2.1 Frekvenčna kriptoanaliza v angleškem jeziku

Dešifriranje šifropisov izdelanih po substitucijskih postopkih je različno zahtevno. Pri Cezarjevi šifri je dokaj preprosto, preizkusiti moramo le L-1 ključev (premikov abecede), kar je izvedljivo tudi brez uporabe računalnika. Že pri preprosti monoalfabetski zamenjalni šifri z naključno abecedo je ta naloga mnogo zahtevnejša, saj je vseh možnih ključev kar $L!$. Na primer, pri dolžini abecede L=25 je to 15.511.210.043.330.985.984.000.000 ključev, kar ustreza 88 bitnemu ključu (2^{88}). Kljub temu ta metoda še zdaleč ni varna, saj se navadno tako šifriranje razbije že s preprosto frekvenčno analizo. Pogoji za to so predvsem:

- da poznamo jezik čistopisa,
- da poznamo relativne frekvence znakov in njihovih najpogostejših kombinacij v tem jeziku,
- imamo dovolj dolg šifropis (navadno zadošča že okoli 100 znakov).

V angleških besedilih veljajo relativne frekvence znakov, ki jih najdemo v tabelah, ki so priložene v dodatku in podajajo:

- relativno frekvenco posameznih znakov,
- relativno frekvenco začetnic besed,
- relativno frekvenco *bigramov* (parov znakov),
- pogostost pojavljanja *trigramov* (kombinacij treh znakov).

Naloga:

1. Iz podanega šifropisa poskušaj s pomočjo frekvenčne analize ter spodnjih predpostavk in pogojev pridobiti čistopis.

- Analizo bomo izvajali nad besedili v angleščini (lažje kot v slovenščini).
- Šifropis vsebuje le velike črke, je brez ločil, besede so med seboj ločene s presledki, ki pa so ravno tako šifrirani (zamenjani z nekim drugim znakom abecede).
- Čistopis je šifriran z neznano enoabecedno (monoalfabetsko) zamenjalno šifro.
- Pri analizi si bomo pomagali s pythonom in funkcijami, izdelanimi na prejšnjih vajah.

Šifropis:

HPBKVLZRSOVGZYLBGBDMFDRHRMRHSPBKHZY VBHDBGBJ RYSNBSUB
PKSNHPOBFLBQYHKYBMPHRDBSUBZAGHPR WRBGV BV ZAGK NBQHRYBKHZY VR
WRBGKKSVNHPOBRSBGBV OMAGVBDLDR JBRY BMPHRDBJGLBF BDHPOA BA RR VDBRY
BJSDRBKSJJSPBZGHVDBSUBA RR VDBRVHZA RDBSUBA RR VDBJHWRMV DBSUBRY BGFSI
BGPNBDSBUSVRYBRY BV K HI VBN KHZY VDBRY BR WRBFLBZ VUSVJHPOBGPBHPI VD
BDMFDRHRMRHSP

Navodilo:

Po izvedeni frekvenčni analizi šifropisa se lahko lotimo razbijanja (dešifriranja) po naslednjem postopku, ki predstavlja zgolj eno izmed mnogo možnih različic:

1. Pripravimo si funkcijo *fnAnaliza(niz)*, ki opravi frekvenčno analizo (štetje znakov, bigramov in trigramov). Ta naj za vsak znak, bigram in trigram, ki nastopa v šifropisu, ugotovi število pojavitev in jih izpiše.
2. Pripravimo funkcijo *fnSubstitucija*, izdelano na prejšnjih vajah, kjer za ključ pri neznanih znakih (za enkrat še vseh) uporabimo črtico »–«.

3. V šifropisu poiščemo najbolj pogost znak in ga zamenjamo s presledkom, ki je na splošno v čistopisu najbolj pogost znak,
4. preostale znake, ki so vsi črke angleške abecede, razvrstimo po frekvenci in vnesemo dva do tri v dešifrirni ključ, ter preverimo ali so zamenjave smiselne, in če niso, poskusimo drugo kombinacijo,
5. preverimo najbolj pogoste bigrame in trigrame v šifropisu in jih poskusimo nadomestiti s čistopisom,
6. poskusimo uganiti posamezne besede čistopisa.

Zgornje korake prvič izvedemo po vrstnem redu, potem pa jih logično uporabljamo v različnih kombinacijah, glede na izide prejšnjih korakov. Na primer, če uganemo eno izmed besed čistopisa lahko takoj zatem naredimo korak 2, 3 ali 4, seveda odvisno od tega kateri izmed njih nam v tem trenutku najbolj ustreza ali je najlažje izvedljiv.

Pri postopku analize angleških besedil si pomagamo tudi s spodnjimi namigi:

- pri uganjevanju besed se najprej osredotoči na krajše, ki vsebujejo 1 do 4 znake:
  - edini enoznakovni besedi v angleščini sta *a* in *I*,
  - najpogostejše dvoznakovne besede so:
    of, to, in, it, is, be, as, at, so, we, he, by, or, on, do, if, me, my, up, an, go, no, us, am,
    opazimo, da imajo vse dvoznakovne besede en samoglasnik,
  - najpogostejši triznakovni besedi sta: *the* in *and*, ki jima sledijo *for*, *was* in *his*,
  - najpogostejša beseda iz štirih znakov je *that*,
- poišči podvojene znake – v angleščini so najpogostejši *ll, ee, ss, oo tt, ff,* and *mm*,
- zamenjave in ugibanja niso vedno pravilni – bodite pripravljeni na popravke že »*dešifriranih*« znakov,
- zelo uporabni pripomočki so še vedno tudi svinčnik, radirka in papir.

Rešitev:

2.2 Frekvenčna kriptoanaliza v slovenskem jeziku

Naloga:

2. S funkcijo *fnAnaliza* in daljšim besedilom v slovenskem jeziku pripravite frekvenčno analizo slovenskega jezika.

Navodilo:

Kot primer daljšega besedila vzamemo Cankarjevo delo »Na klancu« , ki je dostopno za prenos v .txt formatu na: https://sl.wikisource.org/wiki/Na_klancu

Pred izdelavo frekvenčne analize še:

- pretvorimo vse črke v velike tiskane
- odstranimo (prezremo) vse znake, ki NISO:
  - črke slovenske ali angleške besede
  - presledek

Naloga:

3. Spodnji čistopis šifrirajte z monoalfabetsko zamenjalno šifro s poljubnim ključem. Dobljeni šifropis poskusite brez uporabe ključa dešifrirati na podlagi pravkar ustvarjene frekvenčne analize slovenskega jezika, pri čemer upoštevajte zgolj relativne frekvence posameznih znakov. Ugotovite uspešnost »avtomatskega« dešifriranja:
   a. Koliko odstotkov znakov ključa je algoritem pravilno dešifriral?
   b. Koliko odstotkov znakov šifropisa je bilo pravilno dešifriranih?
   c. Bi bila pri drugačnem paru čistopis–šifropis uspešnost verjetno večja ali manjša? Zakaj? Katera lastnost ključno vpliva na zmožnosti dešifriranja s kriptoanalizo?

ČISTOPIS:

JERMAN KAKŠNA JE BILA BESEDA KI STE JO REKLI KAJ STE IZNAŠLI DA BI ME DO DNA RANILI HLAPCI ZA HLAPCE ROJENI ZA HLAPCE VZGOJENI USTVARJENI ZA HLAPČEVANJE GOSPODAR SE MENJA BIČ PA OSTANE IN BO OSTAL NA VEKOMAJ ZATO KER JE HRBET SKRIVLJEN BIČA VAJEN IN ŽELJAN GLAS OD OKNA ALI STE MOŽJE DA POSLUŠATE GLASOVI VSIVPREK LAKOT TI NAS BOŠ ZMERJAL NA CESTO UDARITE JERMAN STOPI Z ENO NOGO NA STOL Z DRUGO NA MIZO HLAPCI MED VAS BI KRISTUS NE PRIŠEL Z BESEDO PRIŠEL BI Z BIČEM NOREC KI SE JE NAPRAVIL DA BI VAM ODKLEPAL TO PAMET DEVETKRAT ZAKLENJENO GLASOVI ZADOSTI JE KDO JE HLAPEC ŽENSKI GLASOVI ALI STE MOŽJE ALI NISTE ALI DA GA ME RAZPRASKAMO

Rešitev:

Dodatek - Frekvenčna analiza besedila v angleškem jeziku

Znak	Relativna frekvenca
e	12,70
t	9,05
a	8,16
o	7,50
i	6,96
n	6,74
s	6,32
h	6,09
r	5,98
d	4,2
l	4,0
c	2,7
u	2,7
m	2,4
w	2,3
f	2,2
g	2,0
y	1,9
p	1,92
b	1,49
v	0,97
k	0,77
j	0,15
x	0,15
q	0,09
z	0,07

Znak	Relativna frekvenca začetnic besed
t	16,67
a	11,60
s	7,76
h	7,23
w	6,75
i	6,29
o	6,26
b	4,70
m	4,37
f	3,78
c	3,51
l	2,71
d	2,67
p	2,55
n	2,37
e	2,01
g	1,95
r	1,65
y	1,62
u	1,49
v	0,65
j	0,60
k	0,59
q	0,17
x	0,04
z	0,03

Bigram	Relativna frekvenca
th	1,52
he	1,28
in	0,94
er	0,94
an	0,82
re	0,68
nd	0,63
at	0,59
on	0,57
nt	0,56
ha	0,56
es	0,56
st	0,55
en	0,55
ed	0,53
to	0,52
it	0,50
ou	0,50
ea	0,47
hi	0,46
is	0,46
or	0,43
ti	0,34
as	0,33
te	0,27
et	0,19
ng	0,18
of	0,16
al	0,09
de	0,09
se	0,08
le	0,08
sa	0,06
si	0,05
ar	0,04
ve	0,04
ra	0,04
ld	0,02
ur	0,02

Uvrstitev	Trigram
1	the
2	and
3	tha
4	ent
5	ing
6	ion
7	tio
8	for
9	nde
10	has
11	nce
12	edt
13	tis
14	oft
15	sth
16	men