Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Metódy útoku hrubou silou na TrueCrypt

Diplomová Práca

2016 **Bc. Martin Strapko**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Metódy útoku hrubou silou na TrueCrypt

Diplomová Práca

Študijný program: Informatika
Študijný odbor: 2508 Informatika
Školiace pracovisko: Katedra Informatiky
Vedúci práce: RNDr. Richard Ostertág PhD.

2016 **Bc. Martin Strapko**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Martin Strapko
Študijný program: informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Metódy útoku hrubou silou na TrueCrypt
Methods of brute-force attack on TrueCrypt

Cieľ: Cieľom práce je napísať program, ktorý sa snaží hrubou silou, ale čo najefektívnejšie, nájsť heslo pre dešifrovanie disku zašifrovaného pomocou programu TrueCrypt.

Riešenie projektu je možné rozdeliť do viacerých podproblémov:
* Preštudovanie dokumentácie a implementácie šifrovania v programe TrueCrypt.
* Izolovanie minimálneho kódu potrebného pre overenie korektnosti hesla.
* Navrhnúť popis na generovanie hesiel spĺňajúcich určité pravidlá. Napríklad:
* Minimálna, maximálna dĺžka.
* Počet znakov, cifier, a nealfanumerických znakov.
* Tvar hesla, napríklad pomocou regulárnych výrazov (napr. /(\d){2,3}\[a-z\]+(\[A-Z\]){4,7}/).
* Heslo sa má nachádzať v definovanom slovníku.
* Vytvoriť takýto slovník pre SK.
* Generovanie potenciálnych hesiel na základe navrhnutého popisu.
* Zohľadniť pravdepodobnosť výskytu slov alebo písmen (napr. podľa použitého jazyka).
* Overovanie hesiel (paralelne na viacerých jadrách CPU, GPU, distribuované na viacerých PC)
* Zobrazovanie progresu a odhadovaného času.
* Hľadanie hesla pomocou SAT-solverov na GPU.

Vedúci: RNDr. Richard Ostertág, PhD.
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: doc. RNDr. Daniel Olejár, PhD.

Spôsob sprístupnenia elektronickej verzie práce:
bez obmedzenia

Dátum zadania: 22.10.2014

Dátum schválenia: 12.12.2014 prof. RNDr. Branislav Rovan, PhD.
garant študijného programu

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

....

študent vedúci práce

Poďakovanie

Touto cestou by som sa chcel poďakovať svojmu vedúcemu RNDr. Richardovi Ostertágovi PhD., za námety a prospešné rady pri tvorbe tejto práce. Taktiež by som sa chcel poďakovať priateľom a rodine za podporu a motiváciu počas písania tejto práce.

Abstrakt

Táto práca sa vo všeobecnosti zaoberá optimalizáciou útokov hrubou silou. Špeciálne rozoberá možnosti útoku na program TrueCrypt. Sústredí sa na to ako využiť znalosti o často používaných heslách na vylepšenie efektívnosti útoku. Hlavným cieľom práce je napísať program, ktorý sa snaží pomocou útoku hrubou silou nájsť čo najefektívnejšie heslo pre dešifrovanie disku zašifrovaného pomocou programu TrueCrypt.

KĽÚČOVÉ SLOVÁ: TrueCrypt, útok hrubou silou, bezkontextové gramatiky, Markovov zdroj, používateľské heslá

Abstract

Focus of this work is on optimizing brute-force attacks. It studies diffrent methods to attack TrueCrypt. The main purpose of this thesis is to implement application that is able to effectively recover password for disk encrypted with TrueCrypt.

KEYWORDS: TrueCrypt, brute-force attack, context-free grammar

Obsah

Zoznam listingov

Zoznam obrázkov

Úvod

V dnešnom svete fungujúcom na elektronických dátach, ktoré pre nás majú obrovskú cenu, sa ľudia snažia ochrániť ich dôvernosť. Tieto dáta sa dajú jednoducho ochrániť, ak k ním bude mať fyzický prístup len majiteľ. Toto avšak nie je najpoužiteľnejšie riešenie, keďže takmer každý počítač je pripojený do nejakej siete. Iná možnosť je mať dáta uložené vo forme, ktorá bude dávať zmysel len oprávneným osobám, aj keď fyzický prístup k ním môžu mať aj iný ľudia. Na tento účel primárne slúži šifrovanie.

Bezpečnosť šifrovania veľmi záleží na utajení kľúča. Preto je dôležité, aby nebol ľahko uhádnuteľný. Keďže počítače priniesli so sebou obrovskú výpočtovú silu, sú schopné robiť až niekoľko desiatok tisíc pokusov za sekundu snažiac sa uhádnuť tento kľúč [13]. Keďže rýchlosť tohto hľadania kľúča záleží hlavne od veľkosti prehľadávaného priestoru kľúčov, v praxi sa bežne používajú aspoň 256 bitov dlhé kľúče. Toto je ekvivalent 32 znakového používateľského hesla zloženého z ľubovolných znakov ASCII tabuľky.

Takéto hľadanie kľúča preberaním všetkých možností sa nazýva útok hrubou silou. Jeho podstatou je postupné generovanie možných kľúčov a následne overenie ich správnosti. V našej práci sa zameriame na útok pri ktorom má útočník fyzický prístup k súboru s odtlačkami. Postup tohto útoku je nasledovný:

- Útočník vygeneruje kandidáta na overenie
- Ak existuje, tak pripojí náhodný reťazec spojený s týmto heslom ku kandidátovi. Tieto reťazce zabezpečujú ochranu pred útokom pomocou predpočítaných odtlačkov
- Zahešuje kandidáta pomocou zvoleného hešovacieho algoritmu
- Porovná novovzniknutý odtlačok s odtlačkami nachádzajúcimi sa v súbore z ktorého sa snaží nájsť heslá

Nakoľko je tento algoritmus na overovanie kandidátov dobre paralelizovateľný, keďže overenie jedného kandidáta nezávisí na žiadnom inom overení, jednou z optimalizácií tohto procesu bude počítanie týchto odtlačkov pomocou grafických kariet. Nakoľko existujú práce [11] podrobne sa venujúce sa týmto optimalizáciam, nebudeme vrámci tejto práce implementovať program na overenie správnosti kandidátov.

V tejto práci sa budeme zaoberať skúmaním a implementáciou algoritmov na generovanie týchto kandidátov. Výstupom našej práce bude program, ktorý bude generovať zoznam hesiel, ktoré sa dajú následne použiť ako kandidáti v niektorom z voľne dostupných programov na skúšanie takýchto hesiel ako napríklad *hashCat*, *John The Ripper* alebo *PasswordsPro*.

Tieto programy podporujú viaceré spôsoby útokov hrubou silou na veľké množstvo známych a často používaných hešovacích funkcií. Tieto typy útokov zahŕňajú použitie predom vytvorených slovníkov, spájanie slov z rôznych slovníkov ako aj postupné generovanie všetkých možných reťazcov. V našej práci sa snažíme vyvinúť a implementovať ďalší spôsob generovania kandidátov použitím pravdepodobnostných bezkontextových gramatík.

Efektívnosť nami navrhnutého riešenia budeme porovnávať s algoritmom používajúcim Markovovské zdroje, ktoré sú používané aj vyššie spomenutými programami pri postupnom generovaní kandidátov. Tento algoritmus sme si vybrali, keďže jeho základný princíp priraďovania pravdepodobnosti jednotlivým znakom na základe predošlého stavu je najbližší k tomu, ktorý používame v našom algoritme využívajúcom bezkontextové gramatiky.

KAPITOLA 1

TrueCrypt

TrueCrypt je šifrovací program umožňujúci používateľovi na základe ním zvoleného hesla vytvoriť šifrovaný disk. Taktiež používateľovi umožňuje vytvorenie virtuálneho šifrovaného disku, ktorý bude následne uložený v súbore na fyzickom disku. Vývoj tohto programu bol ukončený v roku 2014 a podľa autorov nie je bezpečný, nakoľko jeho implementácia môže obsahovať bezpečnostné chyby. Následný bezpečnostný audit tohto programu neukázal žiadne závažné bezpečnostné chyby v jeho základnom návrhu.

V septembri 2015 prišiel James Forshaw s informáciou o 2 chybách vo ovládači Windows-u, ktorý používa program TrueCrypt. Jedna z chýb umožňuje útočníkovi plný prístup k zašifrovaným partíciam iných používateľov, ktoré sú na tom istom počítači [5]. Druhá, závažnejšia chyba umožňuje útočníkovi prístup k zvýšeným právam zneužitím tvorby symbolického odkazu na písmená diskov [4]. Obe tieto chyby sa dokážu prejaviť až počas behu samotného programu. Preto ich v tejto práci používať nebudeme, nakoľko pre náš algoritmus nie je potrebné aby TrueCrypt bežal.

Ako sme spomínali v úvode, program TrueCrypt slúži na zašifrovanie dát na používateľskom disku pomocou zvoleného hesla. K tomuto TrueCrypt používa niektorý zo šifrovacích algoritmov medzi ktoré patrí AES, Serpent alebo Twofish. Keďže sa jedná o blokové šifry, TrueCrypt používa blokové šifry v XTS móde na šifrovanie objemu dát väčšieho ako je 1 blok šifry. TrueCrypt taktiež poskytuje možnosť vybrať si jednu z podporovaných hešovacích funkcií ako RIPEMD-160, SHA-512 a Whirlpool.

Samotné šifrovanie partície prebieha vo viacerých fázach:

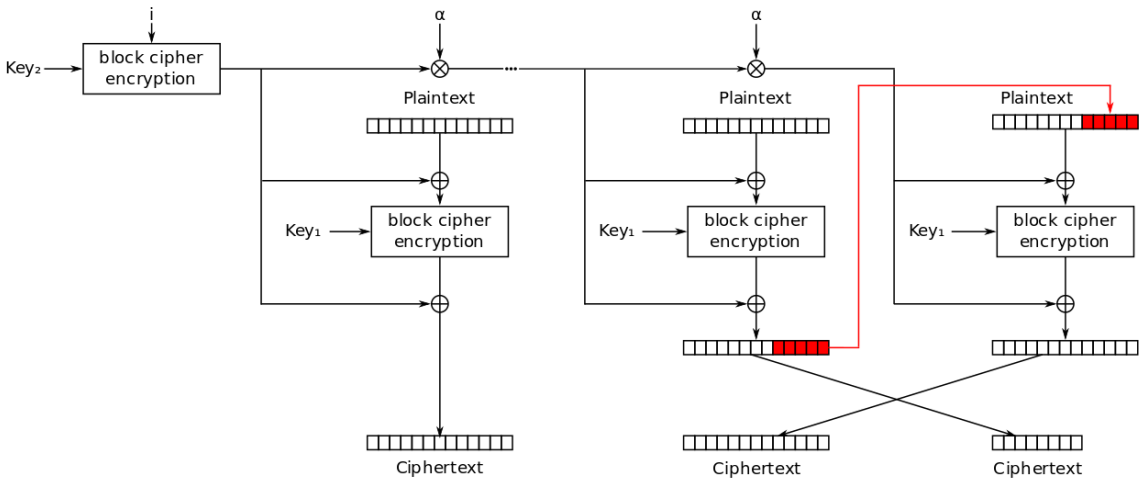
- Vygeneruje sa náhodný kľúč vhodný na šifrovanie pomocou zvoleného algoritmu.
  - V prípade XTS módu veľkosť kľúča zdvojnásobíme. Dôvodom je použitie 2 kľúčov pri šifrovaní v XTS móde ako je znázornené na obrázku 1.1

- Pomocou tohto kľúča sa zašifruje celá požadovaná partícia
- Vytvorí sa hlavička pre túto partíciu obsahujúca:
  - Verziu programu TrueCrypt
  - Verziu hlavičky
  - ASCII reťazec 'TRUE'
  - Vygenerovaný kľúč použitý na zašifrovanie dát
  - CRC-32 kontrolná suma kľúča
  - CRC-32 kontrolná suma zvyšku hlavičky
- Táto hlavička sa zašifruje pomocou kľúča vygenerovaného na základe používateľského hesla, ktoré nemusí byť vhodné na použitie ako šifrovací kľúč
  - K heslu sa pripojí náhodný 512 bitový reťazec - kryptografická soľ
  - Takto upravené heslo sa dá na vstup algoritmu PBKDF2
  - Transformáciami pomocou hešovacích funkcií vznikne vhodný kľúč na šifrovanie
- Vyššie vygenerovaná kryptografická soľ sa nezašifrovaná pridá pred hlavičku partície

Takto zašifrovaná partícia je nakoniec uložená na fyzickom disku v súboru obsahujúcom:

- Kryptografickú soľ v nešifrovanej forme
- Šifrovanú hlavičku partície obsahujúcej kľúč použitý na šifrovanie dát
- Používateľom zvolené dáta zašifrované programom TrueCrypt

XTS mód Na obrázku 1.1 vidíme schému módu XTS pre blokové šifry. Základ tvorí mód XEX, ktorý bol navrhnutý na efektívne spracovanie za sebou nasledujúcich blokov vrámci dátového bloku napríklad diskového sektoru. Na schéme môžeme vidieť použitie dvoch rôznych kľúčov Key_1 a Key_2. Kvôli použitiu týchto kľúčov sa častokrát generuje šifrovací kľúč dvojnásobnej dĺžky, kedy sa následne prvá polovica použije ako Key_1 a druhá ako Key_2. Hodnota i vyjadruje číslo diskového sektoru, ktorý sa práve šifruje, zatiaľ čo α je prvok z konečného poľa $GF(2^128)$. Pre každý blok šifry sa spraví α^j, kde j je číslo bloku vrámci sektoru. Mód XTS prináša módu XEX podporu pre sektory veľkosti nedeliteľnej veľkosťou bloku šifry. Ak posledný blok otvoreného textu nemá dostatočnú veľkosť, pridá sa k nemu potrebný počet bajtov predošlého zašifrovaného bloku. Po zašifrovaní takto upraveného otvoreného textu sa novo zašifrovaný blok vymení so zvyškom predposledného šifrovaného bloku, ako je znázornené na schéme.



Obr. 1.1: Schéma módu XTS

Vďaka zdrojovým kódom voľne prístupným na internete, sme mali možnosť si ich prehliadnuť a hľadať v nich možnosť rýchleho overovania kandidátov na hľadané používateľské heslo. Našim cieľom pri tomto hľadaní bolo nájdenie časti kódu, ktorá je zodpovedná za prijatie používateľského vstupu a jeho následné spracovanie. Toto spracovanie zahŕňa transformáciu tohto vstupu na kľúč, ktorým sa program pokúsi rozšifrovať hlavičku partície 1.1. Následne sa program posnaží rozšifrovať hlavičku pomocou tohto kľúču a overí, či dáta v nej dávajú zmysel. Toto zahŕňa overenie reťazca *'TRUE'*, verzie hlavičky, minimálnej verzie programu a CRC-32 súm kľúčov a ostatných položiek 1.2. Bohužiaľ vrámci tejto práce sme nemali čas izolovať minimálny kód potrebný na úspešne fungovanie tohto overenia.

Zbytok tejto práce sa venuje generovaniu slovníka z ktorého budeme brať kandidátov na hľadané heslo.

```
switch (pkcs5_prf)
{
  case RIPEMD160:
    derive_key_ripemd160 (keyInfo.userKey, keyInfo.keyLength,
            keyInfo.salt, PKCS5_SALT_SIZE,
            keyInfo.noIterations, dk, GetMaxPkcs5OutSize());
    break;

  case SHA512:
    derive_key_sha512 (keyInfo.userKey, keyInfo.keyLength,
            keyInfo.salt, PKCS5_SALT_SIZE,
            keyInfo.noIterations, dk, GetMaxPkcs5OutSize());
    break;

  case SHA1:
    // Deprecated/legacy
    derive_key_sha1 (keyInfo.userKey, keyInfo.keyLength,
            keyInfo.salt, PKCS5_SALT_SIZE,
            keyInfo.noIterations, dk, GetMaxPkcs5OutSize());
    break;

  case WHIRLPOOL:
    derive_key_whirlpool (keyInfo.userKey, keyInfo.keyLength,
            keyInfo.salt, PKCS5_SALT_SIZE,
            keyInfo.noIterations, dk, GetMaxPkcs5OutSize());
    break;

  default:
    // Unknown/wrong ID
    TC_THROW_FATAL_EXCEPTION;
}
```

Zdrojový kód 1.1: Ukážka kódu transformácie hesla na šifrovací kľúč

```
// Copy the header for decryption
memcpy (header, encryptedHeader, sizeof (header));

// Try to decrypt header
DecryptBuffer (header + HEADER_ENCRYPTED_DATA_OFFSET,
      HEADER_ENCRYPTED_DATA_SIZE, cryptoInfo);

// Magic 'TRUE'
if (GetHeaderField32 (header, TC_HEADER_OFFSET_MAGIC) != 0x54525545)
  continue;

// Header version
headerVersion = GetHeaderField16 (header, TC_HEADER_OFFSET_VERSION);

if (headerVersion > VOLUME_HEADER_VERSION)
{
  status = ERR_NEW_VERSION_REQUIRED;
  goto err;
}

// Check CRC of the header fields
if (!ReadVolumeHeaderRecoveryMode
  && headerVersion >= 4
  && GetHeaderField32 (header, TC_HEADER_OFFSET_HEADER_CRC) !=
    GetCrc32 (header + TC_HEADER_OFFSET_MAGIC,
    TC_HEADER_OFFSET_HEADER_CRC - TC_HEADER_OFFSET_MAGIC))
  continue;

// Required program version
cryptoInfo->RequiredProgramVersion = GetHeaderField16 (header,
              TC_HEADER_OFFSET_REQUIRED_VERSION);
cryptoInfo->LegacyVolume = cryptoInfo->RequiredProgramVersion < 0x600;

// Check CRC of the key set
if (!ReadVolumeHeaderRecoveryMode
  && GetHeaderField32 (header, TC_HEADER_OFFSET_KEY_AREA_CRC) !=
  GetCrc32 (header + HEADER_MASTER_KEYDATA_OFFSET,
      MASTER_KEYDATA_SIZE))
  continue;
```

Zdrojový kód 1.2: Ukážka kódu overenia správnosti hesla dešifrovaním hlavičky

KAPITOLA 2

Útoky hrubou silou

Základným princípom útokov hrubou silou je hľadanie správneho riešenia pomocou skúšania veľkého množstva kandidátov. Spôsob skúšania kandidátov sa môže líšiť od situácie, avšak veľmi často máme prístupnú zašifrovanú, respektíve zahešovanú verziu hľadaného reťazca. Keďže hešovacie algoritmy sú dizajnované tak, aby nebolo možné z odtlačku vyrobiť pôvodný reťazec, musíme pre vyskúšanie kandidáta zahešovať tohto kandidáta a následne porovnať výsledný odtlačok s tým od správneho hesla. Metód akými sa dajú títo kandidáti generovať je mnoho a nižšie si predstavíme niektoré z nich.

Všetky v praxi používané algoritmy používajú kľúče dĺžky aspoň 256 bitov, čo je ekvivalent 32 znakového reťazca zloženého z ľubovolných znakov ASCII tabuľky. V praxi je veľmi nepravdepodobné, že používateľ bude mať takto dlhé heslo založené nad tak veľkou abecedou. Práve preto sa v tejto práci sústredíme na používateľské heslá, pretože majú omnoho menší počet možných reťazcov. Možností pre 256 bitový kľúč je 2^{256} zatiaľ čo možností pre 16 miestne heslo zložené z veľkých, malých písmen, číslic a niektorých často používaných znakov je približne 2^{101}, čo už je dosť signifikantné zmenšenie počtu možností (na $2.18953 * 10^{-45}\%$ pôvodnej veľkosti).

2.1 Útok úplným prehľadávaním

Tento spôsob hľadania hesla patrí medzi najzákladnejšie útoky hrubou silou a častokrát sa práve on myslí pod pojmom útok hrubou silou. Podstatou tohto útoku je vyskúšanie všetkých kandidátov. Ak prejdeme cez celý priestor reťazcov, museli sme určite prejsť aj cez konkrétny reťazec, ktorý hľadáme. Táto metóda má tým pádom 100% úspešnosť. Jej problém avšak spočíva v množstve reťazcov, ktoré je potrebné vyskúšať. Vo väčšine prípadov vieme obmedziť hľadanie maximálnou dĺžkou hľadaného výrazu a abecedou znakov z ktorej sa daný výraz skladá. Používateľské heslá mávajú maximálnu dĺžku okolo 16 znakov a sú zložené prevažne z asi 80 rôznych znakov. Pre takéto re-

ťazce, ktorých je 80^{16}, by nám vyskúšanie všetkých trvalo približne $8.92 * 10^{13}$ rokov pri skúšaní 1 miliardy reťazcov za sekundu.

2.2 Slovníkový útok

Častokrát existuje ešte menšia množina reťazcov, heslá z ktorej majú omnoho väčšiu šancu, že medzi nimi bude hľadaný výraz. Toto je pravda špeciálne pri hľadaní používateľských hesiel, nakoľko používatelia volia heslá tak, aby boli zapamätateľné. Vďaka tomu existuje relatívne malá množina reťazcov, ktoré keď vyskúšame, máme vysokú šancu úspechu. V takomto prípade je najlepšie zostaviť slovník takýchto reťazcov, ktoré potom postupne skúšame. Táto metóda väčšinou nájde heslo rýchlejšie ako vyššie spomínané úplne prehľadávanie. Avšak jej úspešnosť záleží hlavne od tohto vstupného slovníka. V dnešnom svete, keď takmer každá služba vyžaduje heslo od používateľa, existuje veľa verejne prístupných zoznamov najčastejšie používaných hesiel, ktoré slúžia ako veľmi dobrý základ pre tento útok.

2.2.1 Prekrúcanie slov

Samotný slovníkový útok väčšinou pokrýva takmer zanedbateľné percento všetkých možných výrazov spadajúcich do priestoru hesiel danej abecedy a dĺžky. Preto sa spolu s touto metódou často používa prekrúcanie slov. Podstatou je rozšírenie vstupného slovníka o alternatívne verzie vstupných hesiel za účelom rozšírenia prehľadaného priestoru reťazcov. Bežne sa to dosahuje definovaním zoznamu pravidiel popisujúcich transformáciu slova. Tieto pravidlá budú následne aplikované na jednotlivé vstupné slová a tým vzniknú potenciálne nové reťazce, ktoré sa nenachádzajú vo vstupnom slovníku. Tieto pravidlá môžu transformovať slovo rôznymi spôsobmi od pridania prefixu či sufixu cez zmenu veľkostí písmen alebo vynechanie spoluhlások. Mnohé programy zaoberajúce sa útokmi hrubou silou podporujú vlastný jednoduchý jazyk na popísanie týchto pravidiel.

2.3 Hybridný útok

V tejto práci sa venujeme implementácii útoku, ktorý je spojením vyššie uvedených. Našim hlavným cieľom je nájdenie správneho hesla k partícii zašifrovanej programom TrueCrypt. Keďže chceme toto heslo nájsť v ľubovoľne veľkom konečnom čase, budeme náš algoritmus implementovať tak, aby vygeneroval všetky možné reťazce zo vstupnej abecedy kratšie ako nami stanovená dĺžka, podobne ako pri úplnom prehľadávaní. Avšak v našom prípade predpokladáme, že na vstupe dostaneme ešte slovník obsahujúci zoznam hesiel. Predpokladáme, že tento zoznam je usporiadaný podľa pravdepodobnosti správnosti hesiel v ňom. Náš algoritmus si na základe týchto hesiel upraví pravde-

podobnosti vygenerovania jednotlivých reťazcov aby následne mohol na výstup dávať heslá od najpravdepodobnejšieho po najmenej pravdepodobné.

2.4 SAT Solver

Táto metóda útoku hrubou silou je založená na probléme splniteľnosti boolovského výrazu. Ako vstup tohto algoritmu je boolovský výraz, väčšinou v konjuktívnom normálnom tvare, pre ktorý sa daný algoritmus snaží nájsť také ohodnotenie boolovských premenných, aby všetky formuly tohto výrazu boli pravdivé. Algoritmus postupne rekurzívne prehľadáva všetky možnosti nastavenia jednotlivých premenných. Po nastavení niektorej premennej skontroluje, či žiaden výskyt tejto premennej nespôsobil konflikt, čiže ohodnotil formulu tak, že sa stala nesplniteľnou. V tomto prípade sa algoritmus vráti do momentu kedy bol výraz nekonfliktný a odtiaľ sa snaží postupovať inou cestou.

Náš problém sa dá pretransformovať na vstup pre takýto SAT solver. V našom prípade by sme vedeli šifrovací algoritmus prepísať do konjuktívnej normálnej formy. Ako výstup tohto algoritmu je reťazec takmer náhodných bitov, ktoré dopredu poznáme, pretože tieto sú fyzický uložené na disku. Neznámymi v tomto boolovskom výraze sú vstupné dáta a kľúč, pomocou ktorého boli tieto dáta zašifrované. Našim predpokladom je, že veľkú časť týchto dát by sme vedeli určiť, keďže ide o dopredu známe informácie ako reťazec TRUE, použitú verziu programu TrueCrypt a podobne. Vďaka týmto informáciám by sa vedel SAT solver skôr rozhodnúť, že niektoré ohodnotenie premenných nie je možné, kvôli konfliktu, ktorý by vznikol.

V našej práci sme túto metódu hlbšie neanalyzovali, keďže existuje viacero prác, ktoré do podrobna rozoberajú správanie týchto solverov v prípade, že majú dopredu určené niektoré bity vstupu alebo výstupu. Taktiež existujú práce, ktoré sa venujú optimalizácii behu SAT solveru, čo zahŕňa optimalizáciu poradia ohodnotenia premenných alebo miesta, na ktoré sa algoritmus vráti v prípade konfliktu.

KAPITOLA 3

Používateľské heslá

Aj napriek tomu, ako sú používateľské heslá vo všeobecnosti ľahko prelomiteľné, sú dnes najčastejšie používaný spôsob autentifikácie používateľa. Tento trend sa pravdepodobne ani v najbližšej budúcnosti nebude meniť. Hlavným dôvodom slabých používateľských hesiel sú obmedzenia ľudskej pamäte. Ak by si používatelia nemuseli pamätať heslo, tak by používali heslá s najväčšou entropiou. Tie by boli najdlhšie možné povolené systémom, zložené z náhodne vybraných znakov povolených týmto systémom a neexistovala by žiadna iná možnosť ako túto sekvenciu dostať na základe inej informácie.

Tento spôsob tvorby hesiel je presný opak toho k čomu je prispôsobená ľudská myseľ. Ľudia sú schopní zapamätať si sekvenciu znakov dlhú približne sedem znakov plus mínus dva znaky vo svojej krátkodobej pamäti. Taktiež, aby si človek zapamätal takúto sekvenciu, táto sekvencia nemôže byť kompletne náhodná, ale musí sa skladať zo známych kusov informácie ako sú napríklad slová. Nakoniec ľudská myseľ funguje veľmi dobre vďaka redundancii informácie, čiže človeku sa ľahšie pamätajú veci, ktoré si vie odvodiť z viacerých iných kusov informácii.

Mnohé systémy používajúce heslá ako spôsob autentifikácie používateľa dávajú používateľovi rady ako si zvoliť bezpečné heslo. Na základe informácií spomenutých na začiatku tejto kapitoly by heslo malo byť dostatočne dlhé, skladajúce sa z rozumne veľkej abecedy znakov a malo by byť ľahko zapamätateľné. Väčšina týchto systémov sa sústredí hlavne na prvé dve podmienky tvorby hesla a to že by malo používateľove heslo spĺňať minimálnu dĺžku a obsahovať aspoň jeden z každej kategórie veľké, malé písmena, číslice a špeciálne znaky. Týmto dávajú dôraz na ochranu proti útokom hrubou silou oproti zapamätateľnosti hesla.

Mnoho používateľov, ktorí boli prezentovaní minimálnymi nárokmi na heslo, si vyvinulo spôsob na generovanie takýchto hesiel. Tento spôsob zahŕňal výber slova, ktorému

zväčšili prvé písmeno a pridali sufix skladajúci sa z číslic a špeciálnych znakov. Takéto spôsoby zakladajúce na jednoduchej transformácii slova sa veľmi rýchlo ukázali neefektívne, keď sa počas posledného desaťročia náramne zvýšil výkon počítačov. Tie boli schopné skúsiť veľké množstvo transformácií pre každé slovo slovníka.

Začali sa objavovať mnohé mnemotechnické pomôcky umožňujúce generovať používateľské heslá. Jedna z často sa vyskytujúcich doporučovala vytvorenie si extrémne dlhého slovného spojenia. Účelom tejto metódy bola ochrana proti útokom hrubou silou zväčšením priestoru potenciálnych hesiel pomocou zvýšenia dĺžky samotného hesla. Ďalšia veľmi často používaná metóda bola založená na tvorbe hesla zobratím prvých znakov slov z fráze, ktorú používateľ vymyslel. Heslá založené na mnemotechnických pomôckach sa málokedy vyskytujú v slovníkoch používaných pri útokoch hrubou silou. To však neznamená, že sú bezpečnejšie ako bežné heslá [8].

V tejto práci sa snažíme vyvinúť algoritmus, ktorý dostane na vstupe slovník s heslami. Tento slovník ma slúžiť na naučenie algoritmu metódy tvorby a používania hesiel pre konkrétneho používateľa. Mal by zahŕňať ukážky hesiel, ktoré sú vytvorené podobnými metódami ako používateľ vytvára heslá pre potreby svojej autentifikácie.

KAPITOLA 4

Učenie

Ako sme spomínali, budeme sa zaoberať útokom, ktorý dostane na vstupe slovník a na základe tohto slovníka bude generovať heslá zoradené podľa pravdepodobnosti. Kvalita výsledného zoznamu bude závisieť od schopnosti algoritmu správne sa naučiť ohodnotiť pravdepodobnosti jednotlivých reťazcov. V tejto práci kladieme dôraz na skúmanie možností generovania reťazcov použitím bezkontextových gramatík, avšak implementovali sme taktiež algoritmus používajúci Markovovské zdroje, ktorý použijeme na porovnanie s bezkontextovými gramatikami.

4.1 Pravdepodobnostné bezkontextové gramatiky

Bezkontextové gramatiky sú definované štyrmi parametrami. Množinou neterminálov, ktoré slúžia ako premenné pri odvodzovaní vetnej formy. Množinou terminálov, ktoré tvoria reálny obsah výslednej vetnej formy. Túto množinu tvorí vstupná abeceda symbolov a je disjunktná s neterminálmi. Vetná forma obsahujúca len terminálne symboly sa nazýva terminálna vetná forma. Ďalej je potrebné zadefinovať počiatočný neterminál z ktorého sa bude každá vetná forma odvádzať. Nakoniec potrebujeme poznať množinu prepisovacích pravidiel, ktoré definujú spôsob akým sa menia neterminály na ďalšie vetné formy. Pri bezkontextových gramatikách majú prepisovacie pravidlá tvar

$$N -> (N \cup \Sigma)^*$$

kde N vyjadruje množinu neterminálov a Σ je množina terminálov. Tieto pravidlá vyjadrujú schopnosť neterminálu zmeniť sa na ľubovolnú vetnú formu, bez ohľadu na kontext v ktorom sa nachádza. V našej práci sa budeme venovať špeciálnym bezkontextovým gramatikám, ktorých každé prepisovacie pravidlo má priradenú pravdepodobnosť. Suma pravdepodobností jedného neterminálu bude vždy rovná 1. Vďaka týmto pravdepodobnostiam dokážeme ohodnotiť nami generované heslá a zoradiť ich podľa ich

pravdepodobností. Pravdepodobnosť ľubovolnej vetnej formy získame súčinom pravdepodobností prepisovacích pravidiel použitých na jej odvodenie.

Odvodenia vetných foriem, čiže sekvencie použitých prepisovacích pravidiel, tvoria strom odvodenia daného slova. Je možné aby v takomto strome existovali 2 rôzne cesty odvodenia, ktoré nakoniec vygenerujú rovnakú vetnú formu. Tejto vlastnosti bezkontextových gramatík sa budeme snažiť vyhnúť vytvorením pravidiel tak, aby ľubovolná terminálna vetná forma mala práve jeden spôsob odvodenia v danej gramatike. Zámerom tohto obmedzenia je zamedzenie generovania duplikátov, keďže predpokladáme, že pri skúšaní jedného hesla viac krát sa výsledok tohto pokusu nezmení.

4.2 Markovovské zdroje

Bezkontextové gramatiky, ktoré sme implementovali si odvádzajú vetné formy výberom prepisovacieho pravidla s najvyššou pravdepodobnosťou. Taktiež si pamätajú, ktoré pravidlá už použili, aby sa vyhli odvodeniu jednej terminálnej vetnej formy viackrát. Keďže gramatika si počas generovania reťazca pamätala celú postupnosť použitých prepisovacích pravidiel, vyžadovala veľmi veľa pamäte. Preto sme implementovali náhodný proces prechádzajúci cez priestor stavov. Tento náhodný proces spĺňa Markovovskú vlastnosť, ktorá je popísaná ako takzvaná bezpamäťovosť. Hovorí o tom, že distribúcia pravdepodobností nasledujúceho stavu závisí len od terajšieho stavu a nezáleží na sekvencií udalostí, ktoré mu predchádzali. Vďaka tejto vlastnosti je potrebná pamäť konštantná. Jediné čo si tento algoritmus pamätá, je tabuľka pravdepodobností pomocou ktorej sa rozhoduje aký najbližší symbol vygeneruje. Pre konštantne veľký prefix si Markovovský zdroj vypočíta pravdepodobnosti následujúcich znakov.

Pomocou základnej implementácie Markovovho zdroja dokáže algoritmus používajúci tento zdroj generovať len heslá zložené z kombinácií znakov, ktoré videl na vstupe. Toto pre nás spôsobuje problém, keďže by sme chceli aby náš algoritmus v konečnom čase vygeneroval všetky možné reťazce kratšie ako zadaná dĺžka. Na vyriešenie tohto problému sme sa rozhodli definovať pravdepodobnosti pre kombinácie znakov, ktoré sa na vstupe nevyskytli.

KAPITOLA 5

Implementácia

5.1 Pravdepodobnostná bezkontextová gramatika

Pravdepodobnostná bezkontextová gramatika je bezkontextová gramatika, ktorej pravidlá majú priradenú pravdepodobnosť. Pravdepodobnostná bezkontextová gramatika G je pätica $G = (M, T, R, S, P)$, kde:

- $M = N^i : i = 1, ..., n$ je množina neterminálov
- $T = w^k : k = 1, ..., V$ je množina terminálov
- $R = N^i \rightarrow \zeta^j : \zeta^j \epsilon (M \cup T)^*$ je množina pravidiel
- $S = N^1$ je počiatočný neterminál
- P je množina pravdepodobností pravidiel, pre ktoré platí $\forall i \sum_i P(N^i \rightarrow \zeta^j) = 1$

5.1.1 Tvorba gramatiky

Pri tvorbe gramatiky sme si dali za cieľ zaistiť, aby gramatika spĺňala určité podmienky:

- Generovať všetky reťazce zo vstupnej abecedy kratšie ako používateľom zadaná maximálna dĺžka
- Vygenerovať každý reťazec práve raz

V nami vytvorenej gramatike bude kategorizovať slová podľa ich zloženia z jednotlivých typov znakov. Pre každý takýto typ znakov vygenerujeme všetky možné reťazce zložené zo znakov tohto typu. Následne vytvoríme všetky možné predpisy zložené z kombinácií týchto typov. Uvažujme, že typ 1 obsahuje t_1 znakov a typ 2 obsahuje t_2 znakov a budeme vytvárať reťazce dĺžky 2, kde každý znak patrí do iného typu. Všetkých

takýchto reťazcov je $t_1 * t_2 * 2$. Naša gramatika si bude pamätať len $t_1 + t_2 + 2$ položiek, z ktorých dokáže spájaním vygenerovať všetkých $t_1 * t_2 * 2$.

Jednoduché neterminály Prvý typ neterminálov, ktoré budeme nazývať jednoduché, obsahujú pravidlá, ktoré majú na pravej strane pravidla iba terminálne symboly. Keďže naša vstupná abeceda obsahuje okolo 70 znakov, medzi ktoré patria veľké a malé písmena, cifry a niektoré často používané symboly, rozhodli sme sa ich rozdeliť do jednotlivých skupín. Pre každú z týchto skupín sme vytvorili neterminál, ktorý bude reprezentovať sekvenciu pevnej dĺžky zloženú zo znakov danej skupiny. V gramatike tieto neterminály vyjadrujeme pomocou prvého písmena anglického názvu danej skupiny.

- U - veľké písmená
- L - malé písmená
- D - cifry
- S - symboly

Každý jednoduchý neterminál sa teda skladá z písmena vyjadrujúceho skupinu znakov, ktoré generuje, a čísla popisujúceho dĺžku sekvencie na pravej strane pravidiel tohto neterminálu. Ako napríklad jednoduchý neterminál D_1 vyjadruje pravidlá $D_1 \rightarrow 1|2|...|9|0$. Keďže všetkých reťazcov terminálnych znakov dĺžky k nad abecedou veľkosti n je n^k, rozhodli sme sa zadefinovať maximálnu veľkosť jednoduchého neterminálu, vyjadrujúcu maximálné povolené k. Na základe testov zo sekcií 6.1 a 6.3 sa ako vhodná hodnota ukázalo $k = 5$. Avšak v prípade generovania malého počtu hesiel sa kvôli optimalizácií času oplatí použiť hodnotu $k = 4$.

Zložené neterminály Jednoduché neterminály nám pomáhajú vyjadrovať sekvenciu znakov práve jedného z vyššie vymenovaných typov. Aby sme boli schopní popísať ľubovolný reťazec tvorený znakmi vstupnej abecedy, budeme tieto jednoduché neterminály skladať do skupín, tieto skupiny budeme nazývať zložené neterminály. Tieto neterminály vyjadrujú vždy jeden možný predpis pre terminálne slovo. Napríklad zložený neterminál $U_1L_3D_4$ vyjadruje všetky terminálne slová začínajúce na veľké písmeno nasledované tromi malými písmenami, ukončené štvoricou cifier.

Kvôli dodržaniu jednoznačnosti generovania nedovoľujeme, aby sa vyskytovali 2 jednoduché neterminály rovnakého typu za sebou. V prípade, že potrebujeme popísať sekvenciu terminálnych znakov jedného typu dlhšiu ako povolené maximum (popísane vyššie), rozdelíme túto sekvenciu do viacerých jednoduchých neterminálov pažravým algoritmom, čiže každý z týchto neterminálov zoberie maximálny možný počet znakov sekvencie. Ak zoberieme heslo pozostávajúce z 9 cifier ako napríklad jedno z najpoužívanejších *123456789* a máme najvyššiu povolenú dĺžku jednoduchého neterminálu

nastavenú na 4, toto heslo bude v našej gramatike zapísané ako $D_4D_4D_1$. Tento spôsob nám zaručí, že nevzniknú dva rôzne zložené neterminály vyjadrujúce ten istý predpis terminálneho slova.

Počiatočný neterminál gramatiky Z bude obsahovať pravidlá prepisujúce tento neterminál na zložené neterminály vyjadrujúce všetky možné predpisy slov kratších ako zadaná maximálna dĺžka.

5.1.2 Počítanie pravdepodobností

Aby sme vedeli čo najlepšie vyhovieť potrebám používateľa, potrebujeme im prispôsobiť našu gramatiku. Tu začnú zohrávať rolu pravdepodobnosti jednotlivých pravidiel našej gramatiky. Našim cieľom je nastaviť našu gramatiku tak, aby generovala heslá podľa pravdepodobnosti použitia daným používateľom. Úspešnosť tohto učenia gramatiky bude záležať od kvality vstupných dát.

Vzhľadom na to, že v dnešnom svete používatelia používajú rôzne služby, ktoré každá odporúča mať jedinečné heslo, používatelia používajú niekoľko hesiel naraz. Tieto heslá by si radi všetky pamätali a preto, ako sme už v úvode spomínali, si často vytvoria pre seba charakteristický spôsob tvorby a zapamätania si týchto hesiel. V ideálnom prípade by sme chceli aby naše vstupné dáta pozostávali z čo najväčšieho počtu hesiel vytvorených pomocou tohto charakteristického spôsobu, keďže každé upresnenie informácií o hľadanom hesle nám zvýši rýchlosť nájdenia tohto hesla.

Keďže cieľom našej práce je nájsť heslo so 100% pravdepodobnosťou, čo v najhoršom prípade znamená vygenerovať všetky možné reťazce kratšie ako zadaná maximálna dĺžka hesla, tak základnú gramatiku s pravidlami vieme vygenerovať dopredu a pravidlá tejto gramatiky sa budú meniť len pri zmene maximálnej dĺžky hesla. Preto používanie nášho algoritmu s rôznymi slovníkmi nevyžaduje vyrábanie novej gramatiky až do momentu kedy sa rozhodneme generovať heslá s inou maximálnou dĺžkou. Pravdepodobnosti prepisovacích pravidiel generujúcich terminálne sekvencie budeme rátať ako percento výskytov danej terminálnej sekvencie spomedzi všetkých sekvencií spadajúcich pod tento neterminál. Práve kvôli tomuto spôsobu sme pridali v implementácií možnosť napísať do vstupného slovníku počty výskytov jednotlivých hesiel, aby mal používateľ možnosť zdôrazniť dôležitosť hesla. Vstupné slovníky, ktoré neskôr používame v našich testoch majú formát, kde na každom riadku je heslo s počtom jeho výskytov oddelené medzerou.

Pri počítaní pravdepodobností zložených neterminálov máme viacero možností ako postupovať.

```
# ideme po pismenach slova zo slovnika
for i in range(1, len(word)):
  # ak sa zmenil typ znaku na male pismeno
  if (word[i] in lower) and (currentNet != 'L'):
    # k zlozenemu neterminalu pridame jednoduchy posledneho
    # videneho typu a velkosti
    rule += currentNet + str(i-startI)
    # pripocitame pocet vyskytov retazca daneho posledneho
    # jednoducheho neterminalu
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    # pripocitame pocet vyskytov daneho jednoducheho neterminalu
    ruleCount[currentNet + str(i-startI)] += occ
    # dalsi typ zacina na i-tej pozicii
    startI = i
    # je to retazec malych pismen
    currentNet = 'L'
    # budujeme od zaciatku
    currentSubstring=''
  # ak sa zmenil typ znaku na velke pismeno
  elif (word[i] in upper) and (currentNet != 'U'):
    # k zlozenemu neterminalu pridame jednoduchy posledneho
    # videneho typu a velkosti
    rule += currentNet + str(i-startI)
    # pripocitame pocet vyskytov retazca daneho posledneho
    # jednoducheho neterminalu
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    # pripocitame pocet vyskytov daneho jednoducheho neterminalu
    ruleCount[currentNet + str(i-startI)] += occ
    # dalsi typ zacina na i-tej pozicii
    startI = i
    # je to retazec velkych pismen
    currentNet = 'U'
    # budujeme od zaciatku
    currentSubstring=''
```

Zdrojový kód 5.1: Úprava pravidiel na základe vstupného slova

Priamo zo vstupného slovníka Prvý spôsob ako postupovať je identický s tým pre jednoduché neterminály. Pre každé pravidlo gramatiky prepisujúce počiatočný neterminál na zvolený zložený neterminál vypočítame jeho pravdepodobnosť ako pomer počtu výskytov tohto neterminálu a výskytov všetkých neterminálov dohromady. Existuje taktiež viacero spôsobov ako počítať výskyty zložených neterminálov.

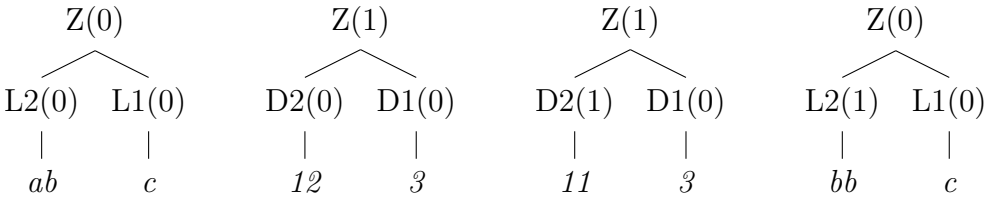
- Do výskytov počítame len výskyty hesiel ktoré sú presne reprezentované daným neterminálom
- Do výskytov započítame aj výskyty kedy je zvolený neterminál podreťazcom

iného neterminálu

V oboch týchto variantoch počítame počty výskytov jednoduchých neterminálov. Rozdiel medzi týmito variantami ukážeme na príklade. Majme na vstupe heslo, ktoré je reprezentované zloženým neterminálom $U_2L_3D_2$, použitím prvého variantu tento zložený neterminál vygeneruje jedno zvýšenie počtu výskytov a to pre tento konkrétny neterminál. Ukážka kódu implementujúceho prvý variant 5.1. Druhý variant by na tomto neterminali vyvolal 2 navýšenia počtu výskytov a to osobitne pre zložené neterminály U_2L_3 a $U_2L_3D_2$. Aby sme upravili náš program na druhý variant pridali sme pre každú zmenu typu neterminálu pripočítanie výskytov k doteraz vytvorenému zloženému neterminálu, ukážka takto upraveného kódu je v 5.3.

Rekurzívne Ďalší spôsob spočíva v tom, že zo vstupného slovníka vypočítame pravdepodobnosti len pre jednoduché netermiály. Následne pre zložené neterminály počítame pravdepodobnosti ako súčin pravdepodobností jednoduchých neterminálov, ktoré daný neterminál obsahuje.

5.1.3 Generovanie hesiel



Dôležitým aspektom používania bezkontextových gramatík je práve spôsob generovania hesiel. Našim hlavným cieľom bolo generovanie hesiel pomocou gramatiky od najpravdepodobnejšieho z nich. Aby sme toto vedeli robiť čo najefektívnejšie, ako prvé sme si usporiadali pravidlá jednotlivých neterminálov vzostupne podľa pravdepodobnosti. Na diagramoch na začiatku sekcie 5.1.3 vidíme stromy odvodenia pre 4 najpravdepodobnejšie slová gramatiky popísanej v tabuľke 5.1. Čísla v zátvorkách určujú index použitého pravidla pre daný neterminál. Výpisom týchto hodnôt do poľa v preorder poradí dostaneme vektor určujúci odvodenia daného terminálneho slova. Napríklad na základe diagramov na začiatku sekcie 5.1.3 slovu *bbc* prislúcha vektor $(0, 1, 0)$.

V 5.6 vidíme pseudokód pre algoritmus generujúci heslá z gramatiky. Algoritmus začína pridaním dvojice rád vektora a vektor do prioritnej fronty. Táto fronta zoraďuje prvky na základe ich pravdepodobnosti. Následne algoritmus vojde do cyklu, z ktorého vyjde len ak vygeneruje požadovaný počet hesiel alebo vyprázdni celú frontu, čo sa stane v prípade, že by vygeneroval všetky možné heslá pre danú gramatiku.

Tabuľka 5.1: Ukážka počítania pravdepodobností - gramatika

Slovo	Počet výskytov
abc	20
123	14
113	13
bbc	12
ab	8
ca	8
c2c	6
bb3	5
3a3	5
ca2	5
a	2
b	2
2	2
31	1
122	1

Pravidlo	$p_{zakladna}$	$p_{podretazce}$	$p_{rekurzivne}$
$L1 \rightarrow c$	0,8035	0,8035	0,8035
$L1 \rightarrow a$	0,1428	0,1428	0,1428
$L1 \rightarrow b$	0,0535	0,0535	0,0535
$L2 \rightarrow ab$	0,4328	0,4328	0,4328
$L2 \rightarrow bb$	0,2686	0,2686	0,2686
$L2 \rightarrow ca$	0,2089	0,2089	0,2089
$L2 \rightarrow aa\|ac\|ba\|bc\|cb\|cc$	0,0149	0,0149	0,0149
$D1 \rightarrow 3$	0,7288	0,7288	0,7288
$D1 \rightarrow 2$	0,2542	0,2542	0,2542
$D1 \rightarrow 1$	0,0169	0,0169	0,0169
$D2 \rightarrow 12$	0,4210	0,4210	0,4210
$D2 \rightarrow 11$	0,3680	0,3680	0,3680
$D2 \rightarrow 13\|21\|22\|23\|31\|32\|33$	0,0263	0,0263	0,0263
$Z \rightarrow L2L1$	0,2796	0,2558	0,0061
$Z \rightarrow D2D1$	0,2457	0,2248	0,0004
$Z \rightarrow L2$	0,1440	0,1317	0,1440
$Z \rightarrow L2D1$	0,0932	0,0852	0,0036
$Z \rightarrow L1D1L1$	0,0593	0,0542	0,00004
$Z \rightarrow D1L1D1$	0,0508	0,0465	0,00002
$Z \rightarrow L1$	0,0423	0,0387	0,0423
$Z \rightarrow D1$	0,0254	0,0232	0,0254
$Z \rightarrow D2$	0,0169	0,0155	0,0169
$Z \rightarrow L1D1$	0,0084	0,0542	0,0010
$Z \rightarrow L1D2$	0,0084	0,0077	0,0007
$Z \rightarrow D2L1$	0,0084	0,0077	0,0007
$Z \rightarrow D1L1$	0,0084	0,0465	0,0010
$Z \rightarrow D1L2$	0,0084	0,0077	0,0036

Tabuľka 5.2: Ukážka krokov algoritmu pre neterminál $U_1L_3D_2$

Ľavá strana	Pravá strana	p	i
U_1	A	0,7	0
U_1	B	0,2	1
U_1	C	0,1	2
L_4	minf	0,6	0
L_4	fmfi	0,3	1
L_4	dipl	0,1	2
D_2	47	0,8	0
D_2	42	0,2	1

p	(rád, vektor)	slovo
0.336	(0, [0, 0, 0])	Aminf47
0.168	(1, [0, 1, 0])	Afmfi47
0.096	(0, [1, 0, 0])	Bminf47
0.084	(2, [0, 0, 1])	Aminf42
0.096	(0, [1, 0, 0])	Bminf47
0.084	(2, [0, 0, 1])	Aminf42
0.056	(1, [0, 2, 0])	Adipl47
0.042	(2, [0, 1, 1])	Afmfi42

Ako je vidieť na diagramoch na začiatku sekcie 5.1.3, stromy odvodenia pre heslá s rovnakým predpisom majú takú istú štruktúru. Pri zmene použitého pravidla pre počiatočný neterminál musíme ošetriť potenciálnu zmenu štruktúry stromu odvodenia (riadky 11 až 14). Následne algoritmus prejde všetky susedné vektory práve spracovaného vektoru a pridá ich do fronty s aktualizovaným rádom a pravdepodobnosťou (riadky 16 až 20).

Dôležitým prvkom tohto algoritmu je rád vektora, bez ktorého by sme generovali veľké množstvo rovnakých vektorov, ktoré by sme dostali úpravou hodnôt vektora v inom poradí. Napríklad ak by sme zvýšili najprv index na pozícii 1 a následne 3, dostali by sme to isté, ako keby sme zvýšili na pozícii 3 a potom 1. Rád vektoru je číslo určujúce pozíciu najvyššieho zmeneného indexu. Pri generovaní susedných vektorov dovolíme meniť indexy len na pozíciach vyšších alebo rovných ako je aktuálny rád vektora. Týmto zaručíme zmenu použitých pravidiel v preorder poradí vzhľadom na strom odvodenia.

V tabuľke 5.2 demonštrujeme dva kroky nášho algoritmu na generovanie hesiel. V tomto príklade sa sústredíme na generovanie rôznych terminálnych slov zo zloženého neterminálu $U_1L_3D_2$. V ľavej časti tabuľky môžme vidieť zadefinované prepisovacie pravidlá pre tento neterminál aj s pravdepodobnosťami, ktoré majú priradené. V pravej časti simulujeme obsah našej prioritnej fronty, kde dvojitou vodorovnou čiarou sú oddelené stavy tejto fronty v rôznych krokoch algoritmu. V počiatočnom stave máme vo fronte prvý prvok ukazujúci na najpravdepodobnejšie heslo generované z definovaných pravidiel. Algoritmus tento prvok vyberie z fronty a následne tam vloží prvky označujúce heslá vzdialené práve na jednu zmenu použitého pravidla. Tieto novo pridané prvky sú automaticky zoradené podľa pravdepodobností vďaka tomu, že na pozadí je naša fronta reprezentovaná prioritnou haldou.

V druhom kroku algoritmu vyberie prvok z najvyššou pravdepodobnosťou. Opäť do fronty pridáme prvky vyjadrujúce heslá vzdialené na jednu zmenu použitého pravidla.

Tu si treba všimnúť, že sme nepridali prvok s vektorom [1, 1, 0], keďže rád práve spracovaného prvku je 1, čiže môžme meniť len indexy 1 a 2, ktoré sú väčšie alebo rovné ako rád vektora. Týmto spôsobom algoritmus pokračuje až pokiaľ sa nevygeneruje požadovaný počet hesiel alebo sa nevyprázdni fronta. Fronta sa môže vyprázdniť len ak prejdeme cez všetky možné heslá, keďže jediný moment kedy nepribudne žiaden prvok do fronty je ak rád vektora bude rovný jeho dĺžke a v poslednom jednoduchom neterminál sme použili už všetky jeho pravidlá.

V zdrojovom kóde 5.7 môžme vidieť časť kódu zodpovednú za napĺňanie prioritnej fronty ďalšími kandidátmi na najbližšie vygenerované heslo. Premenná *task* je usporiadaná dvojica (rád, vektor). Algoritmus prejde od člena určeného rádom vektora až po koniec vektora (riadok 1) a pre každý prvok posunie index ukazujúci na aktuálne použitý prvok (riadok 4). Taktiež vypočíta pravdepodobnosť hesla reprezentovaného novým stavom vektora (riadok 3 a 7). Túto pravdepodobnosť spolu s usporiadanou dvojicou obsahujúcou zmenený rád vektora a samotný vektor vloží do prioritnej fronty (riadok 8 a 9).

Veľký nedostatok použitia bezkontextových gramatík je veľkosť pamäte, ktorá je potrebná na samotný zápis gramatiky na disku. V tvare JSON mal pri gramatike generujúcej 12 znakové heslá približne 1 gigabajt. Väčší problém nastáva s pamäťou použitou pri samotnom behu algoritmu. Počas testov na zistenie vplyvu jednotlivých parametrov na veľkosť potrebnej pamäte sme zistili, že najväčšia časť použitej pamäte je potrebná na uloženie našej gramatiky v asociatívnom poli.

Poslednou vecou čo sme riešili v implementácií bezkontextových gramatík bola možnosť prerušovaného generovania. Vtedy má používateľ možnosť generovať požadované heslá po ľubovoľne veľkých častiach.

Pri ukončení generovania bez ohľadu na dôvod, splnenie počtu vygenerovaných hesiel alebo prerušenie od používateľa, zapíšeme do súboru aktuálny stav generovania. Tento stav zahŕňa:

- Cesta ku gramatike, ktorá bola použitá na generovanie
- Aktuálny zoznam prvkov prioritnej fronty

Vďaka týmto dvom informáciam dokáže náš algoritmus následne pokračovať v generovaní slov od posledného vygenerovaného.

5.2 Markovovský zdroj

Po implementácií vyššie uvedeného algoritmu na generovanie hesiel pomocou pravdepodobnostných bezkontextových gramatík a odhalení nedostatkov čo sa týka pamäťovej náročnosti, sme sa rozhodli implementovať ešte jednu metódu. Tou je Markovovský zdroj. Ako sme písali v predošlej kapitole, jedná sa o náhodný proces, pre ktorý platí, že nasledujúce vygenerované písmeno záleží len od posledného stavu a nie od postupnosti stavov, ktoré mu predchádzali. Markovovské zdroje sa veľmi často používajú práve pri generovaní prirodzeného jazyka. Práve preto boli vhodným kandidátom pre generovanie hesiel na základe znalostí získaných zo vstupného slovníka.

Bohužiaľ táto metóda nespĺňa ani jednu z podmienok, ktoré sme si dali za cieľ pri bezkontextových gramatikách:

- Generovať všetky reťazce zo vstupnej abecedy kratšie ako používateľom zadaná maximálna dĺžka
  - Pravdepodobnosti jednotlivých znakov sú inicializované na 0, Markovovský zdroj ich nikdy nevygeneruje
- Vygenerovať každý reťazec práve raz
  - Keďže tomuto zdroju nič nebráni v tom vygenerovať viac krát počas behu to isté slovo

Aj keď druhú podmienku nevedia Markovovské zdroje splniť už priamo z definície, s prvou sme sa pokúsili niečo vymyslieť. Najprv sme sa pokúšali inicializovať pravdepodobnosti všetkých znakov na nenulovú hodnotu. Toto však spôsobilo, že sa zdroj relatívne ľahko dostal medzi prefixy, ktoré neboli definované. Pri takýchto prefixoch majú všetky znaky rovnakú pravdepodobnosť. Dôsledkom tohto nastávalo cyklenie sa v týchto neznámych stavoch, čo spôsobovalo generovanie dlhých nezmyselných reťazcov znakov. Preto sme sa snažili nájsť spôsob ako nastaviť pravdepodobnosti nevidených stavov na nenulové, avšak dostatočne malé aby sa v nich samotný algoritmus necyklil.

Pri takto definovanom Markovovskom zdroji dokážeme všetky stavy tohto zdroja rozdeliť do dvoch množín prefixov. Videné prefixy sú také, ktoré aspoň raz nastali pri učení podľa vstupného slovníka. Takéto stavy majú pre aspoň jeden znak slovníka nenulovú pravdepodobnosť. Druhou väčšou skupinou prefixov sú nevidené prefixy, ktoré sa nevyskytli nikde vo vstupnom slovníku a preto pre všetky znaky našej abecedy je pravdepodobnosť prechodu nulová. Keďže chceme upraviť náš Markovovský zdroj tak, aby mal možnosť generovať všetky možné heslá, potrebujeme tieto nulové pravdepodobnosti zmeniť na nenulové.

Tabuľka 5.3: Ukážka Markovovského zdroja - prefix dĺžky 2

Slovník	Výskyty
abc	3
12	2
b1c	2
bb1	2

	a	b	c	1	2	\n
ab	-	-	3	-	-	-
bc	-	-	-	-	-	3
c\n	-	1	-	2	-	-
\n1	-	-	-	-	2	-
12	-	-	-	-	-	2
2\n	-	2	-	-	-	-
\nb	-	-	-	2	-	-
b1	-	-	2	-	-	1
1c	-	-	-	-	-	2
bb	-	-	-	1	-	-

Videné stavy nemusia mať určené pravdepodobnosti pre všetky znaky nášho vstupného slovníka. Pre tieto znaky nastavíme pravdepodobnosti, ktoré v pôvodnom algoritme majú nulovú pravdepodobnosť, na hodnotu $\varepsilon > 0$. Táto hodnota by mala vyjadrovať pravdepodobnosť prechodu zo stavu v ktorom je prefix známy (z dát vo vstupnom súbore) do stavu kedy je prefix neznámy a pravdepodobnosti všetkých znakov sú nulové. Domnievame sa, že pre správne fungovanie algoritmu by hodnota ε mala byť niekoľkonásobne menšia ako najnižšia známa pravdepodobnosť pre daný prefix. Vplyv tejto konštanty na výkon Markovovského zdroja je znázornený na grafe 6.12.

Po prechode nášho algoritmu do stavu, ktorý nebol videný počas inicializácie programu potrebujeme nastaviť pravdepodobnosti všetkých znakov našej abecedy. Pravdepodobnosť jednotlivých znakov nastavíme na hodnotu ε ak sa jedná o znak pomocou ktorého v ďalšom kroku algoritmu vznikne videný prefix. V prípade, že znak dostane náš algoritmus do stavu s iným neznámym prefixom, nastavíme tomuto znaku pravdepodobnosť δ, ktorá je niekoľkokrát menšia ako ε. Opäť vplyv tejto konštanty na výsledky algoritmu je znázornený na grafe 6.12. Použitím hodnôt ε a δ by sme mali dostať algoritmus používajúci Markovovský zdroj do stavu, kedy v konečnom čase dokáže vygenerovať ľubovolný počet hesiel.

Jediná informácia, ktorú si algoritmus používajúci Markovovský zdroj pamätá, je tabuľka pravdepodobností nasledovania znakov po danom prefixe. V pôvodnej implementácií tohto algoritmu sme si zoznam prefixov pamätali ako kľúče v asociatívnom poli. Hodnoty týchto kľúčov boli zoznamy pravdepodobností o veľkosti vstupnej abecedy. Index pravdepodobnosti zodpovedal indexu znaku v zápise vstupnej abecedy, ktorému táto pravdepodobnosť prislúcha.

Pri implementácií upraveného Markovovského zdroja sme spravili optimalizáciu, pri ktorej si informáciu o pravdepodobnostiach pamätáme len pre znaky a prefixy, ktoré sa

vyskytujú vo vstupnom slovníku, ktorý dostal na vstupe. Týmto nám vznikne riedka matica, ktorú si v programe pamätáme pomocou asociatívneho poľa. Toto asociatívne pole obsahuje kľúče pre znaky ku ktorým vieme priradiť pravdepodobnosť na základe vstupného slovníka. Taktiež sme si definovali množinu videných stavov, ktorá bola implementovaná pomocou pythonovského typu *set*. Toto avšak neprinieslo očakávane zrýchlenie a overovanie, či sme stav videli vo vstupnom slovníku priamo v asociatívnom poli s pravdepodobnosťami.

V prípade, že sa algoritmus dostane do stavu, kedy sa aktuálny prefix nenachádzal vo vstupnom slovníku, prejde cez všetky znaky vstupnej abecedy a každému priradí pravdepodobnosť ε ak sa vygenerovaním tohto znaku dostane do známeho prefixu alebo pravdepodobnosť δ ak pridanie tohto znaku vedie do ďalšie stavu s neznámym prefixom. Rozdiely v čase a potrebnej pamäti na vygenerovanie pevne daného počtu hesiel medzi pôvodnou implementáciou a našou optimalizáciou je ukázaný v grafe 6.5.

Rozdiely vo veľkosti ε oproti najmenšej nenulovej pravdepodobnosti pre ten prefix a δ od ε by mali zaručiť, že algoritmus sa snaží preferovať heslá, ktoré sa skladajú z kombinácií znakov videných na vstupe. Výsledky tohto algoritmu pre rôzne nastavené hodnoty koeficientov ε a δ sú znázornené v kapitole Evaluácia výsledkov.

V tomto prípade sme zvažovali aj použitie možnosti zmeny týchto pravdepodobností počas behu programu podobne ako pri simulovanom žíhaní. Zo začiatku by sme tieto pravdepodobnosti nastavili na relatívne nízke hodnoty a s počtom hesiel vygenerovaných našim algoritmom by sa tieto hodnoty zväčšovali aby mal algoritmus vyššiu tendenciu dostať sa aj k menej pravdepodobným heslám. Bohužiaľ z dôvodu časovej tiesne sme nenašli priestor na implementáciu a preskúmanie takto upraveného algoritmu.

```
  # ak sa zmenil typ znaku na cislicu
  elif (word[i] in digit) and (current != 'D'):
    # k zlozenemu neterminalu pridame jednoduchy posledneho
    # videneho typu a velkosti
    rule += currentNet + str(i-startI)
    # pripocitame pocet vyskytov retazca daneho posledneho
    # jednoducheho neterminalu
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    # pripocitame pocet vyskytov daneho jednoducheho neterminalu
    ruleCount[currentNet + str(i-startI)] += occ
    # dalsi typ zacina na i-tej pozicii
    startI = i
    # je to retazec cislic
    currentNet = 'D'
    # budujeme od zaciatku
    currentSubstring=''
  # ak sa zmenil typ znaku na symbol
  elif (word[i] in special) and (current != 'S'):
    # k zlozenemu neterminalu pridame jednoduchy posledneho
    # videneho typu a velkosti
    rule += currentNet + str(i-startI)
    # pripocitame pocet vyskytov retazca daneho posledneho
    # jednoducheho neterminalu
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    # pripocitame pocet vyskytov daneho jednoducheho neterminalu
    ruleCount[currentNet + str(i-startI)] += occ
    # dalsi typ zacina na i-tej pozicii
    startI = i
    # je to retazec symbolov
    currentNet = 'S'
    # budujeme od zaciatku
    currentSubstring=''
  # presiahli sme velkost jednoducheho neterminalu
  elif len(currentSubstring) >= maxNetSize:
    # k zlozenemu neterminalu pridame jednoduchy posledneho
    # videneho typu a velkosti
    rule += currentNet + str(i-startI)
    # pripocitame pocet vyskytov retazca daneho posledneho
    # jednoducheho neterminalu
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    # pripocitame pocet vyskytov jednoducheho daneho neterminalu
    ruleCount[currentNet + str(i-startI)] += occ
    # dalsi typ zacina na i-tej pozicii
    startI = i
    # budujeme od zaciatku
    currentSubstring=''
  currentSubstring += word[i]
...
```

Zdrojový kód 5.2: Úprava pravidiel na základe vstupného slova - Pokračovanie

```
for i in range(1, len(word)):
  if (word[i] in lower) and (currentNet != 'L'):
    rule += currentNet + str(i-startI)
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    ruleCount[currentNet + str(i-startI)] += occ
    # este sme takyto zlozeny neterminal nevideli
    if not rule in rulez['Z']:
      rulez['Z'][rule] = 1
    # pripocitame pocet vyskytov tohto zlozeneho netermialu
    rulez['Z'][rule] += occ
    # pripocitame pocet vyskytov nejakeho neterminalu
    ruleCount['Z'] += occ
    startI = i
    currentNet = 'L'
    currentSubstring=''
  elif (word[i] in upper) and (currentNet != 'U'):
    rule += currentNet + str(i-startI)
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    ruleCount[currentNet + str(i-startI)] += occ
    # este sme takyto zlozeny neterminal nevideli
    if not rule in rulez['Z']:
      rulez['Z'][rule] = 1
    # pripocitame pocet vyskytov tohto zlozeneho netermialu
    rulez['Z'][rule] += occ
    # pripocitame pocet vyskytov nejakeho neterminalu
    ruleCount['Z'] += occ
    startI = i
    currentNet = 'U'
    currentSubstring=''
```

Zdrojový kód 5.3: Pripočítanie výskytov k podmnožinám zložených neterminálov

```
  elif (word[i] in digit) and (current != 'D'):
    rule += currentNet + str(i-startI)
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    ruleCount[currentNet + str(i-startI)] += occ
    # este sme takyto zlozeny neterminal nevideli
    if not rule in rulez['Z']:
      rulez['Z'][rule] = 1
    # pripocitame pocet vyskytov tohto zlozeneho netermialu
    rulez['Z'][rule] += occ
    # pripocitame pocet vyskytov nejakeho neterminalu
    ruleCount['Z'] += occ
    # dalsi typ zacina na i-tej pozicii
    startI = i
    # je to retazec cislic
    currentNet = 'D'
    # budujeme od zaciatku
    currentSubstring=''
  # ak sa zmenil typ znaku na symbol
  elif (word[i] in special) and (current != 'S'):
    # k zlozenemu neterminalu pridame jednoduchy posledneho
    # videneho typu a velkosti
    rule += currentNet + str(i-startI)
    # pripocitame pocet vyskytov retazca daneho posledneho
    # jednoducheho neterminalu
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    # pripocitame pocet vyskytov daneho jednoducheho neterminalu
    ruleCount[currentNet + str(i-startI)] += occ
    # este sme takyto zlozeny neterminal nevideli
    if not rule in rulez['Z']:
      rulez['Z'][rule] = 1
    # pripocitame pocet vyskytov tohto zlozeneho netermialu
    rulez['Z'][rule] += occ
    # pripocitame pocet vyskytov nejakeho neterminalu
    ruleCount['Z'] += occ
    startI = i
    currentNet = 'S'
    currentSubstring=''
```

Zdrojový kód 5.4: Pripočítanie výskytov k podmnožinám zložených neterminálov - Pokračovanie

```
  elif len(currentSubstring) >= maxNetSize:
    rule += currentNet + str(i-startI)
    rulez[currentNet + str(i-startI)][str(currentSubstring)] += occ
    ruleCount[currentNet + str(i-startI)] += occ
    # ak uz mame zlozeny neterminal
    if len(rule) > 2:
      # este sme taky nevideli
      if not rule in rulez['Z']:
        rulez['Z'][rule] = 1
      # pripocitame pocet vyskytov tohto zlozeneho netermialu
      rulez['Z'][rule] += occ
      # pripocitame pocet vyskytov nejakeho neterminalu
      ruleCount['Z'] += occ
    startI = i
    currentSubstring=''
  currentSubstring += word[i]
```

Zdrojový kód 5.5: Pripočítanie výskytov k podmnožinám zložených neterminálov - Pokračovanie

```
pocet = 0
fronta.push(pravdepodobnost([0,0,0]), (0, [0,0,0]))
while fronta is not empty:
  # prvok je dvojica (rad vektora, vektor)
  aktualnyPrvok = fronta.pop()
  vypisTerminalneSlovo(aktualnyPrvok[1])
  pocet += 1
  if pocet > pozadovanyPocetHesiel:
    break
  if aktualnyPrvok[0] == 0:
    # index aktualne pouziteho pravidla pre neterminal 'Z'
    tmp = aktualnyPrvok[1][0]
    novyVektor = [tmp + 1] + [0] * velkost(Neterminal['Z'][tmp])
    novyPrvok = (novyVektor, 0)
    fronta.push(pravdepodobnost(novyVektor), novyPrvok)
    aktualnyPrvok[0] += 1

  for x in (aktualnyPrvok[0], velkost(aktualnyPrvok[1])):
    novyVektor = aktualnyPrvok[1]
    novyVektor[x] += 1
    novyPrvok = (novyVektor, x)
    fronta.push(pravdepodobnost(novyVektor), novyPrvok)
```

Zdrojový kód 5.6: Pseudokód generovania hesiel

```
for x in range(task[0],len(task[1])):
  tmp = copy.deepcopy(task[1])
  newpriority = priority / rulez[net[(x-1)*2:x*2]][tmp[x]][1]
  tmp[x] += 1
  if tmp[x] >= len(rulez[net[(x-1)*2:x*2]]):
    continue
  newpriority = newpriority * rulez[net[(x-1)*2:x*2]][tmp[x]][1]
  newtask = (x, tmp)
  add_task(newtask, newpriority)
```

Zdrojový kód 5.7: Generovanie všetkých susedných vektorov

KAPITOLA 6

Evaluácia výsledkov

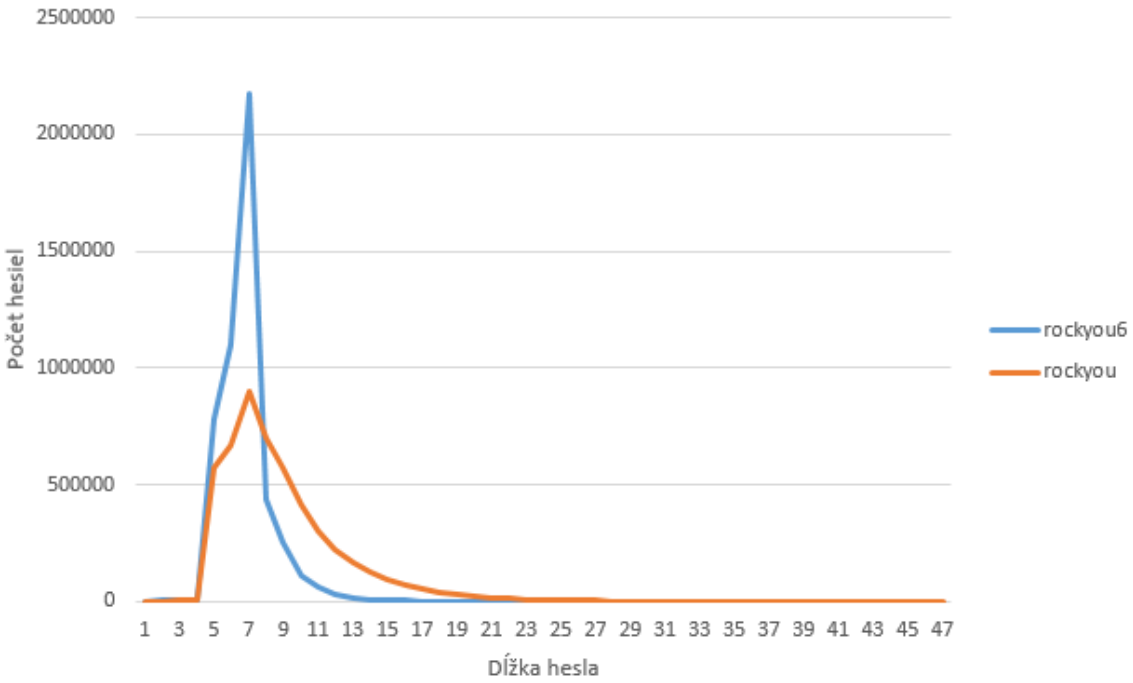
V predošlej kapitole sme bližšie popísali implementáciu našich algoritmov. Táto kapitola sa zameriava na evaluáciu výsledkov týchto algoritmov. Keďže práca sa zaoberá implementáciou algoritmov na generovanie hesiel pomocou vstupného slovníka, bolo potrebné si nájsť vhodný vstupný slovník. Podarilo sa nám nájsť online zdroj slovníkov [2], ktorý obsahuje slovníky s usporiadanými heslami podľa pravdepodobnosti výskytu. Slovník je formátovaný v dvoch stĺpcoch, kde prvý obsahuje informáciu o počte výskytov daného heslá a druhý je samotné heslo.

Testy boli spúšťané na procesore Intel® Core™ i5-4690K s rýchlosťou 3.50GHz so 16 gigabajtami operačnej pamäte na operačnom systéme Windows 10. Implementácia jednotlivých algoritmov bola písaná v jazyku Python 3.5.1 a spúšťaná 64-bitovou verziou interpretéra.

6.1 Časová náročnosť

V tomto základnom teste sme merali čas behu algoritmov pre jednotlivé parametre spustenia. Pod názvom *MarkovV2* rozumieme nami upravený Markovovský zdroj, ktorý generuje všetky možné kombinácie. Pre všetky testy používame slovník *rockyou-withcount* stiahnutý z nášho zdroja slovníkov [2]. Slovník sme upravovali aby obsahoval len heslá zodpovedajúce maximálnej dĺžke hesiel, ktoré generuje bezkontextová gramatika. Týmto sme znížili pravdepodobnosť Markovovského zdroja generovať heslá dlhšie ako stanovené maximum pre gramatiku ako je vidieť na grafe 6.1, kde *rockyou* je nezmenený slovník a *rockyou6* je slovník obmedzený na slová kratšie ako 7 znakov. Stĺpec d tabuľky 6.1 označuje maximálnu dĺžku generovaných hesiel zatiaľ čo stĺpec p vyjadruje maximálnu veľkosť jednoduchého neterminálu v prípade bezkontextovej gramatiky zatiaľ čo pri Markovovskom zdroji vyjadruje dĺžku prefixu podľa ktorého sa tento zdroj rozhoduje. V stĺpci *GEN* je čas potrebný na natrénovanie jednotlivých algoritmov na

vstupný slovník. Ostatné stĺpce ukazujú čas, ktorý jednotlivé algoritmy potrebujú na vygenerovanie daného počtu hesiel. Grafy 6.4, 6.2 a 6.3 ukazujú vplyv parametrov na rýchlosť behu algoritmov.



Obr. 6.1: Distribúcia dĺžok vygenerovaných hesiel

6.2 Pamäťové nároky

Popri meraní času, ktorý algoritmy strávia generovaním pevne daného počtu hesiel sme sledovali použitú pamäť počas behu programu. Výkyvy v niektorých hodnotách boli spôsobené manažovaním pamäte operačným systémom, kedy časti použitej pamäte odkopíroval na disk.

Ďalej sme taktiež porovnali množstvo pamäte, ktorú potrebuje Markovovský zdroj ak si pamätá kompletnú maticu prefixov a znakov abecedy s optimalizáciou, ktorú sme popísali v sekcii 5.2. Na výslednom grafe 6.5 môžme vidieť, že pri krátkych prefixoch není táto optimalizácia efektívna nakoľko pamäť navyše potrebná na použitie asociatívneho poľa je relatívne vysoká. Pri väčšej dĺžke prefixu už začína byť rozdiel v potrebnej pamäti signifikantný.

6.3 Výstupné heslá

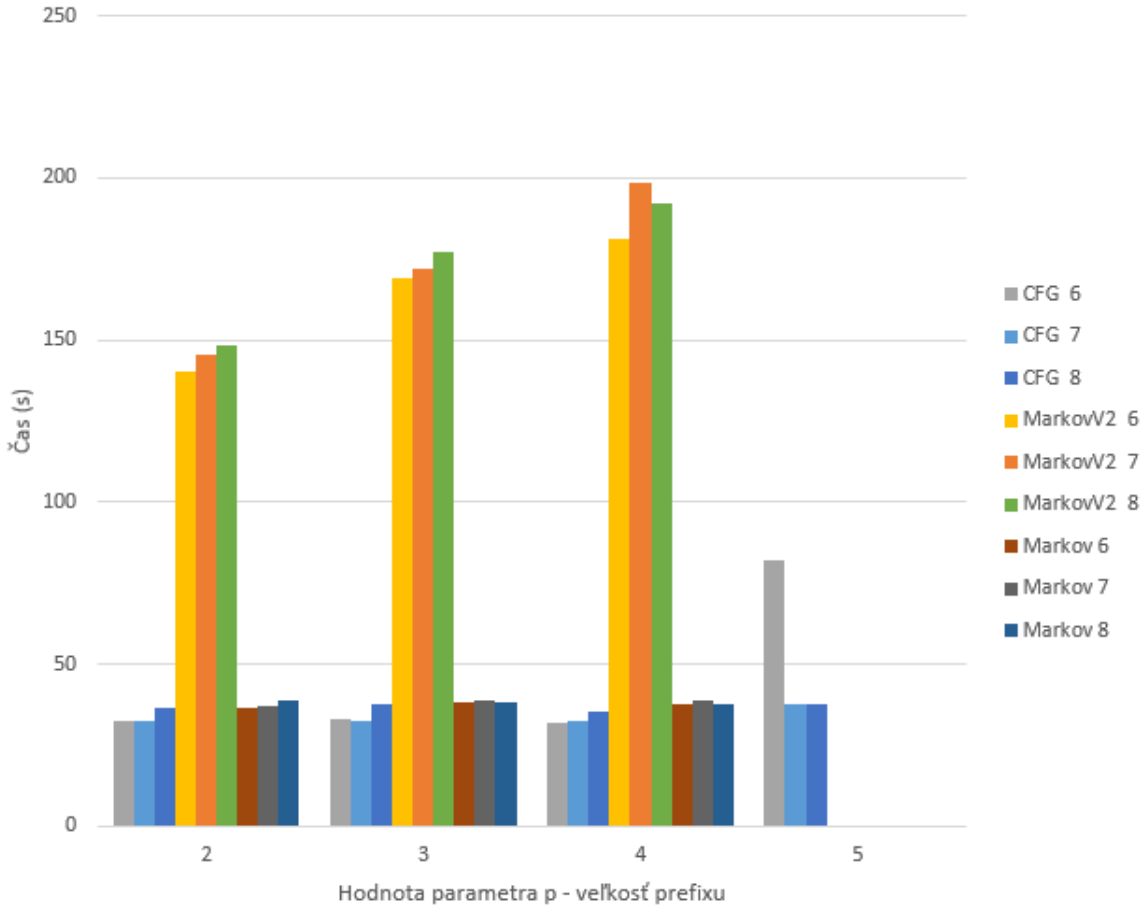
Po overení časovej zložitosti generovania hesiel pomocou jednotlivých algoritmov sme skúmali rôzne vlastnosti vygenerovaných slovníkov s heslami. Cieľom týchto testov bolo

Tabuľka 6.1: Časy v sekundách pre slovník rockyou-withcount

	d	p	GEN	10000	50000	100000	500000	1000000	5000000
CFG	6	2	20,983	0,576	3,116	6,235	32,584	65,005	345,186
CFG	6	3	18,592	0,632	3,290	6,730	33,276	69,801	357,483
CFG	6	4	18,128	0,623	3,508	6,576	31,971	65,760	331,291
CFG	6	5	44,653	0,615	3,179	54,054	81,957	116,699	397,522
Markov	6	2	14,864	0,699	3,774	7,398	36,241	74,244	384,801
Markov	6	3	17,705	0,755	4,098	7,612	38,43	75,553	379,393
Markov	6	4	23,36	0,726	3,768	7,564	37,882	77,203	383,682
MarkovV2	6	2	13,170	2,709	13,168	27,195	140,057	278,877	1394,651
MarkovV2	6	3	15,446	3,350	16,948	33,538	169,359	340,262	1681,150
MarkovV2	6	4	16,957	3,606	18,043	36,435	181,323	360,489	1819,145
	d	p	GEN	10000	50000	100000	500000	1000000	5000000
CFG	7	2	45,993	0,602	3,099	6,382	32,200	65,867	336,915
CFG	7	3	40,868	0,574	3,185	6,271	32,379	64,441	350,051
CFG	7	4	39,640	0,630	3,269	6,496	32,512	66,543	337,664
CFG	7	5	65,633	0,564	3,284	6,835	37,462	72,955	353,728
Markov	7	2	36,147	0,742	3,848	7,69	36,996	75,469	386,964
Markov	7	3	40,321	0,744	3,714	7,515	38,572	75,271	378,42
Markov	7	4	49,719	2,335	3,855	7,423	38,86	77,215	390,586
MarkovV2	7	2	31,485	2,925	14,684	29,606	145,325	298,865	1456,182
MarkovV2	7	3	34,116	3,478	16,919	33,483	172,056	339,607	1699,369
MarkovV2	7	4	37,530	3,728	18,540	37,156	198,664	390,051	1897,855
	d	p	GEN	10000	50000	100000	500000	1000000	5000000
CFG	8	2	80,10	0,640	3,209	6,545	36,207	73,112	375,780
CFG	8	3	69,53	0,667	3,472	7,066	37,623	73,635	379,328
CFG	8	4	65,34	0,684	3,567	7,155	35,082	72,635	372,207
CFG	8	5	90,795	0,614	3,590	7,132	37,472	73,995	366,613
Markov	8	2	62,266	0,727	3,788	7,684	38,614	76,573	380,002
Markov	8	3	70,651	0,719	3,962	7,37	38,298	77,763	391,823
Markov	8	4	84,152	0,736	3,713	7,648	37,6	76,351	380,466
MarkovV2	8	2	51,49	2,965	15,080	30,405	148,388	294,671	1498,122
MarkovV2	8	3	57,23	3,522	17,446	34,795	176,967	354,904	1777,306
MarkovV2	8	4	64,75	4,011	19,731	39,810	192,103	395,619	2002,240

Tabuľka 6.2: Pamäť v MB pre slovník rockyou-withcount

	d	p	GEN	10000	50000	100000	500000	1000000	5000000
CFG	6	2	17,30	19,05	33,26	46,72	117,80	181,35	518,33
CFG	6	3	22,25	29,97	36,38	41,95	70,78	96,19	462,72
CFG	6	4	148,04	263,18	269,29	275,54	310,76	346,15	608,03
CFG	6	5	4661,65	6116,98	6117,87	6119,98	6143,38	6155,72	6349,73
Markov	6	2	37,64	37,66	37,64	37,63	37,82	37,89	37,80
Markov	6	3	488,00	488,04	488,04	488,02	488,14	488,28	488,10
Markov	6	4	3661,08	3661,22	3661,31	3661,08	3661,39	3661,28	3661,25
MarkovV2	6	2	27,60	31,55	31,62	31,66	31,54	31,52	31,41
MarkovV2	6	3	122,99	152,29	152,43	152,53	152,25	152,12	152,52
MarkovV2	6	4	604,64	711,12	710,98	711,14	711,20	711,02	711,26
	d	p	GEN	10000	50000	100000	500000	1000000	5000000
CFG	7	2	18,76	23,77	34,06	44,46	108,55	198,31	693,54
CFG	7	3	24,29	32,77	39,38	45,56	88,24	137,11	464,00
CFG	7	4	149,79	267,46	275,87	284,02	330,79	377,35	669,59
CFG	7	5	4664,87	6120,33	6124,66	6128,76	6156,38	6188,21	6430,26
Markov	7	2	38,25	38,41	38,49	38,36	38,28	38,31	38,46
Markov	7	3	643,87	644,04	643,85	643,77	643,89	643,96	643,96
Markov	7	4	5278,48	5278,41	5278,29	5278,44	5278,47	5278,63	5278,55
MarkovV2	7	2	31,01	36,36	36,16	36,13	36,17	36,10	36,07
MarkovV2	7	3	168,70	210,99	211,10	211,02	210,94	210,85	210,79
MarkovV2	7	4	928,83	1098,33	1098,08	1098,03	1098,09	1098,05	1098,26
	d	p	GEN	10000	50000	100000	500000	1000000	5000000
CFG	8	2	25,26	36,61	47,14	58,01	121,60	184,49	1540,53
CFG	8	3	37,71	46,67	60,41	75,75	185,75	317,56	1274,87
CFG	8	4	214,79	279,68	289,30	298,76	352,48	407,48	757,99
CFG	8	5	4674,50	6133,58	6139,11	6144,55	6174,67	6204,51	6422,41
Markov	8	2	38,57	38,73	38,82	38,7	38,63	38,68	38,74
Markov	8	3	787,69	787,9	787,97	787,91	787,99	787,89	787,96
Markov	8	4	6926,67	6927,08	6927,04	6926,72	6926,68	6926,82	6924,64
MarkovV2	8	2	34,33	40,45	40,46	40,35	40,62	40,37	40,60
MarkovV2	8	3	211,16	266,37	266,44	266,60	266,44	266,38	266,35
MarkovV2	8	4	1232,79	1467,32	1467,44	1468,33	1467,73	1467,65	1467,46



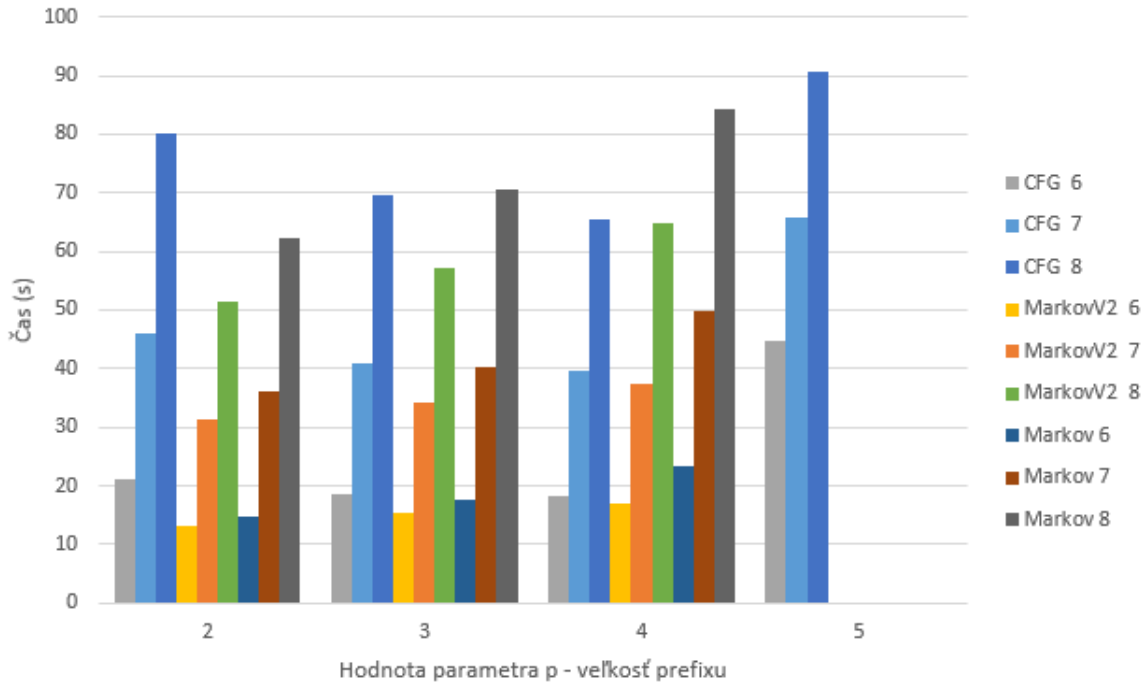
Obr. 6.2: Čas potrebný na vygenerovanie 500 000 hesiel podľa parametru p

ukázať výhody a slabiny jednotlivých algoritmov a spraviť ich vzájomne porovnanie v zmysle šancí na čo najrýchlejšie nájdenie hľadaného hesla.

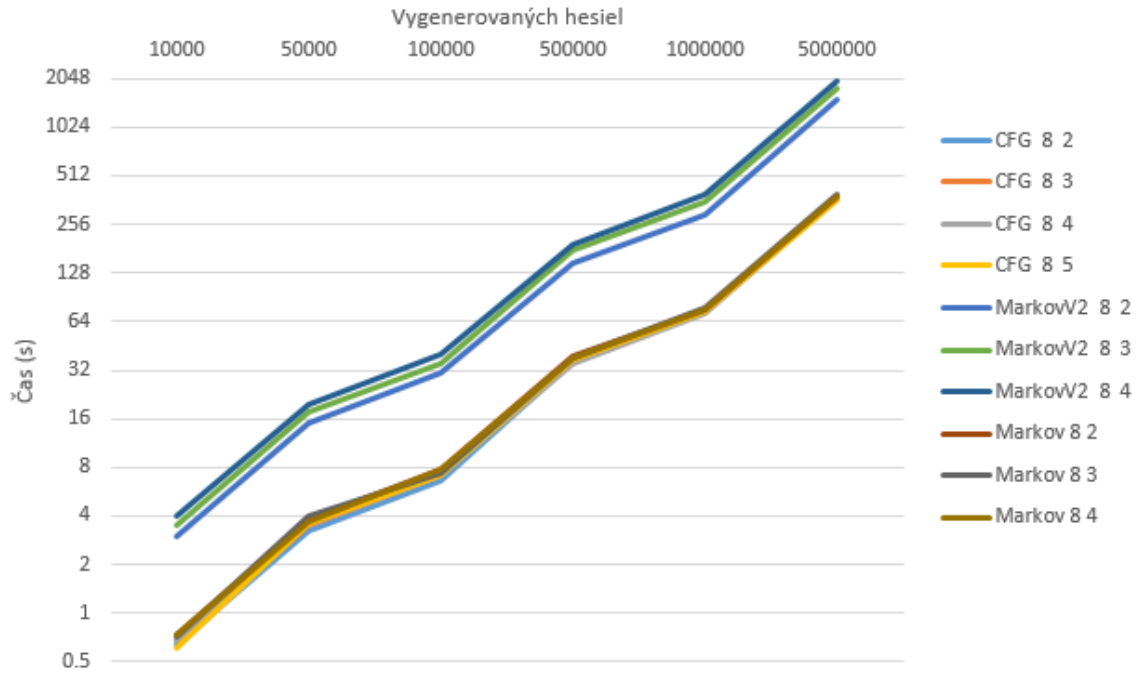
6.3.1 Heslá zo vstupného slovníka

Ako prvé sme skúmali koľko hesiel zo slovníka gramatika vygenerovala po vygenerovaní určitého počtu hesiel. Na grafoch, ktoré boli výstupom tohto testu sme na vodorovnej osi znázornili počet hesiel vygenerovaných gramatikou v desiatkach tisícov hesiel zatiaľ čo na vertikálnej osi je percento hesiel slovníka, ktoré sa medzi nimi nachádzajú. Pri tomto teste sme nechali oba algoritmy generovať 100 miliónov hesiel k čomu bol použitý slovník *rockyou-withcount* obsahujúci 13 331 008 rôznych hesiel dĺžky 12 a menej znakov.

Zatiaľ čo na obrázku 6.6 vidíme koľko unikátnych hesiel bolo vygenerovaných pomocou algoritmu využívajúceho Markovovské zdroje, nasledujúce grafy 6.7 a 6.8 ukazujú vyššie popísanú vlastnosť zaoberajúcu sa počtom vygenerovaných hesiel patriacich do vstupného slovníka.

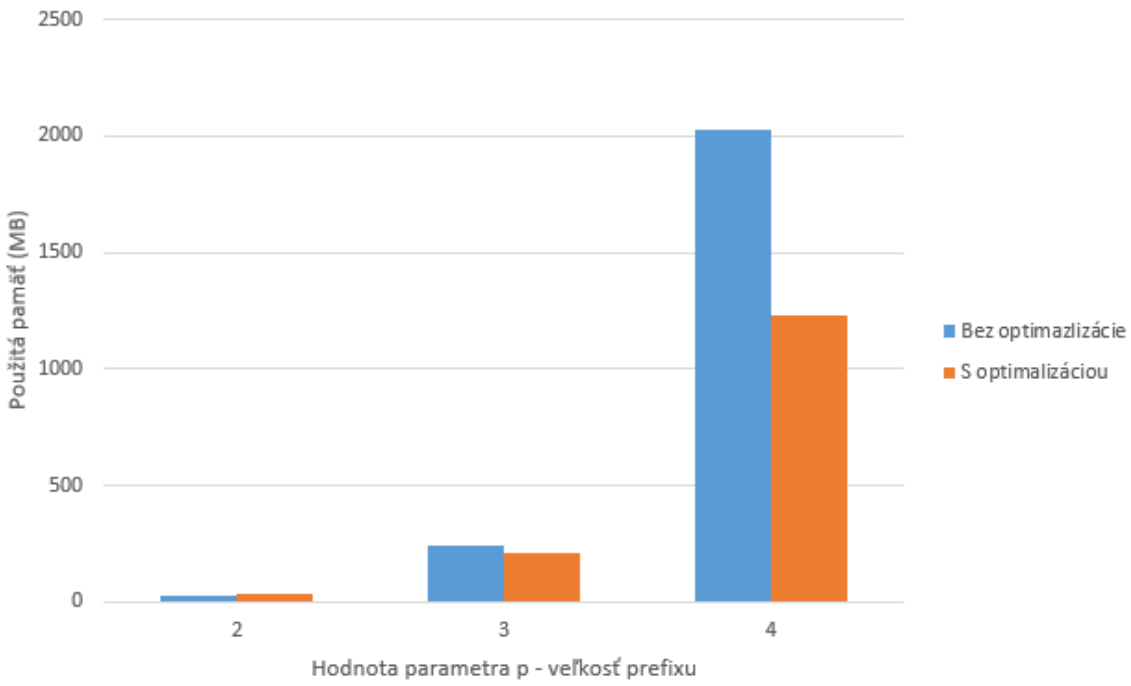


Obr. 6.3: Čas učenia pravdepodobností podľa parametru p



Obr. 6.4: Čas generovania hesiel

Heslá zo vstupného slovníka Na obrázku 6.7 vidíme priebeh hodnôt, kde heslá boli porovnávané so vstupným slonvíkom. Vidíme, že nami definované a implementované riešenie pomocou bezkontextových gramatík má na rozdiel od Markovovského zdroja omnoho pomalší rast počtu hesiel patriacich do slovníka. Pri 100 miliónoch generovaných hesiel to je niečo málo pod 700 tisíc.



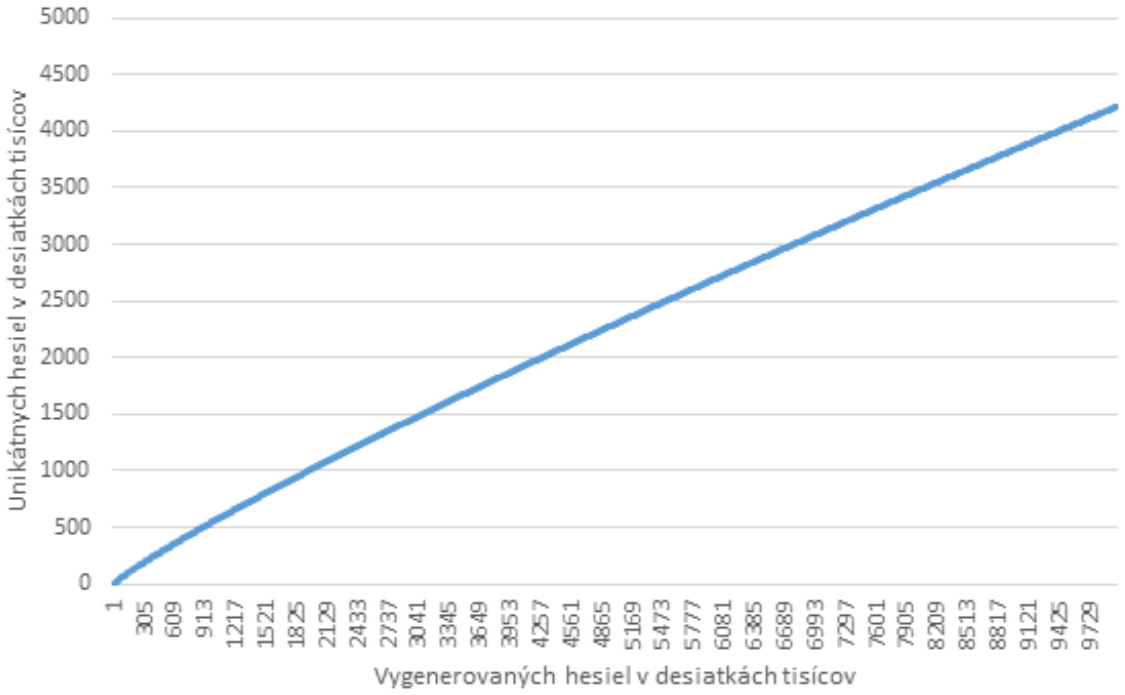
Obr. 6.5: Potrebná pamäť bez optimalizácie a s optimalizáciou

Heslá z nezávislého slovníka Graf 6.8 znázorňuje hodnoty po porovnaní vygenerovaných hesiel s iným nezávislým slovníkom. Pri tomto teste sme zobrali heslá vygenerované našimi algoritmami, ktoré na vstupe dostali slovník *rockyou-withcount*. Následne sme spravili testy počtu vygenerovaných hesiel, tentokrát avšak z iného ako vstupného slovníka. V tomto prípade sme použili slovník *phpbb*. Týmto sme chceli ukázať schopnosť nášho algoritmu vygenerovať slovník veľmi podobný tým používaným v praxi.

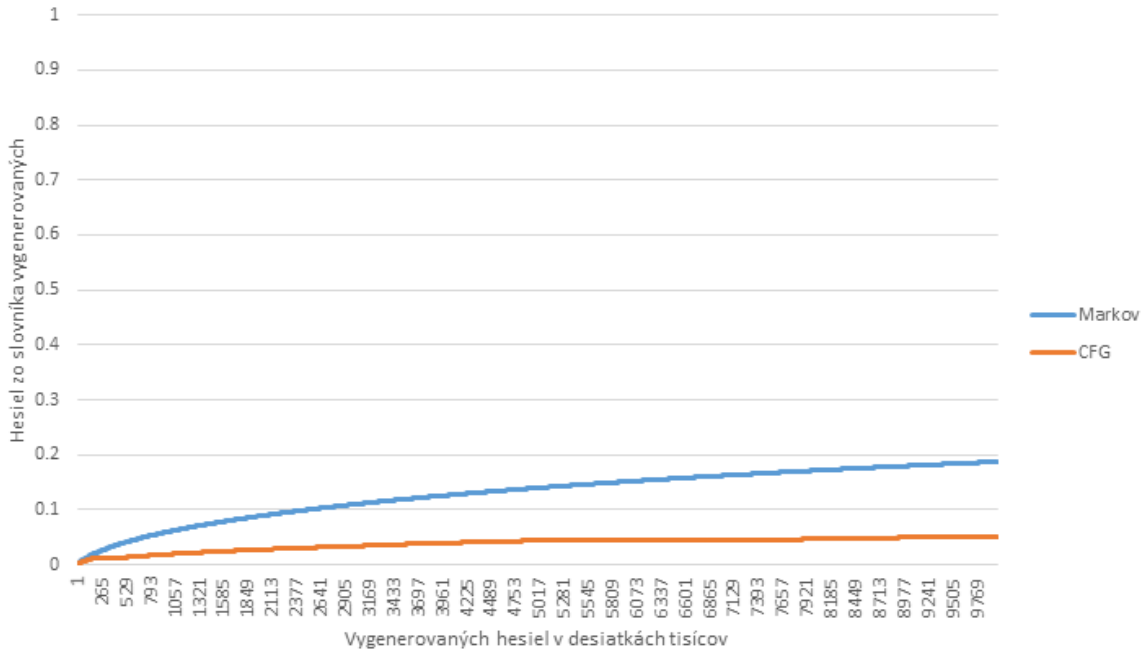
Ďalej sme taktiež skúmali ako sa správajú nami implementované algoritmy na menších dátach. Na obrázku 6.9 je znázornený graf priebehu generovania hesiel, ktoré sa nachádzajú vo vstupnom slovníku. Na vodorovnej osi je ukázaný počet vygenerovaných hesiel, v tomto prípade to bolo 5 miliónov hesiel. Výška čiar určuje množstvo hesiel, ktoré boli nájdene vo vstupnom slovníku. Pre Markovovské zdroje sa toto číslo počíta z počtu unikátnych hesiel, ktoré boli vygenerované. Taktiež si môžme všimnúť, že algoritmus používajúci Markovovské zdroje je v tomto teste opäť lepší ako algoritmus používajúci pravdepodobnostné bezkontextové gramatiky. Opäť sme použili slovník *phpbb-withcount* stiahnutý z nášho zdroja [2], ktorý sme upravili aby všetky heslá mali dĺžku najviac 6 znakov. Takto upravený slovník mal nakoniec 55 745 rôznych hesiel.

Graf 6.10 zobrazuje namerané percentá pre nami implementované algoritmy spustené tak, aby generovali heslá s dĺžkou 7. Ako vstupný slovník bol použitý *phpbb-withcount* upravený tak, aby obsahoval len heslá kratšie ako 8 znakov. Tento slovník obsahoval 88 416 unikátnych hesiel.

Nakoniec sme tento istý test opäť spustili na všetkých algoritmoch. Tentokrát

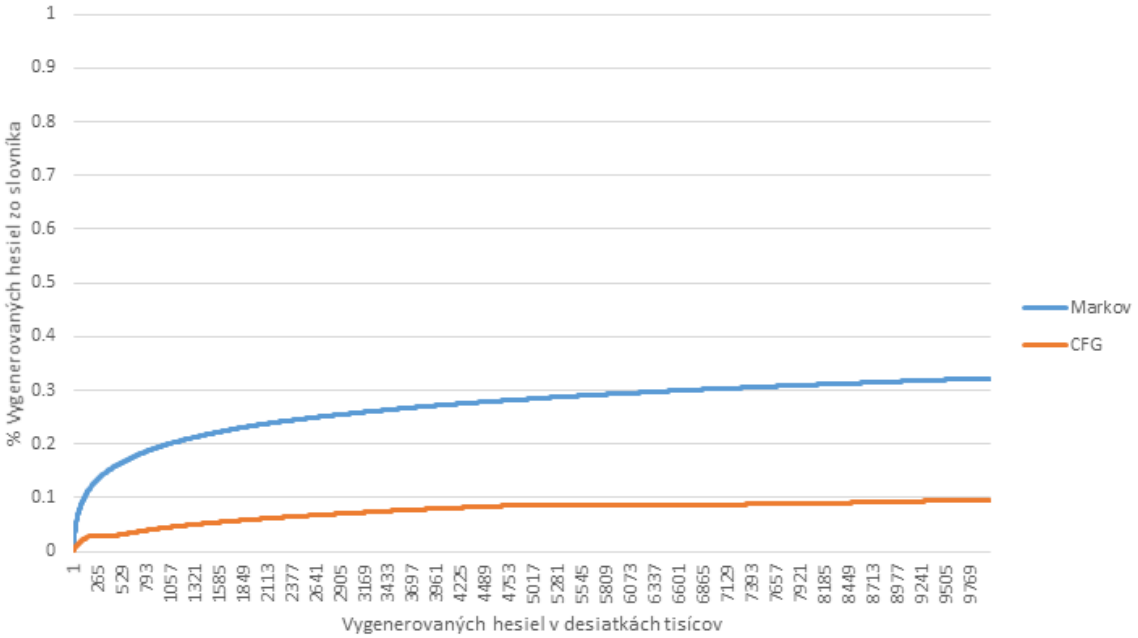


Obr. 6.6: Počet unikátnych hesiel



Obr. 6.7: Pomer vygenerovaných hesiel zo vstupného slovníka

vstupné parametre a slovník boli nastavené na generovanie hesiel maximálnej dĺžky 8. Takto upravený slonvík *phpbb-withcount* obsahoval 143 675 hesiel. V grafe 6.11 zobrazujúcom tieto dáta sme zobrazili výsledky tohto testu aj pre nami upravenú verziu Markovovských zdrojov, ktorá by mala byť schopná v konečnom čase vygenerovať heslá z celého priestoru možností. Označili sme ju ako *M2-8-4* keďže sa jedná o druhú verziu Markovovských zdrojov použitých v tejto práci.



Obr. 6.8: Pomer vygenerovaných hesiel z nezávislého slovníka

V prípade algoritmu, ktorý používa nami upravený Markovovský zdroj pri generovaní hesiel sme taktiež testovali vplyv nastavenia konštánt δ a ε. V nasledujúcom grafe 6.12 zobrazujeme rozdiel v počte vygenerovaných hesiel zo vstupného slovníka pre jednotlivé parametre označené v legende grafu ako $\delta - \varepsilon$. Vidíme, že tieto parametre nemajú žiaden vplyv na prvých 5 miliónov hesiel vygenerovaných týmto algoritmom.

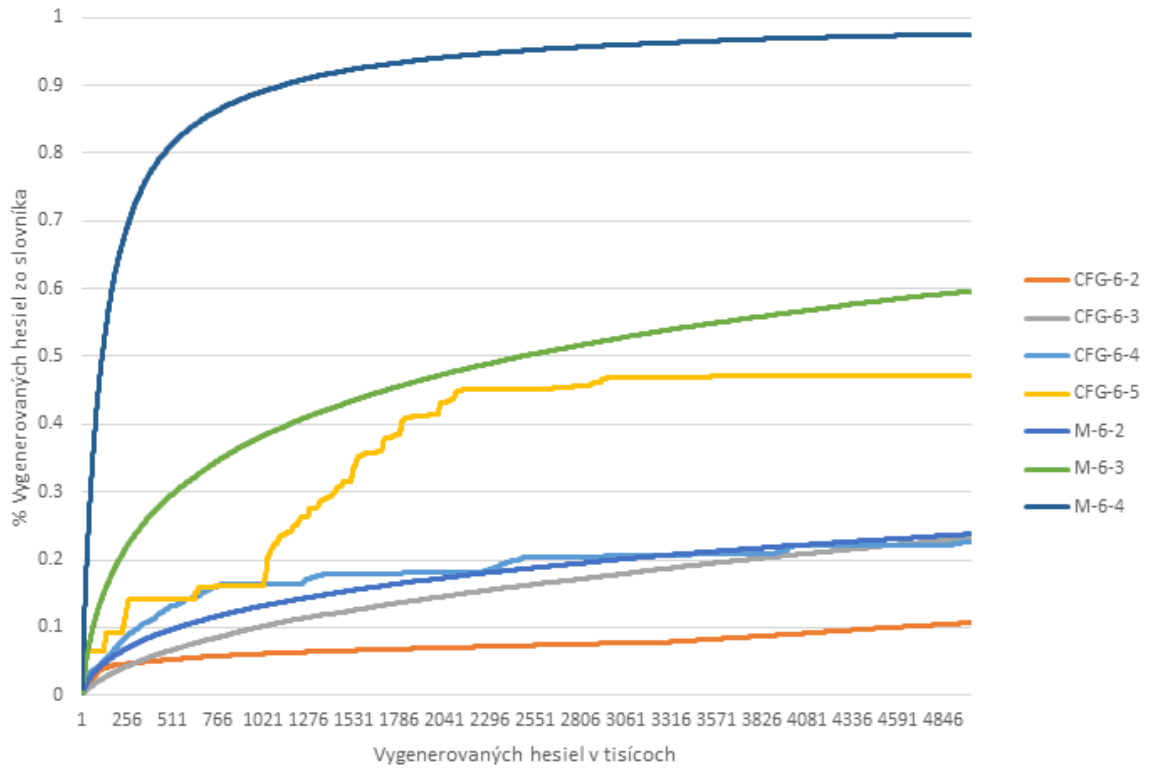
Na základe 6.7 a 6.8 vidíme, že nami navrhnutá metóda pomocou bezkontextových gramatík negeneruje veľa hesiel zo vstupného slovníka počas prvých miliónov vygenerovaných hesiel. Tento problém by sa dal vyriešiť tým, že by sme vždy ako prvé na výstup poslali všetky heslá zo slovníka, keďže ten býva zanedbateľne malý oproti veľkosti priestoru hesiel, ktorý musíme prehľadať aby sme definitívne našli hľadané heslo.

6.3.2 Miery presnosti

Pod týmto pojmom rozumieme metriky popisujúce nielen kvantitu nami generovaných hesiel patriacich do slovníka, ale snažia sa bližšie vyhodnotiť ako rýchlo sa gramatika dostane k heslám, ktoré boli podľa vstupného slovníka označené za najpravdepodobnejšie.

Stĺpce tabuľky 6.3 vyjadrujú hodnoty jednotlivých metrík pre daný algoritmus.

- *PPS* - Priemerná Pozícia v Slovníku - Vyjadruje priemernú pozíciu vo vstupnom slovníku pre heslá, ktoré boli vygenerované algoritmom na výstupe



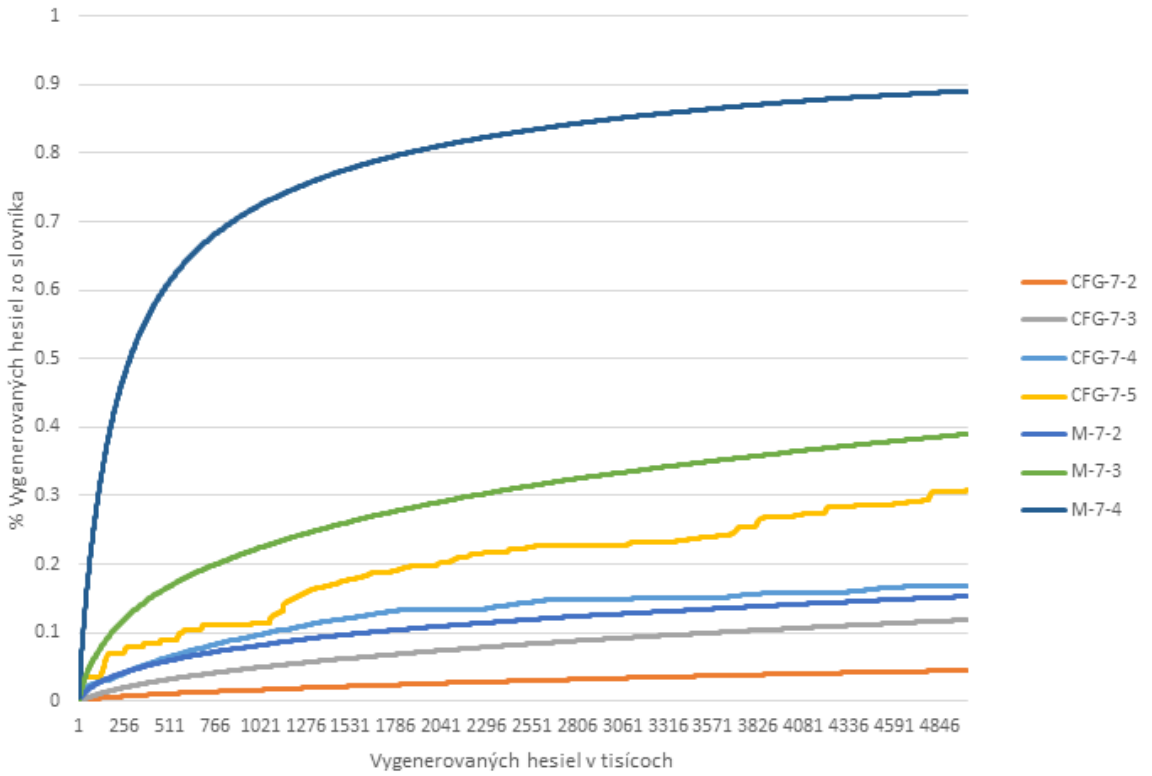
Obr. 6.9: Počet vygenerovaných hesiel zo slovníka - dĺžka 6

- *PPS %* - Priemerná Pozícia v Slovníku - Vyjadruje percentuálnu pozíciu vrámci slovníku pre heslá, ktoré boli vygenerované algoritmom na výstupe
- *RPSV* - Rozdiel Pozície v Slovníku a na Výstupe - Rozdiel v pozícii na vstupe a na výstupe algoritmu prenásobený percentuálnym počtom výskytov vo vstupnom slovníku
- *OPSV* - Odchýlka Pozície v Slovníku a na Výstupe - Absolútna hodnota rozdielu v pozícii na vstupe a na výstupe algoritmu prenásobená percentuálnym počtom výskytov vo vstupnom slovníku

$$\frac{\sum_{i=1}^{k}((indG_i - indS_x) * \frac{v_{ind_x}}{\sum_{j=1}^{n} v_j})}{k}$$

Vzorec vyjadrujúci mieru RPSV, kde n vyjadruje počet hesiel vo vstupnom slovníku a k je počet výskytov hesiel zo vstupného slovníka medzi generovanými. Hodnota $indG_i$ určuje poradie i-tého hesla vrámci generovaného slovníka. Hodnota $indS_i$ vyjadruje tú istú hodnotu pre vstupný slovník. Hodnoty v_i sú počty výskytov hesiel zadefinované vo vstupnom slovníku.

Vzorec pre mieru OPSV je takmer identický s vyššie uvedeným vzorcom, jediný rozdiel je v absolútnej hodnote rozdielu medzi pozíciami na vstupe a na výstupe.

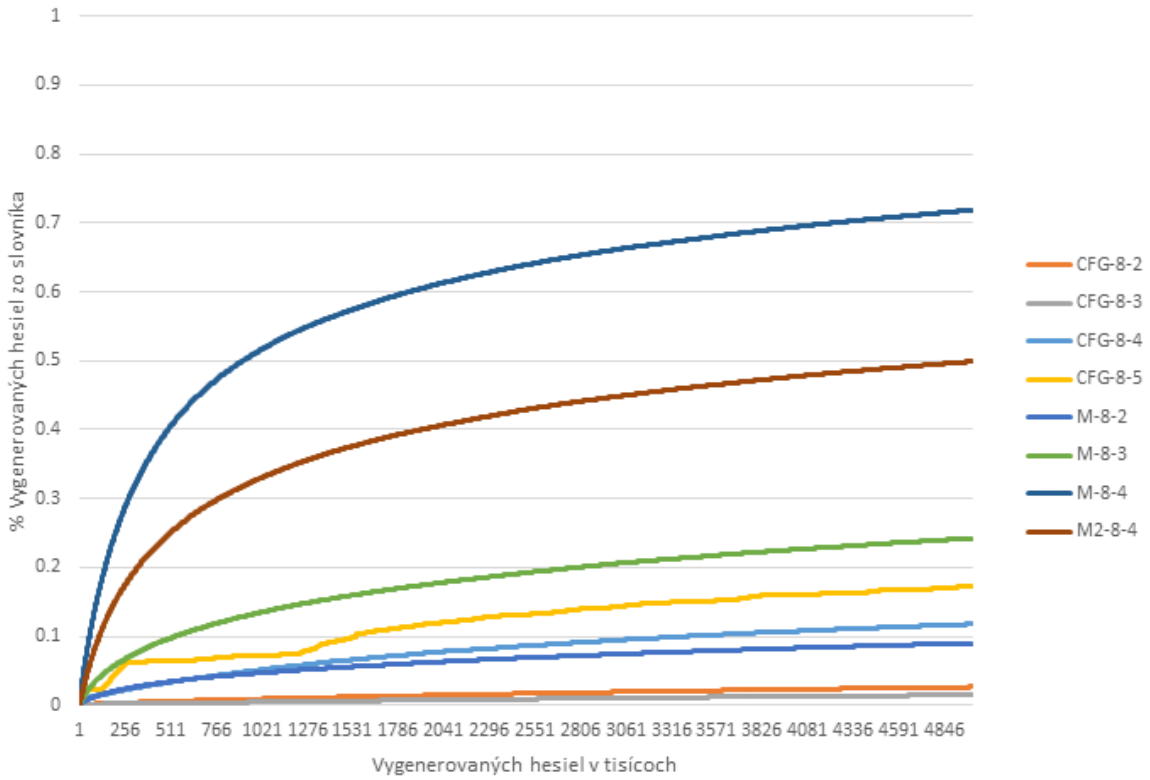


Obr. 6.10: Počet vygenerovaných hesiel zo slovníka - dĺžka 7

$$\frac{\sum_{i=1}^{k}(|indG_i - indS_x| * \frac{v_{ind_x}}{\sum_{j=1}^{n} v_j})}{k}$$

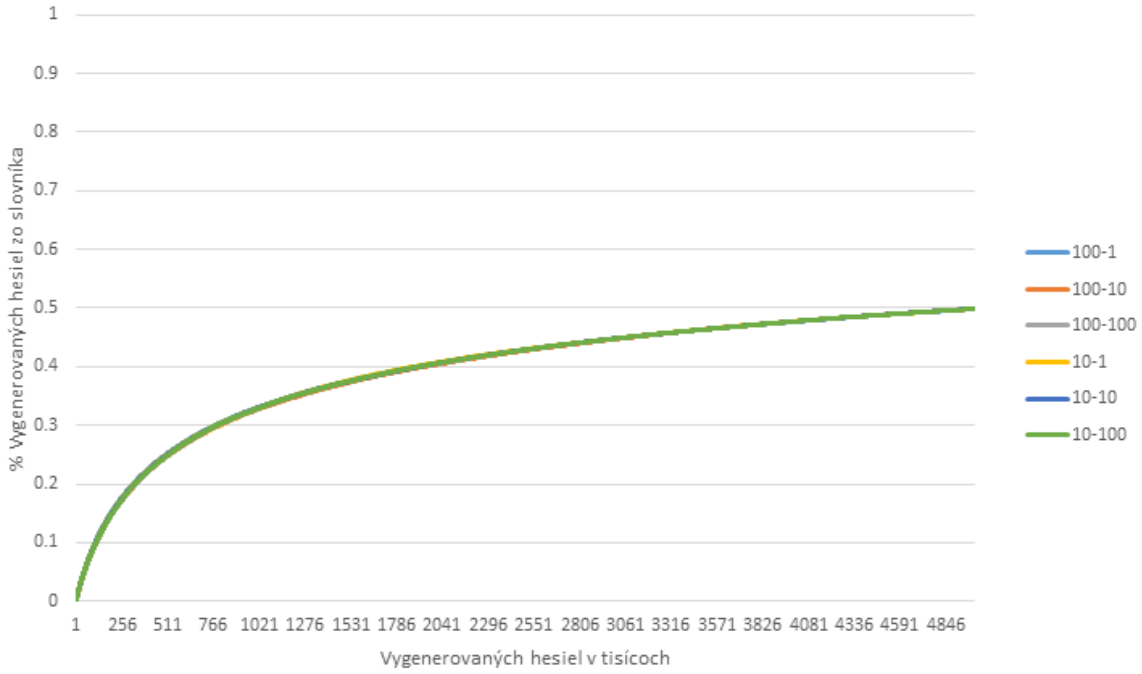
Priemerná pozícia v slovníku ukazuje priemernú pozíciu vygenerovaných hesiel vo vstupnom slovníku. Keďže toto číslo je závislé od veľkosti vstupného slovníka, prikladáme k nemu v druhom stĺpci jeho percentuálnu hodnotu. Hodnoty d a p vyjadrujú maximálnu dĺžku hesiel v slovníku a dĺžku použitých prefixov v Markovovskom zdroji. Pre hodnoty 6, 7, 8 parametru *d* sme použili slovník *phppbb* upravený na heslá relevantnej dĺžky. Pre hodnoty *d* rovné 12 sme použili omnoho robustnejší slovník *rockyou*, skladajúci sa z takmer 14 miliónov unikátnych hesiel.

Z tabuľky 6.3 môžme vidieť, že obom našim algoritmom prospieva navýšenie vstupnej informácie o heslách. Toto je vidieť ako na percentuálnych hodnotách priemernej pozície v slovníku, tak aj na rozdieloch pozícií medzi vstupným a vygenerovaným slovníkom. Hodnota rozdielov pozícií má vyjadrovať presnosť generovania hesiel v správnom poradí kedy algoritmus je odmenený znížením skóre ak sa mu podarí vygenerovať niektoré heslo skôr, než sa nachádza vo vstupnom slovníku. Posledná miera *odchýlka pozície v slovníku* má vyjadrovať absolútny rozdiel pozícií oproti vstupnému slovníku bez ohľadu na to, či heslo bolo vygenerované skôr alebo neskôr ako vo vstupnom slov-



Obr. 6.11: Počet vygenerovaných hesiel zo slovníka - dĺžka 8

níku. Malý rozdiel týchto hodnôt naznačuje, že väčšina hesiel, ktoré boli generované našimi algoritmami bola vygenerovaná neskôr ako bol ich výskyt vo vstupnom slovníku.



Obr. 6.12: Počet vygenerovaných hesiel zo slovníka pre rôzne hodnoty δ a ε - dĺžka 8

Tabuľka 6.3: Miery presnosti

	d	p	PPS	PPS %	RPSV	OPSV
CFG	6	2	31328	56,1	36,567	36,671
CFG	6	3	26358	47,2	37,749	37,762
CFG	6	4	31033	55,6	19,673	19,725
CFG	6	5	28554	51,2	21,397	21,484
CFG	7	2	47454	53,6	26,489	26,531
CFG	7	3	40343	45,6	25,831	25,848
CFG	7	4	46739	52,8	16,648	16,700
CFG	7	5	46718	52,8	21,715	21,804
CFG	8	2	88521	61,6	16,139	16,199
CFG	8	3	61113	42,5	22,384	22,436
CFG	8	4	70500	49,0	15,028	15,071
CFG	8	5	75570	52,5	14,749	14,865
Markov	6	2	25219	45,2	28,307	28,332
Markov	6	3	25959	46,5	14,815	14,840
Markov	6	4	27705	49,7	3,649	3,706
Markov	7	2	39090	44,2	21,289	21,323
Markov	7	3	40283	45,5	13,321	13,353
Markov	7	4	43232	48,8	4,700	4,757
Markov	8	2	61688	42,9	15,745	15,792
Markov	8	3	66401	46,2	9,701	9,748
Markov	8	4	68291	47,5	4,596	4,658
CFG	12	4	3781788	28,3	5,879	5,925
Markov	12	4	4609712	34,5	2,191	2,224

KAPITOLA 7

Záver

V tejto práci sme sa zaoberali útokmi hrubou silou na program TrueCrypt. Podrobne sme si naštudovali fungovanie tohto programu a usúdili, že najlepším miestom útoku, na disk zašifrovaný týmto programom, bude používateľské heslo pomocou ktorého sa generujú šifrovacie kľúče. Na základe tohto poznatku sme si naštudovali možnosti útokov hrubou silou, ktoré by sme mohli pri riešení nášho problému použiť.

Pri implementácií nášho riešenia používajúceho bezkontextové gramatiky sme používali znalosti o používateľských heslách k nastaveniu nášho algoritmu tak, aby optimalizoval poradie generovaných hesiel podľa pravdepodobnosti ich správnosti. Tento algoritmus sme následne porovnávali so zaužívanou metódou Markovovských zdrojov. Keďže nami navrhnutý algoritmus využívajúci bezkontextové gramatiky spĺňal konkrétnejšie podmienky pri generovaní hesiel (ako sú determinizmus, generovanie celého priestoru hesiel a vyhnutie sa duplikátom), navrhli sme zmeny v modeli Markovovských zdrojov. Tieto zmeny sme taktiež implementovali a výsledný algoritmus sme taktiež porovnali s našim riešením.

Nami navrhnutá metóda využívajúca bezkontextové gramatiky dopadla v testoch veľmi podobne ako Markovovské zdroje, čo hodnotíme ako pozitívny výsledok tejto práce. Zdrojové kódy všetkých nami implementovaných algoritmov sú voľne dostupné na `https://github.com/Wryxo/Diplomovka`

7.1 Možnosti zlepšenia nášho riešenia

7.1.1 Izolovanie kódu na skúšanie kandidátov

Podarilo sa nám nájsť časti kódu overujúce správnosť hesla. Bohužiaľ z dôvodu časovej tiesne sa nám nepodarilo izolovať kód programu TrueCrypt, ktorý by bez použitia samotného TrueCryptu overoval správnosť používateľom zadaného hesla. Nájdenie

tohto kódu by mohlo pomôcť pri implementácií rýchleho algoritmu na overovanie nami generovaných kandidátov. Tento nedostatok sa dá však nahradiť použitím niektorého z voľne dostupných programov určených na útoky hrubou silou, ktorému ako vstupný slovník dodáme slovník vygenerovaný nami implementovaným programom.

7.1.2 Veľkosť potrebnej pamäte

Najväčším nedostatkom samotného algoritmu využívajúceho bezkontextové gramatiky je množstvo pamäte potrebné na jeho beh. To dovoľuje použiť tento algoritmus len v prostredí s obrovským množstvom operačnej pamäte. Odstránenie tohto nedostatku vyžaduje ďalší vývoj algoritmu tak, hlavne v oblasti optimalizácie použitých dátových štruktúr a efektívnosti algoritmu na generovanie hesiel pomocou týchto gramatík.

7.1.3 Kompletne generujúci Markovovský zdroj

Podarilo sa nám implementovať Markovovský zdroj, ktorý v konečnom čase vygeneruje všetky možné heslá zo vstupnej abecedy. Pri riešení tohto problému sme zadefinovali konštanty δ a ε, ktoré slúžia na zadefinovanie prechodov medzi známymi a neznámymi stavmi. Možným vylepšením by bolo upravovanie týchto konštánt za behu programu podľa počtu hesiel, ktoré sme už vygenerovali, aby sme zvýšili šancu vygenerovania zbytku existujúcich hesiel. Toto však zahŕňa výskum, ktorý by sa dal rozobrať v samostatnej práci.

Literatúra

[1] Francesco Bergadano, Bruno Crispo, and Giancarlo Ruffo. High dictionary compression for proactive password checking. *ACM Trans. Inf. Syst. Secur.*, 1(1):3–25, November 1998.

[2] Ron Bowes. Slovníky s heslami - `https://wiki.skullsecurity.org/Passwords`, 2015.

[3] François Delebecque and Jean-Pierre Quadrat. Optimal control of Markov chains admitting strong and weak interactions. *Automatica*, 17(2):281–296, 1981.

[4] James Forshaw. Truecrypt 7 derived code/windows: Drive letter symbolic link creation EoP - `https://bugs.chromium.org/p/project-zero/issues/detail?id=538`, 2015.

[5] James Forshaw. Truecrypt 7 derived code/windows: Incorrect impersonation token handling EoP - `https://bugs.chromium.org/p/project-zero/issues/detail?id=537`, 2015.

[6] E. F. Gehringer. Choosing passwords: security and human factors. In *Technology and Society, 2002. (ISTAS'02). 2002 International Symposium on*, pages 369–373, 2002.

[7] Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. Password cracking and countermeasures in computer security: A survey. *CoRR*, abs/1411.7803, 2014.

[8] Cynthia Kuo, Sasha Romanosky, and Lorrie Faith Cranor. Human selection of mnemonic phrase-based passwords. In *Proceedings of the Second Symposium on Usable Privacy and Security*, SOUPS '06, pages 67–78, New York, NY, USA, 2006. ACM.

[9] David Malone and Kevin Maher. Investigating the distribution of password choices. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 301–310, New York, NY, USA, 2012. ACM.

[10] Simon Marechal. Advances in password cracking. *Journal in computer virology*, 4(1):73–81, 2008.

[11] T. Murakami, R. Kasahara, and T. Saito. An implementation and its evaluation of password cracking tool parallelized on gpgpu. In *Communications and Information Technologies (ISCIT), 2010 International Symposium on*, pages 534–538, Oct 2010.

[12] Eugene H Spafford. Observing reusable password choices. 1992.

[13] ZDNet. 25 gpus devour password hashes at up to 348 billion per second - `http://www.zdnet.com/article/25-gpus-devour-password-hashes-at-up-to-348-billion-per-second`, 2012.

[14] Moshe Zviran and William J. Haga. A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal*, 36(3):227–237, 1993.