbf{x}) \approx 0.067$ natomiast dla tekstu całkowicie losowego $I_c(\mathbf{x}) \approx 0.038$. Ponieważ szyfr Vigenere'a jedynie przesuwa kolejne litery tekstu $\mathbf{x}$ według pewnego klucza o długości d, to $\mathbf{x}$ można zapisać w postaci:

$$\begin{array}{ccccc} x_0 & x_1 & x_2 & \ldots & x_{d-1} \\ x_d & x_{d+1} & x_{d+2} & \ldots & x_{2d-1} \\ x_2 d & x_{2d+1} & x_{2d+2} & \ldots & x_{3d-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{array}$$

Tabela 3: Kolumnowy podział szyfrogramy na fragmenty zakodowane tą samą literą klucza.

Każda z kolumn utworzona jest przez część tekstu x, która została przesunięta tą samą literą klucza. Możliwe więc jest sprawdzenie czy podział szyfrogramu na zadaną ilość kolumn powoduje, że zapisane w poszczególnych kolumnach litery pojawiają się z prawdopodobieństwem zbliżonym do prawdopodobieństwa dla danego języka (np. angielskiego) tj. czy zachodzi:

$$I_c(x_0 x_d x_{2d} \ldots) \approx I_c(x_1 x_{d+1} x_{2d+1} \ldots) \approx \ldots \approx 0.067. \tag{2}$$

Podział d najlepiej spełniający warunek (2) wyznacza długość szukanego klucza.
Uwaga! Metoda wymaga aby szyfrogram był „dostatecznie długi" a klucz „stosunkowo krótki".
Przykład
tekst: intheprinceton...
klucz: yes
szyfrogram: grlfihpmfailmr...
Podział na 2 kolumny:

nr.	szyfrogram	I_c		nr.	szyfrogram	I_c
1	glipfim ...	0.0459		1	gfpam ...	0.0654
2	rfhmalr ...	0.0475		2	rimir..	0.0688
				3	lhfl ..	0.0659

Tabela 4: Podział przykładowego szyfrogramu na dwie oraz trzy kolumny oraz wartość współczynnika koincydencji dla każdej z kolumn danego podziału.

Wyznaczanie klucza
Mając ustaloną długość klucza można podejść do próby jego złamania. Niech $n_0^0, n_1^0, \ldots, n_{25}^0$ oraz $n_0^1, n_1^1, \ldots, n_{25}^1$ określają ile razy w kolumnach (odpowiednio) $x^0 = (x_0, x_d, x_{2d}, ...)$ oraz $x^1 = (x_1, x_{d+1}, x_{2d+1}, ...)$ o długościach N^0 oraz N^1 pojawiały się znaki a,b,...,z. Wzajemnych indeksem koincydencji nazywa się wyrażenie postaci:

$$MI_c(x^0 x^1) \quad = \quad \frac{\sum_{i=0}^{25} n_i^0 n_i^1}{N^0 N^1}. \tag{3}$$

Wiedząc, że klucz jest długości d tj. składa się z liter $k = (k_0, k_1, ..., k_{d-1})$, można przy wykorzystaniu wzajemnego indeksu zgodności wyznaczyć relację pomiędzy poszczególnymi parami $k_i \leftrightarrow k_j$. Jeżeli $MI_c(x^i x^j) \approx 0.067$ to $k_i = k_j$. Jeżeli nie to należy przesuwać wszystkie dolne indeksy ciągu x^j o jeden (tj. $x_0^j \rightarrow x_1^j$ itd.) aż do uzyskania odpowiedniej wartości MI_c. Metoda ta pozwala na wyznaczenie przesunięć $y = (y_1, y_2, ..., y_{d-1})$ między poszczególnymi literami klucza np.:

$$(k_0, k_1, ..., k_{d-1}) \quad = \quad (k_0, k_0 + y_1, k_0 + y_2, ..., k_0 + y_{d-1}). \tag{4}$$

Aby odczytać klucz (znaleźć k_1) należy próbować deszyfrować szyfrogram przy pomocy kolejnych k_1, do momentu kiedy tekst jest sensowny.
Uwaga! Dobrze jest wyznaczyć wzajemne indeksy zgodności dla wszystkich kombinacji par kolumn $x^i \leftrightarrow x^j$ bo w zależności od tekstu nie musi być tak że akurat pary $x^1 \leftrightarrow x^j$ dadzą poprawne rozwiązanie.
Kontynuując przykład:

$$\begin{aligned} MI_c(x^0x^1) &= 0.0366 \\ MI_c(x^0x^1_{+1}) &= 0.0305 \\ &\vdots \\ MI_c(x^0x^1_{+6}) &= 0.0650, \end{aligned}$$

gdzie dolny indeks „$+i$" oznacza przesuniecie kolejnych częstości występowania liter w kolumnie x^1 tzn wyznacza y_1.

3 ZADANIA

1. Napisać program szyfrujący, tekst do szyfrowania ma pochodzić z pliku, klucz ustawić można w programie.

2. Napisać program, który wyznacza długość klucza (tekst w języku angielskim) i podaje wzajeumne indeksy zgodności pomiędzy wszystkimi literami klucza. *Uwaga: Nie trzeba wyznaczać relacji między literami klucza typu* $y = (y_1, y_2, \ldots y_{d-1})$.

Punktacja - łącznie 10 punktów

- 3 punkty - prawidłowo działająca pierwsza część zadania, czyli wczytanie tekstu z pliku i zaszyfrowanie go oraz wpisanie zaszyfrowanego tekstu do pliku.
- 3 punkty - wczytanie zaszyfrowanego tekstu z pliku i obliczenie indeksu zgodności dla zadanego przez użytkownika podziału na kolumny
- 4 punkty - dla zadanej przez użytkownika ilości kolumn obliczenie wzajemnego indeksu zgodności między wszystkimi kolumnami.