Klasická kryptografia

Stanislav Palúch / Tomáš Majer

Žilinská univerzite v Žiline / Fakulta riadenia a informatiky

Ceazarovské šifry

Ceasar požíval na šifrovanie túto tabuľku – posun písmena o tri znaky

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C

Zovšeobecnenie – posun o k znakov

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N

Použijeme túto reprezentáciu (kódovanie) znakov abecedy $\{\mathsf{A}, \mathsf{B}, \ldots, \mathsf{Z}\}$

$$\mathsf{A} \equiv 0,\ \mathsf{B} \equiv 1,\ \mathsf{C} \equiv 2,\ \mathsf{D} \equiv 3,\ \ldots,\ \mathsf{Y} \equiv 24,\ \mathsf{Z} \equiv 25$$

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	15	18	19	20	21	22	23	24	25

Potom možno abecedu považovať za okruh zvyškových tried $\mathbb{Z}_{26}$
s operáciami $\oplus$, $\otimes$ definovanými pre $a, b \in \mathbb{Z}_{26}$ takto:

$$a \oplus b = (a + b) \mod 26 \qquad a \otimes b = (a.b) \mod 26 \qquad (1)$$

Ceazarovské šifry

Pôvodná Ceasarova šifra:

šifrovanie: $y = E(x) = x \oplus D$ dešifrovanie: $x = D(y) = y \ominus D$

Zovšeobecnená šifra – Ceasarovská šifra s kľúčom $k \in \mathbb{Z}_{26}$:

šifrovanie: $y = E_k(x) = x \oplus k$ dešifrovanie: $x = D_k(y) = y \ominus k$

Kryptosystém je usporiadaná štvorica $(\mathcal{K}, \mathcal{M}, \mathcal{C}, \mathcal{T})$ kde

- $\mathcal{K}$ je množina kľúčov
- $\mathcal{M}$ je množina priamych textov
- $\mathcal{C}$ je množina zašifrovaných textov
- $\mathcal{T}$ je zobrazenie $\mathcal{T} : \mathcal{K} \times \mathcal{M} \to \mathcal{C}$, ktoré každej dvojici $K \in \mathcal{K}$, $M \in \mathcal{M}$ priradí zašifrovanú správu $C \in \mathcal{C}$ a také, že

V tomto systéme $\mathcal{K} = \{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}\}$, kľúč $k = \mathrm{A} \equiv 0$ je nepoužiteľný.
$\mathcal{M}$ je množina všetkých zrozumiteľných slovenských textov.

Ceazarovské šifry – Kryptoanalýza

Útok hrubou silou – vyskúšať 25 kľúčov, pokiaľ nedostaneme zrozumiteľný dešifrovaný text.

Je to „ciphertext only attack“ a „brute force attack“.

Podstatná je tu skutočnosť, že vieme rozhodnúť, či dešifrovaná správa patrí do množiny $\mathcal{M}$ správ kryptosystému.

Afinná šifra

Kľúč – dvojica prvkov $k_1,\ k_2$ okruhu $\mathbb{Z}_{26}$ taká,
že existuje prvok $k_1^{-1} \in \mathbb{Z}$ inverzný ku k_1 (t.j. $k_1 \otimes k_1^{-1} = 1 \equiv \mathrm{B}$).

$$\begin{aligned} \text{šifrovanie:} \quad y &= E_{k_1,k_2}(x) = (x \otimes k_1) \oplus k_2 \\ \text{dešifrovanie:} \quad x &= D_{k_1,k_2}(y) = (y \ominus k_2) \otimes k_1^{-1} \end{aligned}$$

Množina kľúčov $\mathcal{M}$ – množina všetkých usporiadaných dvojíc
$(k_1,\ k_2)$ taká, že existuje $k_1^{-1} \in \mathbb{Z}$.

$k_1 = 1, 3, 5, 7, 9, 11, 15, 17, 19, 21, 23, 25$ – 12 možností

$k_2 = 0, 1, 2, \ldots, 24, 25$ –26 možností

Slabý kľúč $(k_1,\ k_2) = (1, 0)$.

Množina použiteľných kľúčov obsahuje 12.26-1 = 311 prvkov.

Kryptoanalýza afinnej šifry

„Ciphertext only attack“ hrubou silou vyžaduje vyskúšať 311 kľúčov.

Known plaintext attack:

Uhádneme že $E_{k_1,k_2}(C) = P, \quad E_{k_1,k_2}(F) = H,$

t.j. $E_{k_1,k_2}(2) = 15, \quad E_{k_1,k_2}(5) = 7$

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	15	18	19	20	21	22	23	24	25

$$k_1 \otimes 2 \oplus k_2 = 15 \tag{2}$$

$$k_1 \otimes 5 \oplus k_2 = 7 \tag{3}$$

Odčítaním rovnice (2) od (3)

$$k_1 \otimes 3 = -8 \mod 26 = 18 \qquad / * 9 \equiv 3^{-1} \tag{4}$$

$$k_1 = 18 * 9 \mod 26 = 162 \mod 26 = 6 \tag{5}$$

Dosadením za k_1 do (2)

$$(6 \otimes 2) \oplus k_2 = 15 \tag{6}$$

$$k_2 = 15 \ominus 12 = 3 \tag{7}$$

Všeobecná monoalfabetická šifra

π – ľubovoľná permutácia abecedy $\mathbb{Z}_{26}$

π^{-1} – inverzná permutácia k permutácii π

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
D	P	Q	V	R	M	O	S	H	I	E	F	G	N	J	K	Y	Z	A	B	L	T	U	W	X	C

Šifrujeme znak po znaku predpisom $y = E_\pi(x) = \pi(x)$

Dešifrujeme znak po znaku predpisom $x = D_\pi(y) = \pi^{-1}(y)$

Priestor kľúčov $\mathcal{K}$ je obrovský $|\mathcal{K}| = 26! \approx 10^{27}$

Zdroje informácie

Pri kryptoanalýze všeobecnej monoalfabetickej šifry využívame skutočnosť, že množina priamych textov $\mathcal{M}$ je množinou výstupov z konkrétneho zdrpoja informácie.
Ten je charakterizovaný súborom pravdepodobností $P(x_1, x_2, \ldots, x_n)$ vujadrujúcich že zdroj v okmihoch $1, 2, \ldots, n$ vyšle postupne znaky $x_1, x_2, \ldots, x_n$

$$P() = 1 \tag{8}$$

$$\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} P(x_1, x_2, \ldots, x_n) = 1 \tag{9}$$

$$P(x_1, x_2, \ldots, x_n) = \sum_{y_1}\sum_{y_2}\cdots\sum_{y_m} P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m) \tag{10}$$

Pravdepodobnosť vyslania reťazca $x_1, x_2, \ldots x_m$ od času n

$$P_n(x_1, x_2, \ldots x_m) = \sum_{y_1}\sum_{y_2}\cdots\sum_{y_{n-1}} P(y_1, y_2, \ldots, y_{n-1}, x_1, x_2, \ldots x_m) \tag{11}$$

Zdroje informácie

Stacionárny zdroj – $P_n(x_1, x_2, \ldots x_m)$ nezávisí na n

Nazávislý zdroj – vyslanie ľubovoľnýcn dvoch slov v neprekrývajúcich sa časových intervaloch sú nezávislé javy.

Pre kryptoanalýzu všeobecnej monoalfabetickej šifry sa využívajú hlavne pravdepodobnosti $P(x_1)$, $P(x_1, x_2)$, $P(x_1, x_2, x_3)$.

Informácia jedného znaku x_i abecedy zdroja (SHANNON-HARTLEY)

$$I(x_i) = -\log P(x_i) \tag{12}$$

Stredná informácia na jedno písmeno je

$$H_1 = \sum_{x_1} -P(x_1) \log P(x_1) \tag{13}$$

Pri sledovaní dvojíc za sebou idúcich znakov je stredná informácia na dvojicu

$$H_2 = \sum_{x_1} \sum_{x_2} -P(x_1, x_2) \log P(x_1, x_2) \tag{14}$$

Zdroje informácie

Pri sledovaní n-tíc dvojíc za sebou idúcich znakov je stredná informácia na n-ticu

$$H_n = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} -P(x_1, x_2, \ldots, x_n) \log P(x_1, x_2, \ldots, x_n) \tag{15}$$

Stredná informácia na jeden znak je

$$H = \frac{1}{n} H_n \tag{16}$$

Limita tejto hodnoty pre $n \to \infty$ je entropia zdroja

$$\text{Entropia zdroja} \mathcal{H} = \lim_{n \to \infty} \frac{1}{n} H_n \tag{17}$$

Náš odhad: $\mathcal{H}(\text{slovenského jazyka}) = 1,57[\text{bit/znak}]$, $\kappa = 0,0553$.

Znaky s diakritikou

Písmeno	Pravdepodobnosť		Písmeno	Pravdepodobnosť	
	slovenčina	čeština		slovenčina	čeština
A	0,07340	0,054	Ň	0,00139	0,015
Á	0,01545	0,021	O	0,08308	0,068
Ä	0,00060	—	Ó	0,00075	0,000
B	0,01124	0,014	Ô	0,00128	—
C	0,02295	0,019	P	0,02538	0,027
Č	0,01077	0,008	Q	0,00000	0,000
D	0,02919	0,026	R	0,03783	0,029
Ď	0,00141	0,005	Ŕ	0,00006	—
E	0,06927	0,073	Ř	—	0,009
É	0,00669	0,010	S	0,04051	0,040
Ě	—	0,007	Š	0,00918	0,008
F	0,00266	0,002	T	0,04294	0,039
G	0,00222	0,002	Ť	0,00771	0,007
H	0,02050	0,020	U	0,02327	0,030
I	0,05594	0,034	Ú,Ů	0,00875	0,005
Í	0,00996	0,025	V	0,04057	0,039
J	0,01920	0,022	W	0,00011	0,000
K	0,03172	0,033	X	0,00047	0,001
L	0,02976	0,034	Y	0,01341	0,016
Ĺ	0,00006	—	Ý	0,00981	0,008
Ľ	0,00307	—	Z	0,01811	0,019
M	0,02539	0,029	Ž	0,00817	0,009
N	0,05185	0,040	␣	0,13489	0,163

Tabuľka 3.2.1. Relatívna frekvencia výskytu znakov pre zjednodušenú slovenskú a českú abecedu s medzerou

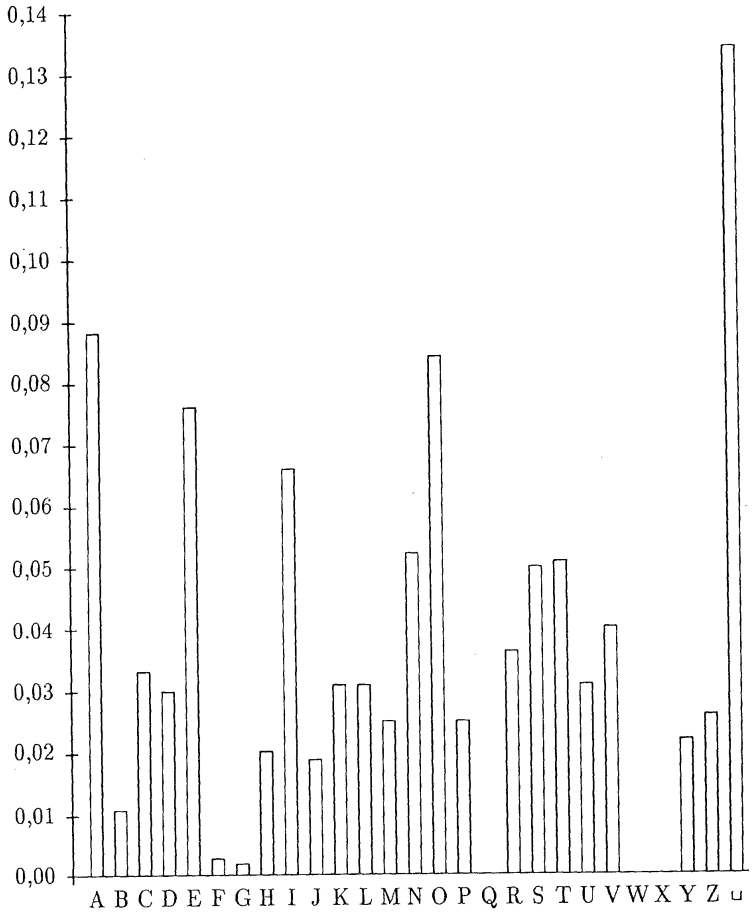
Relatívna frekvencia znakov slov. abecedy s medzerou

Písmeno	Pravdepodobnosť		Písmeno	Pravdepodobnosť	
	slovenčina	čeština		slovenčina	čeština
A	0,08945	0,065	O	0,08511	0,067
B	0,01124	0,012	P	0,02538	0,016
C	0,03372	0,024	Q	0,00000	0,001
D	0,01124	0,031	R	0,03789	0,052
E	0,07596	0,107	S	0,04969	0,050
F	0,00266	0,023	T	0,03265	0,086
G	0,00222	0,013	U	0,03202	0,021
H	0,02050	0,043	V	0,04057	0,008
I	0,06590	0,056	W	0,00011	0,016
J	0,01920	0,001	X	0,00047	0,001
K	0,03172	0,003	Y	0,02322	0,016
L	0,03189	0,028	Z	0,02628	0,001
M	0,02539	0,020	␣	0,13489	0,182
N	0,05324	0,058			

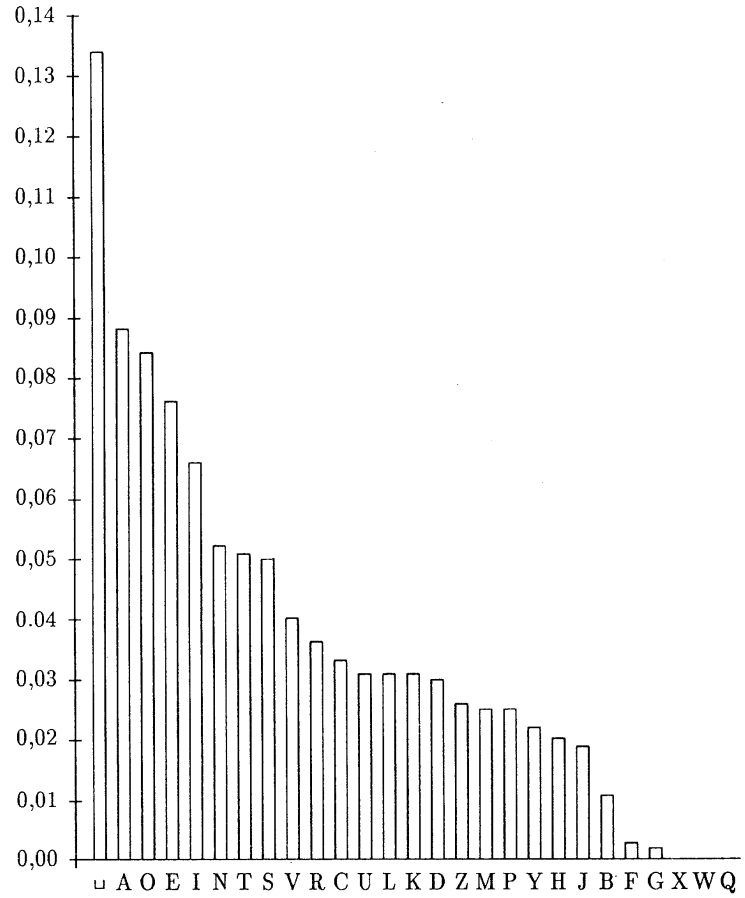
Tabuľka 3.2.2. Relatívna frekvencia výskytu znakov pre telegrafnú slovenskú a anglickú abecedu s medzerou

Zdroj nasledujúcich tabuliek a grafov: Grošek, Porubský : Šifrovanie. Grada 1992, ISBN 80-85424-62-2

Grafické znázornenie frekvencií znakov



Obrázok 3.2.1. Histogram frekvencie výskytu znakov pre telegrafnú slovenskú abecedu



Obrázok 3.2.2. Histogram usporiadaných frekvencií výskytov znakov pre telegrafnú slovenskú abecedu

Relatívna frekvencia znakov slov. abecedy bez medzery

Písmeno	Pravdepodobnosť		Písmeno	Pravdepodobnosť	
	slovenčina	angličtina		slovenčina	angličtina
A	0,11160	0,0856	N	0,05949	0,0707
B	0,01778	0,0139	O	0,09540	0,0797
C	0,02463	0,0279	P	0,03007	0,0199
D	0,03760	0,0378	Q	0,00000	0,0012
E	0,09316	0,1304	R	0,04706	0,0977
F	0,00165	0,0289	S	0,06121	0,0607
G	0,00175	0,0199	T	0,05722	0,1045
H	0,02482	0,0526	U	0,03308	0,0249
I	0,05745	0,0627	V	0,04604	0,0092
J	0,02158	0,0019	W	0,00001	0,0149
K	0,03961	0,0042	X	0,00028	0,0017
L	0,04375	0,0339	Y	0,02674	0,0199
M	0,03578	0,0249	Z	0,03064	0,0008

Tabuľka 3.2.3. Relatívna frekvencia výskytu znakov pre zjednodušenú slovenskú a anglickú abecedu bez medzerv

Počty výskytov dvojíc písmen

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	50	245	238	0	3	16	77	4	222	221	439	160	298
B	56	0	5	6	62	0	0	0	50	13	3	38	5	20
C	99	1	0	0	170	0	0	527	428	0	159	28	1	134
D	160	12	21	2	237	0	0	4	160	0	25	22	18	174
E	16	95	139	408	0	12	14	128	1	317	102	194	132	400
F	9	0	0	0	26	0	0	0	77	0	0	3	0	1
G	26	0	0	0	19	0	0	0	20	0	0	1	2	4
H	81	0	6	0	27	0	0	0	19	2	3	69	3	33
I	408	16	345	38	472	8	2	41	20	19	95	153	101	191
J	63	4	3	7	260	0	0	4	46	0	2	4	18	11
K	181	0	4	13	204	0	0	0	4	0	0	73	5	52
L	340	11	1	4	268	0	1	1	314	0	31	0	7	87
M	174	3	1	0	220	1	0	0	198	0	3	17	0	43
N	613	0	30	7	598	6	6	0	577	0	26	0	1	29
O	2	192	265	329	3	36	32	91	2	116	143	242	338	110
P	68	0	5	0	90	0	0	0	39	0	3	72	0	18
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	441	4	11	15	413	1	14	7	356	0	5	0	15	50
S	286	0	21	0	154	4	0	15	283	0	240	101	33	41
T	391	0	6	5	251	0	1	2	374	0	60	21	12	125
U	11	18	147	99	0	0	6	27	1	118	38	51	25	25
V	380	11	16	11	351	0	0	10	144	0	15	41	1	103
W	4	0	0	0	1	0	0	0	0	0	0	0	0	0
X	0	0	0	0	3	0	0	0	14	0	0	0	0	0
Y	0	20	242	2	0	0	0	19	0	2	43	8	109	17
Z	284	16	0	75	149	0	0	17	173	7	31	20	67	148
␣	650	143	275	364	50	70	26	117	190	202	433	94	293	710

Tabuľka 3.2.4. Relatívna frekvencia výskytu dvojíc znakov pre telegrafnú slovenskú abecedu (časť 1)

	O	P	Q	R	S	T	U	V	W	X	Y	Z	␣
A	4	42	0	152	229	408	22	258	3	5	0	174	1473
B	147	0	0	29	18	1	44	0	0	0	92	2	5
C	111	0	0	4	15	16	36	0	0	0	13	0	46
D	288	28	0	52	47	1	79	28	0	0	85	60	120
E	38	41	0	174	178	200	12	92	0	13	0	80	1242
F	14	0	0	5	0	0	3	0	0	0	1	1	1
G	23	0	0	14	0	0	5	0	0	0	3	0	1
H	297	1	0	30	0	14	41	3	0	0	52	0	406
I	43	31	0	18	174	273	38	125	0	0	0	109	774
J	31	4	0	4	52	9	155	7	0	0	0	0	334
K	380	0	0	72	8	182	131	20	0	0	194	0	159
L	306	0	0	0	60	8	99	4	0	0	47	1	101
M	156	15	0	6	0	6	135	0	0	0	29	0	339
N	385	0	0	1	53	66	105	2	0	0	234	6	79
O	3	54	0	318	350	155	157	577	0	0	0	253	745
P	467	0	0	534	13	3	16	0	0	0	4	0	12
Q	0	0	0	0	0	0	0	0	0	0	0	0	0
R	391	6	0	0	34	16	86	24	0	5	66	11	38
S	151	153	0	7	10	804	110	57	0	0	27	0	138
T	528	0	0	230	16	2	122	96	2	0	88	1	353
U	0	60	0	43	134	106	0	36	0	0	0	66	686
V	277	7	0	17	93	2	24	0	0	0	291	63	294
W	0	0	0	0	0	0	0	0	0	0	0	0	1
X	1	3	0	0	0	0	0	2	0	0	0	0	2
Y	0	16	0	19	85	29	16	34	0	0	0	21	549
Z	115	19	0	32	17	17	28	63	0	0	5	0	110
␣	357	864	0	248	1049	368	234	723	1	2	0	545	0

Tabuľka 3.2.4. Relatívna frekvencia výskytu dvojíc znakov pre telegrafnú slovenskú abecedu (časť 2)

Počty výskytov trojíc písmen

␣PR	455	OVA	166	ICK	131
␣NA	391	STA	166	A␣N	127
CH␣	377	␣JE	166	JE␣	127
␣A␣	362	HO␣	162	NOS	125
␣PO	302	␣ST	162	ENI	124
OST	251	A␣P	160	O␣S	122
EJ␣	248	PRI	157	A␣Z	118
YCH	233	E␣S	156	CIA	115
NE␣	231	TOR	155	OVE	115
NA␣	215	TI␣	150	E␣V	114
IE␣	210	ALI	149	LA␣	114
␣SA	210	␣DO	147	␣VE	114
␣ZA	197	␣V␣	143	EHO	113
A␣S	194	OU␣	142	␣SP	113
SA␣	186	TO␣	141	STR	112
␣VY	186	NIE	140	E␣N	111
PRE	180	␣RO	139	LI␣	110
OM␣	178	VED	137	NY␣	109
STI	176	E␣P	134	E␣A	108
IA␣	172	KTO	133	JU␣	108
␣NE	167	A␣V	132	␣KT	107

Tabuľka 4.3.1. Najčastejšie trojice v abecede s medzerou

YCH	270	IST	113	VAT	85
OST	236	ACI	111	TAT	84
OVA	197	AST	107	ENE	83
STI	181	NAS	107	EPR	82
PRE	180	EJS	105	NIC	82
STA	173	NOV	105	EDN	79
TOR	159	ICH	104	CKE	78
PRI	157	ALE	99	ENA	78
ALI	156	EST	98	ITA	78
ANI	148	SPO	98	NIA	78
NIE	141	NEJ	97	POD	78
ENI	140	LAD	95	RAV	78
VED	140	NYC	94	RED	78
KTO	138	CIT	92	AKO	77
ICK	131	IAL	91	LOV	77
NOS	128	INA	91	SKO	77
PRA	127	APR	90	TIC	77
OVE	126	OCI	90	AJU	76
EHO	122	EDO	87	STO	75
STR	118	VAN	87	VOJ	75
CIA	117	ANA	85	CHO	73

Tabuľka 4.3.2. Najčastejšie trojice v abecede bez medzery

Kryptoanalýza všeobecnej monolafabetickej šifry

Najčastejšie znaky slovenskej abecedy sú medzera a

A, O, E, I, N, T, S

Postup pri kryptoanalýze (Grošek, Porubský):

- Ak bola použitá taká permutácia, ktorý zachováva medzeru, treba analyzovať najskôr kratšie slová, ktoré poskytujú menší priestor pre kombinácie
- Hľadať charakteristické kombinácie znakov (trojice, štvorice). Tie sa najčastejšie vyskytujú na začiatkoch a na koncoch slov.
- Odhadnúť na základe „postranných informácií“, ktoré slová by sa mohli v texte vyskytnúť
- Odhadnúť, ktoré znaky sú samohlásky a ktoré spoluhlásky

Kryptoanalýza všeobecnej monolafabetickej šifry

Vytipovanie samohlások takto:

- samohlásky sú často obkolesené spoluhláskami
- spoluhlásky sú často obkolesené samohláskami
- písmená s malým počtom rôznych susedov sú často spoluhlásky a títo susedia sú často samohlásky
- ak sa dvojica XY vyskytuje často aj v opačnom poradí YX jedno z nich je samohláska
- skoro v každom normálnom slove je samohláska

Pokus o matematickú formuláciu problému kryptoanalýzy

- p_{ij} pravdepodobnosť výskytu dvojice znakov $a_i a_j$ v jazyku.
- r_{pq} relatívna početnosť znakov $a_p a_q$ v zašifrovanom texte
- $x_{ip} = \begin{cases} 1 & \text{ak } a_i \text{ bolo zašifrované na } a_p \\ 0 & \text{inak} \end{cases}$

Minimalizovať

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{p=1}^{n}\sum_{=1}^{n} x_{ip}x_{jq}(p_{ij} - r_{pq})^2$$

za podmienok

$$\begin{aligned} \sum_{i=1}^{n} x_{ip} &= 1 \quad \text{pre } p = 1, 2, \ldots, n \\ \sum_{p=1}^{n} x_{ip} &= 1 \quad \text{pre } i = 1, 2, \ldots, n \\ x_{ip} &\in \{0, 1\} \end{aligned}$$

Polyalfabetické šifry

Nevýhoda monoalfabetických šifier – relatívna početnosť zašifrovaného písmena v zašifrovanom texte závisí na pravdepodobnosti výskytu tohoto písmena v použitom jazyku.

Nová myšlienka – síce šifrovať znak po znaku, ale každý znak priameho textu inak.

Teda zašifrovaný text $y_1 y_1 \ldots y_n$ dostaneme z priameho textu $x_1 x_1 \ldots x_n$ takto:

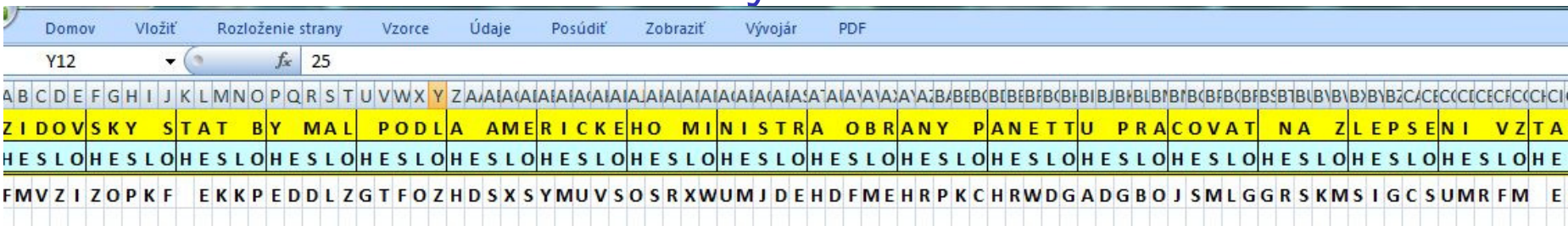
$$\begin{aligned} y_1 &= E_{K_1}(x_1) \\ y_2 &= E_{K_2}(x_2) \\ &\ldots \\ y_n &= E_{K_n}(x_n) \end{aligned}$$

Vigenèrovské šifry

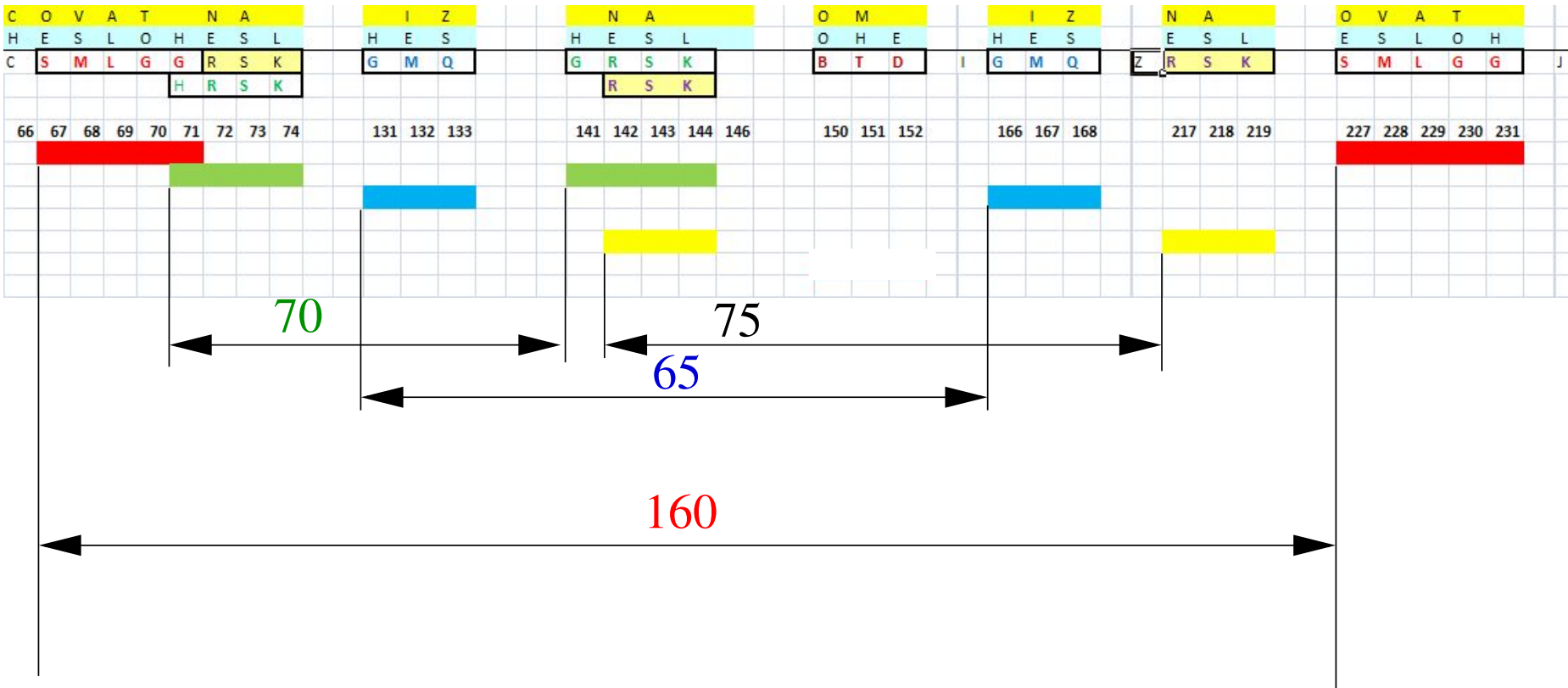
Najjednoduchší spôsob je nasledovný: zvolí sa kľúč – napr. „HESLO“ a potom zašifrovaný text $y_1 y_2 \ldots y_n$ dostaneme z priameho textu $x_1 x_2 \ldots x_n$ takto:

$$
\begin{aligned}
y_1 &= x_1 \oplus H \\
y_2 &= x_2 \oplus E \\
y_3 &= x_3 \oplus S \\
y_4 &= x_4 \oplus L \\
y_5 &= x_1 \oplus O \\
y_6 &= x_6 \oplus H \\
y_7 &= x_7 \oplus E \\
y_8 &= x_8 \oplus S \\
y_9 &= x_9 \oplus L \\
&\ldots
\end{aligned}
$$

Kasiského test na zistenie dĺžky kľúča



Hľadajú sa opakované výskyty toho istého reťazca v zašifrovanom texte.
Je nádej, že vzdialenosť výskytov je násobkom dĺžky kľúča.



Kasiského test na zistenie dĺžky kľúča (2)

Prvvý výskyt	Druhý výskyt	Offset	Trojica
67	227	160	S M L
68	228	160	M L G
69	229	160	L G G
71	141	70	G R S
72	142	70	R S K
72	217	145	R S K
131	166	35	G M Q
142	217	75	R S K
192	244	52	W B L

Dĺžka kľúča je pravdepodobne najväčším spoločným deliteľom vzdialeností rovnakých výskytov

Index koincidencie

Ak by všetky znaky abecedy $A = \{a_1, a_2, \ldots, a_q\}$ s q znakmi mali rovnakú pravdepodobnosť, potom $p(a_i) = \frac{1}{q}$.

Hľadáme spôsob, ako kvantifikovať mieru nerovnomernosti pravdepodobností.

$$\sum_{i=1}^{q}(p(a_i) - \frac{1}{q})^2$$

$$\sum_{i=1}^{q}(p(a_i) - \frac{1}{q})^2 = \sum_{i=1}^{q} p(a_i)^2 - \underbrace{2 \cdot \sum_{i=1}^{q} p(a_i)\frac{1}{q}}_{=2\frac{1}{q}} + \underbrace{\sum_{i=1}^{q}(\frac{1}{q})^2}_{=\frac{1}{q}} = \sum_{i=1}^{q} p(a_i)^2 - \frac{1}{q}$$

Pre $q = 26$

$$\sum_{i=1}^{26} p(a_i)^2 - 0,03846$$

Index koincidencie (2)

Definícia

Číslo $\sum_{i=1}^{q} p(a_i)^2$ *sa nazýva* **index koincidencie**.

Čím je index koincidenci väčší než $\frac{1}{q}$, tým viac sa rozdelenie pravdepodobnosti viac líši od rovnomerného rozdelenia.

Pre slovenskú telegrafnú abecedu bez medzery je index koincidencie asi $0,06027$, pričom $\frac{1}{q} = 0,03846$.
Pre slovenskú abecedu s diakritikou, číslami a interpunkčnými znakmi v kódovaní používanom v počítačoch sme odhadli index koincidencie na $0,0553$.

Index koincidencie (3)

Ďalší význam indexu koincidencie:

Pravdepodobnosť, že dva náhodne vybrané znaky z jazyka (resp. zo zdroja informácie) sa budú oba rovnať a_i je $p(a_i)$.
Jav, že dva náhodne vybrané znaky budú rovnaké je zjednotením nasledujúcicj disjujnktných javov

- že oba znaky sa budú rovnať a_1 – pravdepodobnosť $p(a_1)^2$
- že oba znaky sa budú rovnať a_2 – pravdepodobnosť $p(a_2)^2$
- . . . . . .
- že oba znaky sa budú rovnať a_q – pravdepodobnosť $p(a_q)^2$

Pravdepodobnosť javu, že dva náhodne vybrané znaky budú rovnaké, je súčet pravdepodobností týchto javov, a teda $\sum_{i=1}^{q} p(a_i)^2$

Odhad indexu koincidenci

Máme text (je jedno,či je priamy alebo zašifrovaný) obsahujúci n znakov. Z toho je n_1 znakov a_1, n_2 znakov a_2, atď. až n_q znakov a_q.

Počet neusporiadaných dvojíc, v ktorých sú oba znaky a_i je $\frac{n_i(n_i-1)}{2}$

počet všetkých neusporiadaných dvojíc znakov v danom texte je $\frac{n(n-1)}{2}$.

Pravdepodobnosť, že oba znaky budú a_i je teda

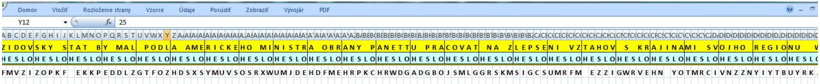
$$p(a_i)^2 \approx \frac{n_i(n_i-1)/2}{n(n-1)/2} = \frac{n_i(n_i-1)}{n(n-1)}$$

Pravdepodobnosť $\sum_{i=1}^{q} p(a_i)^2$ javu, že oba znaky budú rovnaké, odhadneme hodnotou

$$\kappa = \frac{\sum_{i=1}^{q} n_i(n_i-1)}{n(n-1)} \tag{18}$$

Ide o monoalfabetickú alebo polyalfabetickú šifru?

Ak ide o monoalfabetickú šifru, index koincidencie bude zašifrovaného textu bude blázky indexu jazyka (napr. 0,06). Ak bude bližšie hodnote $1/q$, jedná sa pravdepodobne o poyalfabetickú šifru.



Priamy text mal $\kappa = 0,06216$.
Zašifrovaný text $\kappa = 0,04116$, pričom $1/27 = 0,03704$
Záver: s veľkou pravdepodobnosťou ide o polyalfabetickú šifru.

Zisťovanie dĺžky kľúča metódou koincidencie

Majme dva priame texty
$\mathbf{r} = r_1 r_2 \dots r_n$, $\mathbf{s} = s_1 s_2 \dots s_n$
Pravdepodpobnosť, že $r_i = s_i$ je index koincidencie slov. jazyka κ.
Nech teto texty sú zašifrované znak po znaku rovnakým kľúčom.
Príslušné zašifrované texty sú"
$\bar{\mathbf{r}} = T_1(r_1) T_2(r_2) \dots T_n(r_n)$,
$\bar{\mathbf{s}} = T_1(s_1) T_2(s_2) \dots T_n(s_n)$.
Pravdepodobnosť javu , že $T_i(r_i) = T_i(s_i)$, je rovnaká ako pravdepodpobnosť javu, že $r_i = s_i$, lebo $T_i(r_i) = T_i(s_i)$ práve vtedy, keď $r_i = s_i$. Teda

$$P(T_i(r_i) = T_i(s_i)) = P(r_i = s_i) = \kappa$$

Ak sledujeme počet zhôd na rovnakých miestach zašifrovaného a posunutého zašifrovaného textu, počet zhôd by mal nápadne stúpnuť, ak je posun o násobok dĺžky kľúča.

Friedmanov test

Zoradíme zašifrovaný text $\mathbf{s} = s_1 s_2 \ldots s_n$ do tabuľky

1	2	...	...	k
s_1	s_2	...	...	s_k
s_{k+1}	s_{k+2}	...	...	s_{2k}
s_{2k+1}	s_{2k+2}	...	...	s_{3k}
s_{3k+1}	s_{3k+2}	...	...	s_{3k}

Ak sa k rovná dĺžke kľúča jednotlivé stĺpce sú už zašifrované monoalfabeticky, a vtedy by indexy koincidencie počítané zvlášť pre každý stĺpec mali stúpnuť.
Nech $Z_1,\ Z_2, \ldots, Z_t$ sú najčastejšie znaky v prvom stĺpci.
Medzi nimi je s veľkou pravdepodobnosťou zašifrované niektoré z najčastejších znakov - A, O, E, I.
Preto sa medzi znakmi typu $Z_i - A$, $Z_i - O$,$Z_i - E$, $Z_i - I$ nachádza 1. znak kľúča.