Dokumentácia k projektu 1 z predmetu Kryptografie

Timotej Ponek, xponek00@stud.fit.vutbr.cz

Spracovanie argumentov

Na spracovanie argumentov som použil riešenie z projektov, ktoré som vypracoval do predmetov IPK a ISA. Keďže pri testovaní budú použité iba korektné argumenty, toto riešenie je asi trochu zbytočne heavy-weight, ale dovoľuje napríklad zadať argumenty poprehadzované. Ak nie je špecifikovaný vstupný súbor, za vstup je považovaný posledný zadaný argument. Jeden z módov programu (-e, -d alebo -c) musí byť vždy zadaný, inak sa vypíše nápoveda.

Šifrovanie a dešifrovanie so znalosťou kľúča

Implementácia šifrovania je založená čisto na rovnici pre šifrovanie zo zadania. Každý znak vstupného textu sa šifruje samostatne a následne je okamžite zapísaný do výstupného súboru alebo na stdout. Pre dešifrovanie bolo potrebné vytvoriť funkciu, ktorá vypočíta multiplitívny inverz zadaného parametru 'a'. Tu som neobjavoval nový algoritmus, ale použil "extended euclidean" algoritmus, ktorého implementácie sú voľne dostupné na internete. Okrem výpočtu multiplikatívnej inverzie 'a', je implementácia dešifrovania založená na rovnici zo zadania, a až na použitie inej rovnice funguje rovnako ako šifrovanie (čo sa týka spracovania vstupu a generovania výstupu).

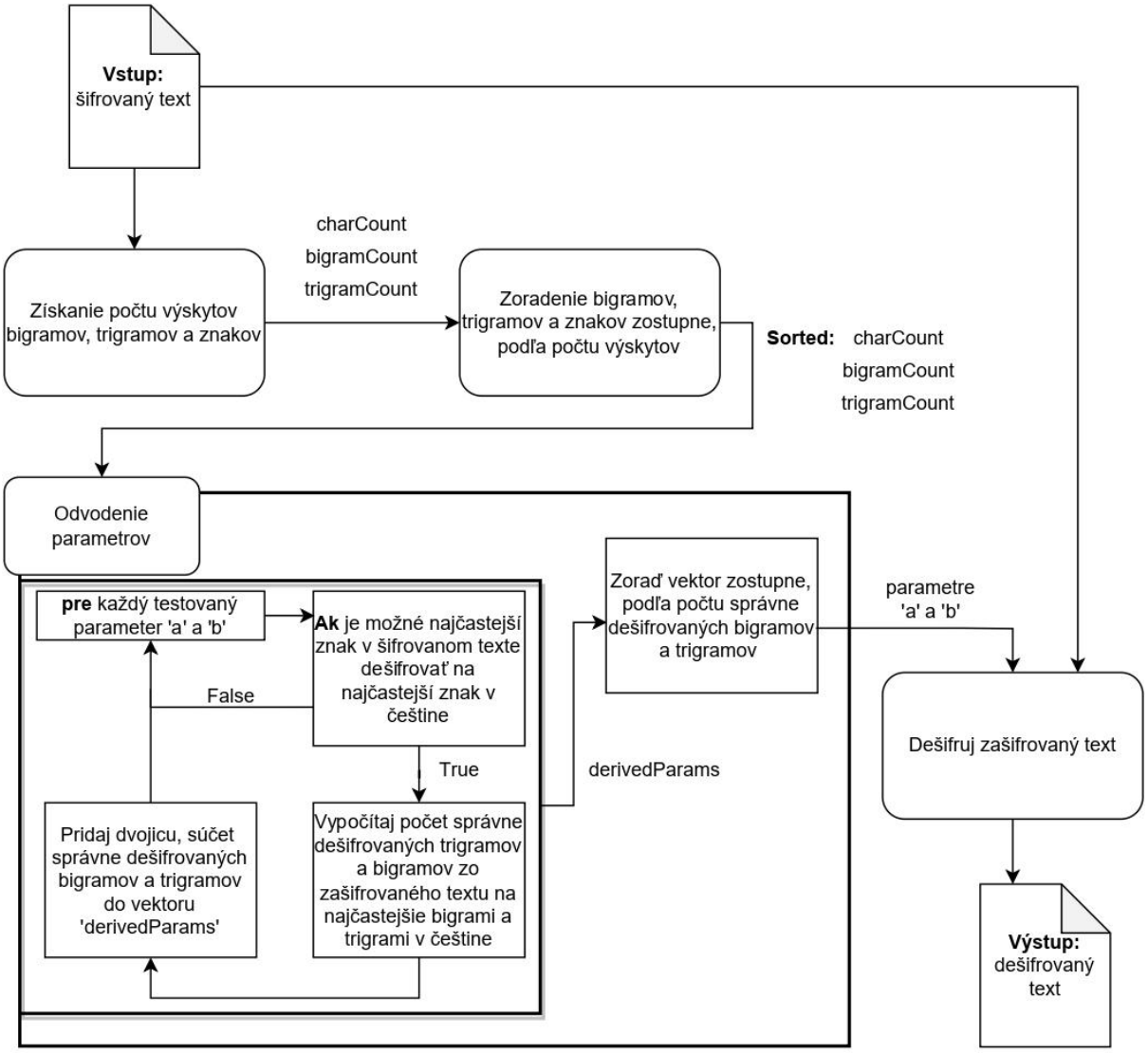
Frekvenčná analýza

Prvé myšlienky

Pri implementácií frekvenčnej analýzy som najskôr vytvoril riešenie, ktoré skúšalo s rôznymi parametrami 'a' a 'b' zašifrovať najfrekventovanejšie písmeno v českej abecede (vybral som 'A', ale v rôznych zdrojoch sa môže líšiť) na najfrekventovanejšie písmeno v zašifrovanom texte (v mojom texte, číslo 222153 to bolo 'F'). Dvojice parametrov, ktoré vyhovovali som si ukladal do vektoru, a následne som dešifroval text s využitím každej dvojice parametrov. Text som ukladal do samostatných súborov s názvami, ktoré obsahovali použité číselné parametre 'a' a 'b' (napr. pre 'a'=9 a 'b'=5 to bolo tmp95). V jednom zo súborov sa nachádzal korektný český text, z čoho som bol šťastný a napadlo mi, že by som mohol správne parametre získať tak, že so všetkými predvypočítanými hodnotami parametrov budem dešifrovať text a kontrolovať, či sa jednotlivé slová, ktoré dešifrujem, nachádzajú medzi najčastejšími českými slovami. Toto riešenie som zavrhol potom, ako som skúsil dešifrovať iný text. Použil som rovnaký postup ako uvedený vyššie, ale ani jeden text neobsahoval korektný český text. Bolo to preto, lebo v tomto texte bolo najfrekventovanejšie české písmeno 'E' a nie mnou zvolené 'A'. Ďalej mi teda napadlo kontrolovať 3 najfrekventovanejšie znaky v šifrovanom texte s tromi najfrekventovanejšími znakmi v češtine, rôzne to kombinovať, ale došlo mi, že tadiaľto cesta nevedie.

Použité riešenie

Po následnom pátraní po ďalších možnostiach, som sa rozhodol počítať, okrem výskytu znakov, aj výskyt bigramov a trigramov v šifrovanom texte. Vytvoril som novú metódu na odvodenie parametrov (`DeriveParameters`), ktorá znovu testuje všetky možné hodnoty 'a' a 'b'. Tieto hodnoty sa testujú v cykle, ktorý funguje tak, že sa najčastejšie vyskytovaný znak v šifrovanom texte dešifruje práve testovanými parametrami a skontroluje sa, či daný dešifrovaný znak odpovedá jednomu z najčastejších znakov v českej abecede (vybral som znaky 'A', 'E', 'O', 'R', 'I', 'N'), ak áno, potom sa následne získa počet najčastejších trigramov a bigramov v šifrovanom texte, ktoré je možné dešifrovať na najčastejšie bigrami a trigrami v českom texte (funkcie `TryTrigrams` a `TryBigrams`). Následný súčet bigramov a trigramov, ktoré sa podarilo dešifrovať na najčastejšie bigrami a trigrami v češtine, je uložený do vektoru spoločne s použitými parametrami pre dešifrovanie. Po otestovaní všetkých možných parametrov sa vektor zoradí, kde prvý prvok zodpovedá parametrom, s ktorými sa podarilo dešifrovať čo najviac bigramov a trigramov. Tieto parametre sú následne prehlásené za parametre s ktorými bol pôvodný text zašifrovaný, a šifrovaný text sa pomocou nich dešifruje rovnakým spôsobom ako pri dešifrovaní so znalosťou kľúča (ako kľúč sa použijú odvodené parametre).



obrázok 1 - zjednodušený control flow graf dešifrovania textu bez znalosti kľúčov

Zdroje

Najčastejšie bigrami a trigrami v češtine boli zvolené z nasledujúceho zdroja:
https://nlp.fi.muni.cz/cs/FrekvenceSlovLemmat
Najčastejšie znaky v českej abecede boli zvolené na základe výskumu zo zdroja:
http://sas.ujc.cas.cz/archiv.php?art=2913