Dokumentace k projektu č. 1 - Vigenerova šifra

Autor: Tomáš Mlčoch (xmlcoc06@stud.fit.vutbr.cz)

Tato dokumentace popisuje implementaci metod použitých v projektu implementujícím cracker Vigenerovy šifry do předmětu KRY (Kryptografie).

Charakteristiky vstupu

Během načítání vstupu ve formě zašifrovaného textu se průběžně vytváří statistické informace o tomto textu. Statická metoda CipherTextFromStdin() třídy CipherTextFactory, která načítá vstup znak po znaku a provádí jeho filtraci (ignoruje jiné znaky než 26 písmen latinské abecedy a velká písmena převádí na malá), pro to používá třídu VigenereStreamAnalysis.

Třída VigenereStreamAnalysis uchovává počet jednotlivých načtených znaků v poli Lalphabet s názvem letter_frequencies_. Zároveň tato třída používá třídy NgramCounter, která buduje hashovací tabulku, obsahující všechny 3gramy vyskytující se v textu. Po skončení načítání vstupu se na třídě NgramCounter zavolá metoda find_longer_ngrams(), která z tabulky 3gramů odstraní všechny 3gramy s pouze jedním výskytem. Nad touto tabulkou pak vytvoří tabulky všech možných 4gramů (protože n+1gram může být pouze tam kde byl předtím ngram). Pak se krok opakuje, z tabulky všech 4gramů se odstraní 4gramy vyskytující se pouze jednou a na základě zbylých ngramů se zkouší hledat 5gramy a tak dokola dokud n < NGRAMMAPS + 2.

Po skončení načítání tedy máme v instanci třídy VigenereStreamAnalysis počty jednotlivých znaku latinské abecedy v textu a několik hashovacích tabulek (pro 3gramy, 4gramy, 5gramy, …), kde klíč je ngram a hodnota vector indexů, na kterých se tento ngram v textu vyskytuje.

Friedmanův test

Hodnota Friedmanova testu je vypočítána dle vzorce:

$$\frac{\kappa_p - \kappa_r}{\kappa_o - \kappa_r}$$

Kde:

- k_p je index koincidence řetězce v angličtině (přibližně 0,067)
- k_r je 1/počet znaků použité abecedy (v našem případě tedy 1/26 = 0,0385)
- k_o je vypočítáno jako:
  $$\kappa_o = \frac{\sum_{i=1}^{c} n_i(n_i - 1)}{N(N-1)}$$
  Kde n_i je počet výskytů i-tého znaku v abecedě a N je celkový počet všech znaků zašifrované zprávy.

Kasiského test

Pro spočítání výsledků kasiského testu jsou použity hashovací tabulky ngramů v textu. Tyto tabulky jsou procházeny postupně, začíná se u té s nejdelšími ngramy. U jednotlivých ngramů se vypočítají jejich vzájemné vzdálenosti a poté největší společný dělitel těchto vzdáleností. Tento největší společný dělitel je pravděpodobně jedena z možných délek klíče. Při prvním průchodu se berou v úvahu pouze ngramy co mají více než 4 výskyty v šifrovaném textu a při dalších průchodech se toto číslo snižuje. Tyto další průchody vůbec nemusí nastat, pokud už při prvním průchodu hashovacích tabulek bude nalezen dostatečný počet možných délek klíče. Stejně tak při tomto procházení hashovacích tabulek ani nemusí dojít k projití všech. Pokud už v té první

zpracovávané nalezneme dostatek hodnot možných vzdáleností klíče, je cyklus procházení těchto tabulek ukončen a tabulky s kratšími ngramy nejsou vůbec zpracovávány.

Výsledkem kasiského testu je vektor možných délek klíče, seřazený vzestupně podle počtu výskytu jednotlivých odhadnutých délek klíče. Posledním prvkem vektoru je tedy nejpravděpodobnější délka klíče, která je poté i použita ve výstupu programu jako výsledek kasiského testu.

Určení délky hesla

Pro určení délky hesla se používá výsledek Kasiského testu. Konkrétně hodnota s nejčastějším výskytem (v seřazené posloupnosti výsledků Kasiského testu jde o poslední prvek).

Odhalení hesla

Pro odhalení hesla je použito kryptoanalýzy založená na indexu koincidence. Podle odhadnuté délky hesla je šifrovaný text rozdělen na několik „pod-zpráv“. U každé z těchto pod-zprávy předpokládáme, že všechny její znaky jsou šifrovány pomocí stejného písmene klíče a proto můžeme použít techniku frekvenční analýzy, kdy na základě frekvencí jednotlivých písmen v šifrované zprávě a na základě znalostí frekvencí jednotlivých písmen v zdrojovém jazyku odhalíme posunutí abecedy. Z tohoto posunutí zjistíme znak, kterým byla pod-zpráva zašifrována. Takto zanalyzujeme všechny pod-zprávy a dostaneme výsledný klíč, který byl použitý k zašifrování zprávy.