Kryptoanalytické rozlišovacie charakteristiky jazykov

Autor: Milan Plančík, Vedúci práce: Ing. Pavol Zajac
plancik.m@gmail.com

Abstrakt

Cieľom mojej práce je poukázať na niektoré zo spôsobov odhadnutia jazyka zašifrovaného textu pomocou jednoduchých metód, ak vieme že text je zašifrovaný práve substitučnou šifrou.

1. Úvod

Potreba utajenia informácii bola pre ľudstvo dôležitá už veľmi dávno. Jej význam mnohonásobne vzrástol v období prvej a druhej svetovej vojny a vznikom počítačov a počítačových sietí. Jednou z možností ako utajiť informáciu bolo šifrovanie. S príchodom šifrovania ale prišla aj otázka, ako narušiť šifrovací systém – zaoberá sa ňou kryptoanalýza.
Jednou z metód šifrovania je substitučná metóda , v ktorej zamieňame znaky, alebo skupiny znakov otvoreného textu inými znakmi, alebo skupinami písmen predpísaným spôsobom. Text zašifrovaný substitučnou šifrou sa dnes dá pomocou viacerých spôsobov dešifrovať. Lúštenie šifry je značne závislé na tom, v akom jazyku je text napísaný. Zaujímalo ma teda, či sa dá reálne odhadnúť, v akom jazyku bola pôvodná správa napísaná, bez toho aby sme ju museli vopred vylúštiť. Takáto pomôcka by mohla významne zjednodušiť ďalšie kroky pri dešifrovaní, keďže by sme poznali jazyk správy a teda by sme sa mohli hneď sústrediť na niektoré ďalšie špeciálne vlastnosti jazyka a zašifrovaný text vylúštiť rýchlejšie. Jednou z možností uvádzaných v literatúre je použitie indexu koincidencie. Bližšie sa mu budeme venovať v časti 3.1.

2. Analýza problému

Na vytvorenie štatistík, ktoré by mohli vypovedať o správnosti určenia jazyka som potreboval dostatočne dlhý text pre každý jazyk. Použité texty sú vybrané zmiešane z rôznych internetových článkov a knižiek, pre každý jazyk individuálne a náhodne. Pre zlepšenie výpovednej hodnoty štatistík boli dané texty prevedené do slovenskej telegrafnej abecedy bez medzier. Takto upravené texty boli skrátene na dĺžku 100 000 znakov.
Pre zachovanie logickej postupnosti by sa dalo riešenie rozdeliť do nasledujúcich častí:

1. Vypočítanie celkovej štatistiky textu.
2. Výpočet jednotlivých štatistík textu.
3. Grafické zobrazenie vypočítaných parametrov.

V 1. kroku som pomocou vlastného programu určil celkovú štatistiku textu. V 2. kroku som potreboval počítať jednotlivé čiastkové výpočty, na ktoré som znova využil vlastný program, čo nie je nutnosťou. Je to však výhodnejšie, lebo má viaceré kroky zautomatizované. V 3. kroku som potom zo získaných výpočtov v tabuľkovom procesore určoval podľa nižšie uvedeného postupu jazyk textu. Nakoniec som pre názornosť určovania jazyka zobrazil výsledky aj graficky.

3. Riešenie

Keďže máme zadané, že text je zašifrovaný substitučnou šifrou, na jeho lúštenie je možné použiť štatistickú analýzu, alebo frekvenčnú analýzu. Pri týchto metódach sa využíva fakt, že sa nezmení počet znakov a ja som sa zameral hlavne na frekvenčnú analýzu. Pri frekvenčnej analýze môžeme využiť index koincidencie (môžeme si ho predstaviť ako pravdepodobnosť, že dve náhodne zvolené písmená v texte sú zhodné), hľadanie vzorov, hľadanie zdvojenia písmen atď.
Sústredil som sa na porovnávanie indexu koincidencie a na hľadanie vzorov, ktoré podľa mňa môžu vo všeobecnosti podať najobjektívnejšie výsledky.

3.1 Riešenie pomocou Indexu koincidencie

3.1.1 Celkový index koincidencie

Prvým krokom bolo určenie celkového indexu koincidencie pre každý z jazykov. Vzorec pre výpočet indexu koincidencie I je uvedený vo vzťahu (1).

$$I = \frac{1}{n(n-1)} \sum_i n_i(n_i - 1) \quad (1)$$

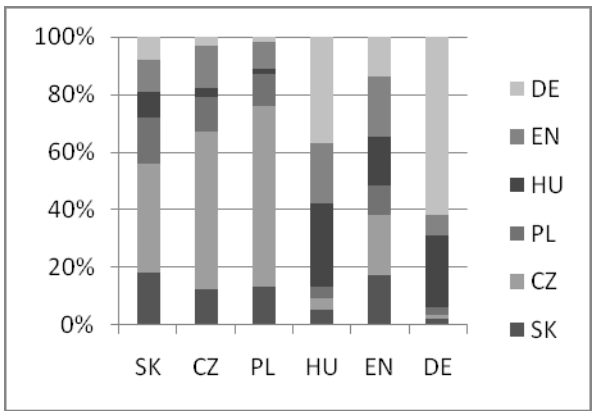
Získaná hodnota pre každý z jazykov bola východisková, teda predpokladal som, že index koincidencie celého textu je najpresnejšia hodnota indexu koincidencie pre daný jazyk.

3.1.2 Index koincidencie textov

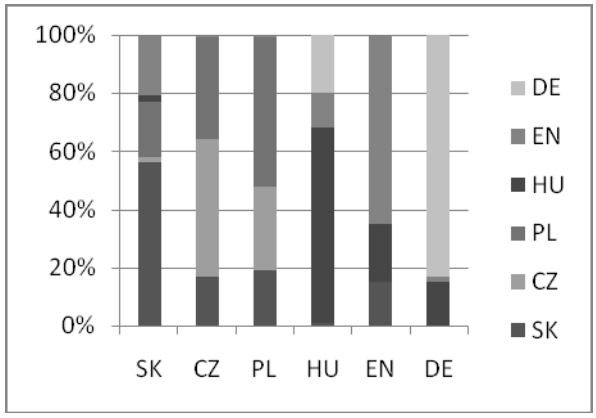
Najskôr bolo treba celý text niektorého z jazykov rozdeliť na rovnako dlhé intervaly. Pre každý interval určitej dĺžky sa vypočíta index koincidencie. Podobne som zobral inak dlhý interval a opäť pre každý vypočítal podľa vzťahu (1) index koincidencie. Dĺžky intervalov som zvolil 100, 200, 300, 500 a 1000. Tieto intervaly dĺžok neboli zvolené náhodne. 100 znakový text môžeme považovať za krátku vetu a 1000 znakový text už za rozvinuté súvetie, poprípade za asi 4 vety. V mnohých prípadoch má správa len niekoľko viet, takže by sa mohla vojsť do niektorého z intervalov. Aby boli výsledky čo najobjektívnejšie, intervaly textov sa vyberali v pravidelných intervaloch po celej dĺžke celkového textu. Takto som postup zopakoval pre každý z jazykov a získal indexy koincidencie pre intervaly textov, ktoré som potom pomocou tabuľkového procesora porovnával s indexom koincidencie každého z jazykov. Za predpokladaný jazyk som označil ten, od ktorého rozdiel bol minimálny.

3.1.3 Grafické zobrazenie

Zobrazuje pravdepodobnosť správneho určenia jazyka pomocou indexu koincidencie.



Obr. 1 Pravdepodobnosť rozlíšenia 100 znakového textu pomocou indexu koincidencie.



Obr. 2 Pravdepodobnosť rozlíšenia 1000 znakového textu pomocou indexu koincidencie.

3.2 Riešenie pomocou vzorov

V tejto práci sa hľadaním vzorov v texte myslí nájsť takú päticu, ktorá má podľa predpisu daný počet opakujúcich sa písmen, pričom nezáleží na poradí písmen v danej pätici. Takto dostaneme sedem možností zostrojenia vzoru.

3.2.1 Celkový index koincidencie vzorov

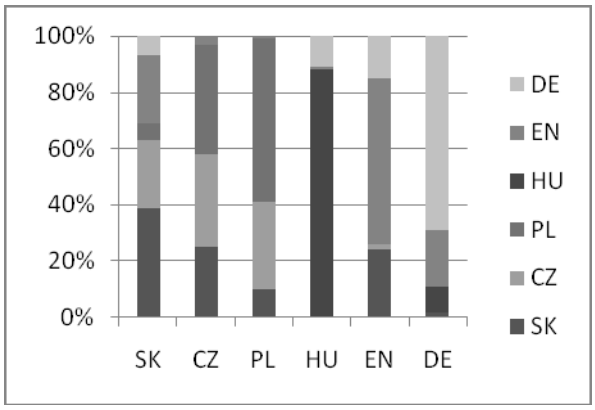
Najskôr som našiel počet všetkých vzorov pre každý jazyk z celého textu. Z jednotlivých počtov vzorov sa potom vypočítal index koincidencie vzorov celkového textu postupne pre všetky jazyky.

3.2.2 Index koincidencie vzorov

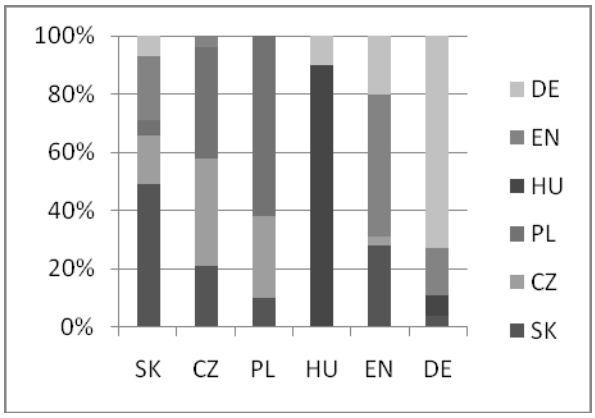
Určenie indexu koincidencie pre vzory spočívalo v spočítaní vzorov intervalu textu niektorého z jazykov. Podobne ako v bode 3.1.2 som zobral určitý interval a vypočítal index koincidencie vzorov. Takto som dostal pre daný jazyk index koincidencie. Postup som znova zopakoval s inou dĺžkou intervalu niekoľkokrát, aby som dostal dostatočný počet vzoriek. Konkrétne som opäť robil 100 vzoriek pre každú dĺžku intervalov 100, 200, 300, 500 a 1000 znakov. Postupne pre každý jazyk som získané hodnoty v tabuľkovom procesore s pomocou funkcie porovnal s celkovým indexom koincidencie každého jazyka. Najmenší rozdiel medzi porovnaným a celkovým indexom koincidencie niektorého z jazykov mi označil daný jazyk ako jazyk správy. Keď som potom porovnal určený jazyk správy s jazykom vybraného textu, overil som správnosť určenia jazyka.

3.2.3 Grafické zobrazenie

Zobrazuje pravdepodobnosť správneho určenia jazyka pomocou hľadania vzorov v texte.



Obr. 3 Pravdepodobnosť rozlíšenia 100 znakového textu pomocou vzorov.



Obr. 4 Pravdepodobnosť rozlíšenia 1000 znakového textu pomocou vzorov.

4. Záver

Správne určenie jazyka pomocou použitých metód nie je v každom prípade rovnaké a správne. Je to zapríčinené charakteristickými črtami jednotlivých jazykov, či dĺžkou analyzovaného textu. Napr. porovnaním obrázkov Obr.1 a Obr. 2, alebo Obr. 3 a Obr. 4 môžeme vidieť ako sa pri dlhšom texte zlepšila pravdepodobnosť správneho určenia jazyka pri rozlišovaní pomocou indexu koincidencie. Je to spôsobené tým, že väčší počet znakov nám umožňuje lepšie priblíženie k charakteristickým črtám jazyka a teda môžeme povedať, že správne určenie jazyka je závislé na dĺžke analyzovaného textu. Skúšal som pri akej dĺžke textu bude toto určenie dostatočne presné. Začal som pri dĺžke textu 100 znakov, 100 znakový text bol ešte relatívne nepresný a určenia jazykov mali podobné pravdepodobnosti. Jazyky sa takmer nedali rozlíšiť. S rastúcim počtom znakov analyzovaného textu sa štatistiky vylepšovali a pri dĺžke 1000 znakov bolo rozlišovanie veľmi dobré. Mohli by sme povedať, že stačia 2 zložené súvetia na určenie jazyka. Rozlíšenie Českého a Poľského jazyka však pri určovaní indexom koincidencie nie je správny ani pri texte s dĺžkou 1000 znakov. Bolo treba vyskúšať iný spôsob, ktorý by lepšie rozlíšil tieto dva jazyky. Pomocou hľadania výrazov sa tieto dva jazyky rozlíšili lepšie, no výrazne sa ani tu neprejavila žiadna vlastnosť jazykov, ktorá by ich dostatočne dobre rozlíšila. To je ale dôkazom, že tieto jazyky sú si blízke, pretože porovnaním slovanských jazykov by boli výsledky bez využitia inej vhodnej metódy rovnako nedostačujúce. Ak si všimneme grafy lepšie, uvidíme, že rozlišovanie slovanských jazykov je pri použití oboch metód podobné a na odlíšenie týchto jazykov by bolo naozaj treba zvoliť nejakú inú metódu. Hľadanie vzorov nám v tomto prípade zlepšilo rozlíšenie nielen poľského jazyka, ale aj maďarského a s kombináciou určovania pomocou indexu koincidencie je pravdepodobnosť určenia správneho jazyka vyššia.

Keďže analyzovaný text bol vlastne vždy iný interval súvislého textu z celkového textu, bolo dôležité ako sa dané intervaly vyberali na analyzovanie. Mohlo by sa totiž stať, že pri náhodnom výbere intervalu textu by sa vybrali intervaly, ktoré by sa prekrývali, poprípade by boli totožné. Preto som využitím vlastného programu vyberal intervaly tak, aby sa neprekrývali a boli od seba čo najrovnomernejšie vzdialené.

Každá pomôcka pri lúštení môže byť významná a index koincidencie, či vzory sú veľmi jednoduchou metódou analyzovania textu, čo som si vďaka tejto práci overil prakticky. Verím, že takáto pomôcka by mohla byť reálne nápomocná a urýchlila by lúštenie zašifrovaného textu.

5. Použitá literatúra

[1] O. Grošek, M. Vojvoda, P. Zajac: Klasické šifry. STU Bratislava, 2007.
[2] Simon Singh :“The Code Book”, Fourth Estate Ltd., London, 1999
[3] B.R.Heap: Permutations by Interchanges. Computer J. 6(1963)