SLOVENSKÁ TECHNICKÁ
UNIVERZITA V BRATISLAVE
STAVEBNÁ FAKULTA

ŠVOČ
Akademický rok 2014/2015

MATEMATICKÉ MODELOVANIE PRIRODZENÉHO JAZYKA

Meno a priezvisko študenta:	Peter Hornáček, 3. ročník
Študijný program:	Matematicko-počítačové modelovanie
Študijný odbor:	9.1.9 Aplikovaná matematika
Katedra/ústav:	Katedra matematiky a deskriptívnej geometrie
Vedúci práce:	Mgr. Jozef Kollár

Bratislava, 28. máj 2015

Abstrakt

Cieľom práce je naštudovať uvedenú problematiku a vytvoriť jednoduché modely prirodzeného jazyka. Modely prirodzeného jazyka majú široké uplatnenie pri spracovaní textu napr. pri webových službách, ďalej v kryptológii a majú súvislosť aj s teóriou kódovania. Pre konkrétne aplikácie je potrebné vybrať vždy najvhodnejší model podľa rôznorodých kritérií ako sú dátová a výpočtová náročnosť, rýchlosť spracovania a samozrejme účel použitia. Práca sa bude zaoberať skúmaním najjednoduchších modelov založených na n-gramou a slovách, ktoré sa používajú napr. na rozpoznávanie jazyka, detekciu zmysluplného textu a pri lúštení klasických šifier.

Kľúčové slová: modely prirodzeného jazyka, n-gramy

Abstract

The aim of this work is to study indicated natured language, modeling and implement simply models. Natural language models have wide applications in word processing, web services, mathematiced cryptology, or coding theory. For each application it is important to select the most appropriate model according to various criteria, including data or computationalg complexity, speed of processing, and intended use. This work is devoted in the most simple models based on n-grams and word dictionaries, which are used to recognize language, detect meaningful text and deciphering classical ciphers.

Keywords: models of natural language, n-grams

Obsah

1 Úvod

Modely prirodzeného jazyka majú v súčasnej dobe široké uplatnenie. Stretávame sa s nimi napr. pri webových službách, v kancelárskych aplikáciach, využívajú sa v kryptológii pri lúštení šifier a majú úzku súvislosť s teóriou kódovania. V rôznych aplikáciach sa využívajú rôzne modely. Najjednoduchšie a najčastejšie využívané sú n-gramové modely. napr. vo vyhľadávači Google sa (okrem iného) jazyk detekuje pomocou 3-gramového modelu jazyka. Potom sú to modely postavené na slovníkoch. Tie sa používajú napr. v už spomenutých kancelárskych aplikáciach. Niektoré komplikovanejšie modely sú kombinované a naviac môžu využívať aj niektoré menej obvyklé štatistiky jazykov ako sú napr. index koincidencie alebo entropia. Pri výbere vhodného modelu pre konkrétnu aplikáciu musia okrem samotného účelu použitia zohľadňovať aj kritéria akými sú dátová a výpočtová náročnosť, rýchlosť spracovania atď. Cieľom je vždy pre danú aplikáciu vybrať dostatočne účinný a pritom čo najjednoduchší model.
V tejto práci sa budeme zaoberať len n-gramovými a veľmi jednoduchými slovníkovými modelmi. Modely založené na n-gramoch sa dajú veľmi dobre využiť na detekovanie zmysluplného textu, na identifikáciu použitého jazyka a generovanie náhodného textu, ktorý má štatistické vlastnosti daného jazyka. Pre uvedené vlastnosti sú n-gramové modely vhodné na mnohoraké využitie v kryptológii, ale samozrejme používajú sa aj iných oblastiach (napr. už spomenutý vyhľadávač Google).
Napríklad pri kryptoanalýze klasických šifier je priestor kľúčov často tak malý, že sme v súčasnosti schopní lúštiť tieto šifry hrubou silou vyskúšaním všetkých možných kľúčov. Pri takomto lúštení však nestačí dešifrovať zašifrovaný text niekoľkými miliónmi možných kľúčov. Nastáva problém detekovania správneho (zmysluplného) textu, ktorý sa rieši vhodným modelom jazyka. V jednoduchších prípadoch, napr. pri jednoduchej zámene, je úplne postačujúci tetragramový model jazyka. Pri šifrách s periodickým heslom (napr. Vigenerova šifra) sa využíva aj index koincidencie jazyka.
V kryptografii sa modely jazyka môžu využívať na vytváranie šumu. Pod šumom tu rozumieme náhodný text, ktorý je (aspoň pre dané štatistické indexy) štatisticky nerozlíšiteľný od zmysluplného textu. Takýto šum sa v šifrách často využíva na oklamanie nepriateľa.
Pri navrhovaní vhodného modelu jazyka sú podstatné kritéria veľkosť modelu a časová náročnosť spracovania. Pre konkrétny jazyk by bol ideálny model korpus všetkých textov, ktoré sa v tomto jazyku môžu vyskytnúť. Toto však nie je prakticky realizovateľné, pretože sa jedná o teoreticky nekonečné množstvo dát, ktoré by nebolo možné ani uložiť a ani s nimi pracovať. Protipólom je najjednoduchší možný model, ktorým sú frekvencie znakov príslušného jazyka. Pre niektoré veľmi jednoduché aplikácie môže byť tento model postačujúci, ale pre väčšinu reálnych aplikácii je nedostatočný. Preto je potrebné hľadať kompromis medzi uvedenými dvoma extrémami. Tu sa objavuje súvislosť medzi teóriou kódovania a modelmi jazyka. Medzi prvými, ktorí túto súvislosť študovali a popísali, bol Claude Shannon v súvislosti so štúdiom entropie a redundancie prirodzeného jazyka.
V práci skonštruujeme niektoré jednoduchšie n-gramové a jeden slovníkový model vybraných jazykov. Následne preskúmame štatistické vlastnosti týchto modelov a vhodnosť ich použitia pre niektoré aplikácie.

2 Príprava

2.1 Vytvorenie zdrojov

Budeme sa zaoberať troma jazykmi: Slovenčina, Španielčina, Angličtina. Zdroje sme vytvorili zozbieraním rôznych kníh z jednotlivých jazykov. Počet znakov pre jednotlivé jazyky je približne 3 000 000[1]. Pracovať budeme s 26 znakovou abecedou od A-Z ktorá sa obvykle nazýva aj telegrafná abeceda (označenie TSA). S touto abecedou budeme pracovať z niekoľkých dôvodov: kvôli tomu, že v kryptografii sú písmena s diakritikou veľmi ľahko odhaliteľné, pretože majú veľmi špecifickú frekvenciu. Ďalší dôvod je technický, lebo napr. v morzeovke sa využíva dvadsaťšesť znaková abeceda.

2.2 Príprava zdrojov

Z textu sme odstránili všetku diakritiku, medzery budeme odstraňovať pri n-gramoch, kde $n \leq 3$ lebo pri nich nezohrávajú podstatnú úlohu, riadky a ešte sme všetky malé písmena zmenili na veľké. Konkrétne vidieť zmenu diakritiky pre Slovenčinu a Španielčinu v príslušných tabuľkách.

Zmena diakritiky v Slovenčine

s diakritikou	á	é	í	ó	ú	ý	Í	ŕ	č	ď	ľ	ň	š	ť	ž	ä	ô	ő	ö
bez diakritiky	a	e	i	o	u	y	l	r	c	d	l	n	s	t	z	a	o	o	o

Zmena diakritiky v Španielčine

s diakritikou	ñ	á	é	ć	ë	ó	ń	ú	ň	ŕ	í	ě	ô	ť	ď	ä	ů	â	ç	ş
bez diakritiky	n	a	e	c	e	o	n	u	n	r	i	e	o	a	d	a	u	a	c	s

Takto nám z textu zostal len jeden riadok, ktorý obsahuje iba veľké písmená TSA abecedy.

3 n-gramy

Zadefinujeme si n-gram ako sled n po sebe idúcich znakov z telegrafnej abecedy a pre $n > 3$ ju rozšírime o medzeru. Príkladmi n-gramov sú AA,YGR, H LK. Čiže n-gramy sú prvky množiny $\mathbb{N}^n$ kde $\mathbb{N} = \{A, B, C, ...\}$ je naša abeceda. Nasledujúce n-gramy budeme ďalej označovať takto: 1-gram = znak, 2-gram=bigram, 3-gram=trigram a 4-gram=tetragram. Pre väčšie čísla to už budú n-gramy (kde n je číslo > 4)

3.1 Frekvencie n-gramov

V skutočnosti budeme zisťovať relatívnu početnosť, ale pre lepšiu zrozumiteľnosť to budeme nazývať frekvenciou. Frekvenciu jednotlivých n-gramov sme získavali nasledovným spôsobom: Prechádzali sme pripravené zdroje n-gram po n-grame. Postupne sme takto narátali koľko krát sa ktorý n-gram vyskytuje v danom zdroji. Frekvenciu f, n-gramu získame vzťahom

$$f = \frac{p}{m}$$

kde p je počet výskytov daného n-gramu v zdroji a m je celkový počet n-gramov v texte. Následne sme frekvencie normovali na interval $[0, 1]$ a zapísali do súboru. Frekvencie

[1] Zoznam kníh je v prílohe

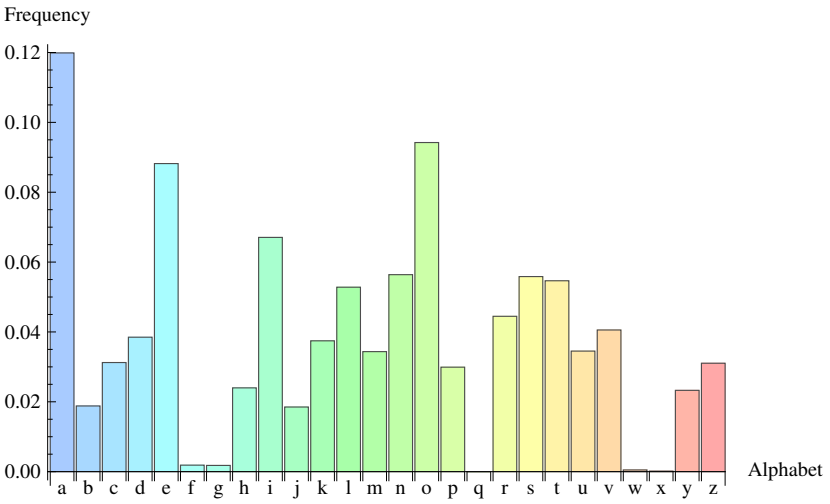
sme normovali kvôli prípadným zaokrúhľovacím chybám. Aby sme mohli generovať n-gramy na základe takto získaných frekvencií. Frekvencie jednotlivých n-gramov z každého jazyka je možné vidieť v tabuľkách frekvencií. Pre znaky sme ich frekvencie zobrazili v grafoch frekvencií ktoré sme vytvorili v programe Mathematica. Pomocou príkazu

```
Import[názov_súboru.txt]
```

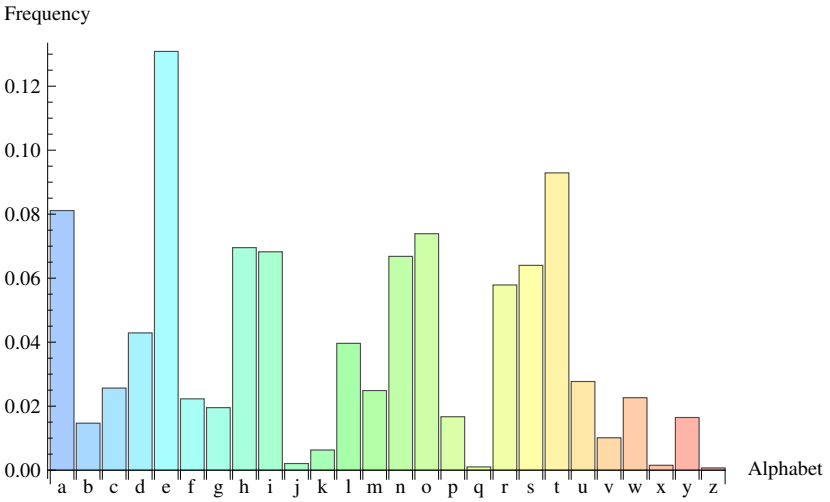
sme načítali frekvencie a potom na vykreslenie grafu sa použil príkaz

```
BarChart[]
```

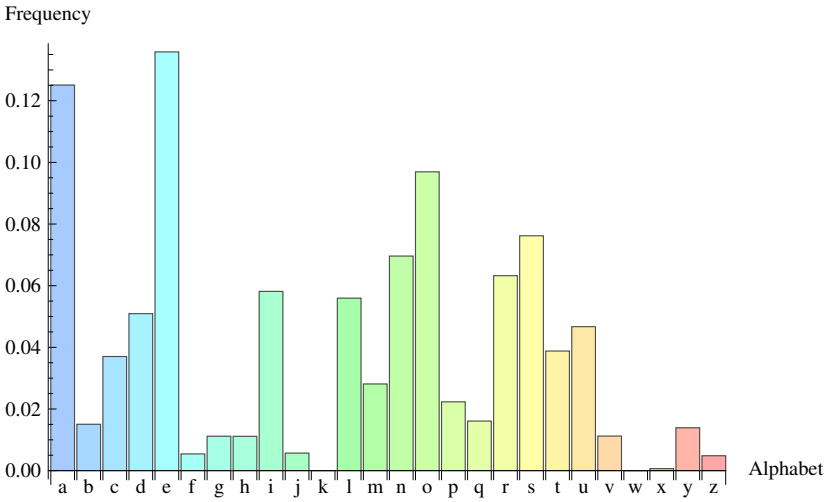
V tabuľke 1 na strane 14 sú frekvencie znakov zoradené od najvyššej po najnižšiu. V tabuľke 12 na strane 18 uvádzame koľko najmenej n-gramov v danom jazyku treba použiť aby súčet ich frekvencií bol aspoň 90 %.



Obr. 1: Frekvencie slovenských znakov



Obr. 2: Frekvencie anglických znakov



Obr. 3: Frekvencie španielskych znakov

3.2 Generovanie n-gramov

Na generovanie n-gramov potrebujeme taký generátor pseudonáhodných čísel (Pseudo-Random Number Generator = PRNG), ktorý generuje čísla čo najrovnomernejšie. Budeme porovnávať dva generátory: `rand()`, `drand48()`. Pre tieto generátory sme sa rozhodli lebo sú najznámejšie a sú v štandardnej knižnici programovacieho jazyka C.

3.2.1 PRNG

PRNG sme overovali nasledujúcim spôsobom. Pomocou `rand()` a `drand48()` sme si vygenerovali 10 000, 100 000, 3 000 000 čísel a následne v programe Mathematica sme použili tieto príkazy:

```
data = Import[názov_súboru.txt]
```

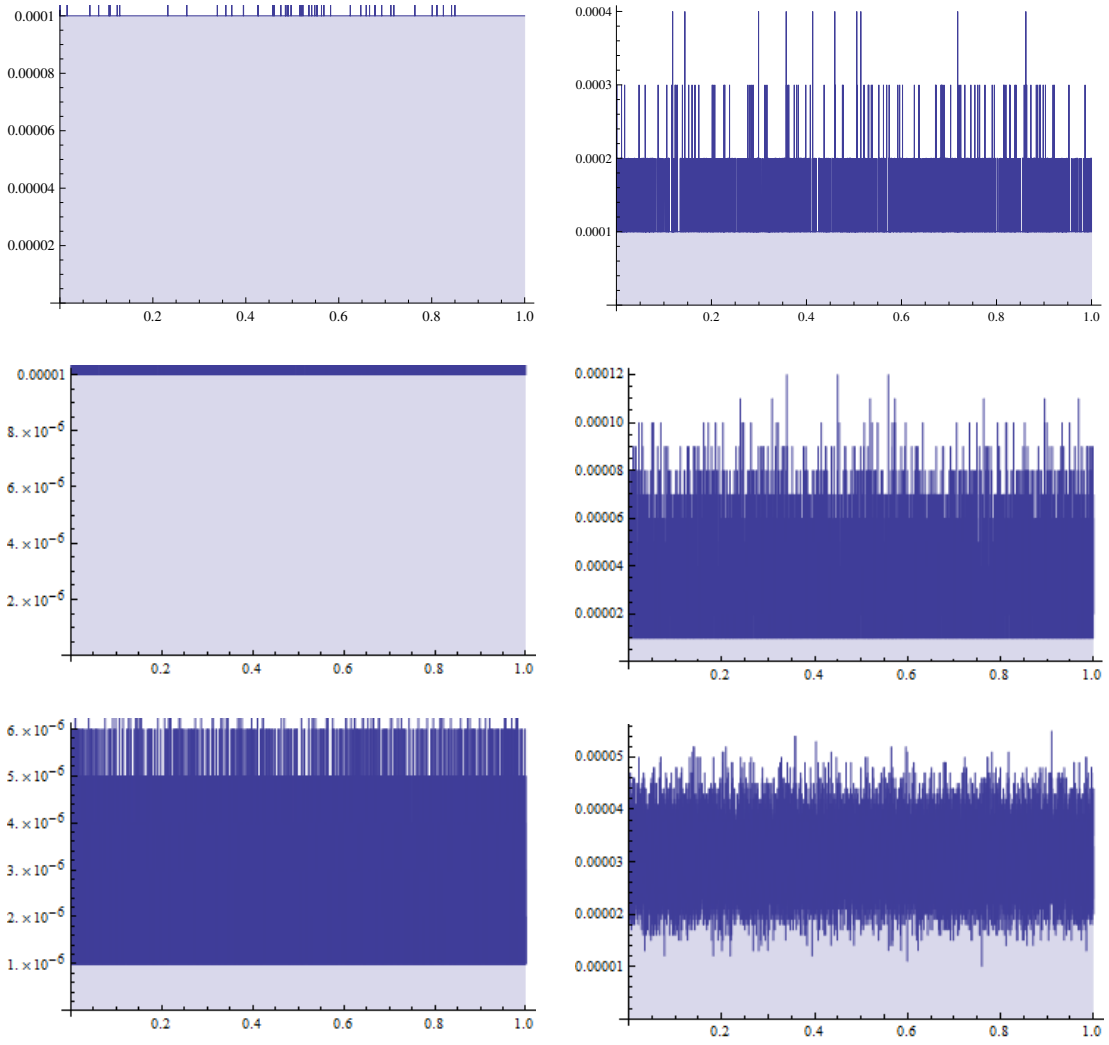
na načítanie číselných dát z .txt súboru do premennej `data`, príkaz

```
dataED = EmpiricalDistribution[data]
```

slúžiaci na zistenie empirickej distribúcie, ďalej

```
DiscretePlot[PDF[data2ED, x], x, data2, AxesOrigin → {0, 0},
```

na vykreslenie grafu, na ktorom je vidieť distribúcia dát a tým vieme určiť ich rovnomernosť. Na obrázku 4 na strane 5 je vidieť že generátor `drand48()` generuje dáta pre naše potreby lepšie než `rand()`.



Obr. 4: Testovanie PRNG –
os x →rozloženie hodnôt na intervale $[0, 1]$
os y →počet pokusov

3.3 Spôsob generovania

Generovali sme n-gramy tak, že interval $[0, 1[$ sme rozdelili na 26^n podintervalov. Veľkosť každého podintervalu sme určili na základe získaných frekvencií. Uvedieme príklad na znakoch.
Príklad: ak by sme mali len tri 3 znaky A,B,C s frekvenciami f_A, f_B, f_C. Následne by sme frekvencie normovali kvôli prípadným zaokrúhľovacím chybám $d_1 = \frac{f_A}{\sum f}, d_2 = \frac{f_B}{\sum f}$, tak podinterval pre A by bol $[0, d_1]$, pre B by bol $[d_1, d_1 + d_2]$, pre C by bol $[d_1 + d_2, 1]$. Potom sa pomocou `drand48()` vygeneruje náhodné číslo z intervalu $[0, 1[$ a pomocou algoritmu sa zistí kam vygenerované číslo padne, podľa čoho sa určí príslušný n-gram. Následne si môžete pozrieť ukážku algoritmu pre znaky naprogramovaného v C (generoval som 180 písmen).

```
srand48(time(NULL));

for(i=0; i<generuj_pismen; i++)
{
        pravdepodobnost = drand48();
        pomsucet = 0;
        pismeno = 0;

        while(pravdepodobnost >= pomsucet)
        {
                pomsucet += intenzity[pismeno];
                pismeno++;
        }
        pismeno--;

        fprintf(fi,"%c", pismeno+65);
        if((i+1)%60 == 0)
                fprintf(fi, "\n");
}
```

3.4 Modely

Modely sme aj implementovali vo forme počítačového programu a toto sú výsledky z nich. Každý model bol spustený na vygenerovanie 10 000 prislúchajúcich n-gramov, pričom nasledujúce ukážky sú tvorené zo 180 n-gramov. V ukážkach sme farebne vyznačili niektoré slová, ktoré sú dĺžky 3 a viac. Robili sme porovnanie frekvencií pomocou Kolmogorovho-Smirnovho testu.
Kolmogorov-Smirnovov test je štatistický test hypotézy

$$H0: \ F_e = F$$

oproti alternatívnej hypotéze

$$H1: \ F_e \neq F$$

kde F_e je empirická distribučná funkcia napočítaná z vygenerovaných dát a F je predpokladaná teoretická distribučná funkcia. Testovacia štatistika je založená na rozdiele

týchto dvoch distribučných funkcií, t.j.

$$\mathrm{TS} = \sup_{x \in \mathbb{R}} |F(x) - F_e(x)|$$

a kritická hodnota sa určuje z Kolmogorovho rozdelenia. Pre dostatočne veľké $n > 35$ je možné urobiť aproximáciu pomocou vzorca

$$\mathrm{KH}(\alpha) = \sqrt{\frac{\ln(\frac{2}{\alpha})}{2n}}$$

kde α je hladina významnosti a n je počet vygenerovaných dát. Robili sme ho na hladine významnosti $\alpha = 0.05$ a $n = 10\,000$. V teste sme porovnávali napočítané frekvencie z normálneho jazyka a frekvencie napočítané z modelov jazyka.

3.4.1 Vygenerované znaky

Výsledky Kolmogorovho-Smirnovho testu pre znaky sú dané v tabuľke 5 na strane 17. Z výsledkov vidíme, že znakový model sa zo štatistického hľadiska vôbec nepodobá reálnemu textu keďže frekvencie sa zhodujú iba pri znakoch.

Slovenčina

ARENALHRRAUKYONJVOLASELEZOETIAEAYTSSDVNCVTODHASYANEEMIRAMELI
ETDAVAKSKOKMRKSHOD OKTJONZEOIYKADCZAEHBDSNAELZYEOOBIINANVHEUS
UAAOOYPDLODZIASLOMEAVTSIMDRLLSHENLSKKAUEAOAOSEOOMEUAEAIKSIJN

Španielčina

QALOQADTNDICELADBESNATJUINSAHPEOISMENIEUEANOEECAYIALIELTMLTO
EEIELREESRNLLSIRAOBIOROMIICQOYNUBDNPCRLRFAGJSVIMLARAAPEFOOEI
UNRCENAIEEADSOOIIOSOMNTODEHDRSDODTSSIASONTNVHYJADNOAEEONENII

Angličtina

TTMGEEAMTTSLNAELEOESIRXTEARHTYORAIODACCOHSEELRMOHIYNAIREAROI
ETSEREOTDCSILTMRXJSRDDSEMAANRTOGHESEEOVCBDHGEHAAFUEETNAEONNL
EIEMRDDEHOBEEPHNSEIITOIECPSOAPAOSOCENHRENECSHWIBASENASTSASRE

3.4.2 Vygenerované bigramy

Výsledky Kolmogorovho-Smirnovho testu pre bigramy sú dané v tabuľke 6 na strane 17. Na základe výsledkov z testu vidíme, že modely bigramov bez pamäti a s pamäťou sú lepšie ako znakový model. Keď sa pozrieme na počet vzniknutých slov tak je vidieť, že tieto modely nie sú úplne ideálne a nepodobajú sa reálnemu textu.

Slovenčina

CETTESSAICINALBIOZOVAUUTRIPOLWDEACYMAPORHNBAADOPLILAPACIOVNO
INISTDRTOMAZLBAMTIVILIODOVCEBOERASPITAZAODTMSTABETTAOTOGKALA
IEAJDOOUTRISLACAKOOTNEUAYSANTOVAUPAALIATATYMABAVANPOYCJUUPTV
CIOSIBJUNUCONIZUCORORVUJJDPRCIEMIHHTTOOKPOMLTAEMLNANLKEBDIAT
HOZLVAAVTUETATKONEAKSIEMRNAMMUASINLKJOVSOOMATKVOELINSIRIELEJ

EBLPIZUPCHNEYFULIDVARHARSEISLIDODUNEDBENSKPOOSPRAHEMUZYVDANA

Španielčina
BUIANDSOGASAIOMEMIRTELSIANLTRDANTONOULFAANCONTLSMUOEVIALAPUE
TASTORUDANONANDEMIJOIOVAANPAELQUOSSESGROSESYCOOSERIGSIOHDETA
OZTEOSSAOEDIDEANNOBAAQNOUROLENNRORSOORSTOQUYRIDUNPSCIOIEECES
ERSLESOTENGUNTRENCISCUYEEFPOUENFSTSOIMANRECOQUESESALOLTIDOST
SDPRQUASOYMAINOIORNENOQUDEALLASEPUUENILRRAMAHASQTONOSIONONCO
EAESDOCENCLTLLIEAURELGUELONLYELD INADDOENUADOS UNDCHDUNRPAEY

Angličtina
DJSLSTNECHSOEDARREILARTTTFRTOFMETAPOROERLDSILLEISTLILDEDGHAP
ITCRUEEADAOHEIATLOENEASPSPPOANINDWCRHIENEAEDESASPAESHAMUOMOU
RSWAISUPANGETFEIINTIIPROATIEEDOWHEHETBSSEBIFLESS TELERILLTIR
TFPOKNFTCAUSEOEAITCHOUSIONTTRQISSTERYBIUANOFHEADRHHARNREATVE
OUMORONGRSOMARHECHEWLOYODCJUEVIMHESSHBPOEOTHREACUMIBTUUASOLT
OGADRTENAKNEBONTLOTGWHTWMIITERSHAVICSIWOITEHSNSSHASFNFOUCKNT

3.4.3 Vygenerované bigramy s pamäťou 1

Bigramy s pamäťou 1 znamená že najskôr sme vygenerovali jeden bigram a pri generovaní ďalších sme sa vždy pozreli na posledný znak, podla neho a príslušnej frekvencie sme generovali nasledujúci znak.
Pre lepšiu predstavu uvedieme príklad:
Ak prvý bigram je AB, pri generovaní nasledujúceho bigramu budeme vyberať len z frekvencií ktoré prislúchajú bigramom začínajúcim znakom B. Následne toto budeme opakovať až kým nevygenerujeme požadovaný počet bigramov.
Výsledky Kolmogorovho-Smirnovho testu pre bigramy s pamäťou 1 sú dané v tabuľke 7 na strane 17.

Slovenčina
UPIALTCHVAMUSINOCINIUSKARAKATASLNAKTOJUTUJNALADALSTAZATAJUNE
LUTANEDUCHLOVIZASASODOKADNKOSAKROVTIVSAZNYCHLAZIAJEDYRASEPOM
DRPITOMIVIEJONTNIESPRAZBONOVIKODNIJUMBOMTATDNAMANVITAZPRAVLA

Španielčina
ATOROESTEDOEJENTODIENLORABREBARASUECAYAUIRABAMOMBIODENOFESTA
RALASACUEMEDEHIDAADASCAMEDENISCHODEMACAPERIRSANZTRRIOADORESI
TAERERABERADOSARTERONOSSHUELESTE RRSDOQUELEETRQUBONGRGUSSEME

Angličtina
NGISAFHICDOUNTHIEDSATASEHETIEPUSPTHIOROUSENGANEGTOVIOFOMASAT
ATEEILDINTHAMENAKSTNASLASOTROFABISOUSUPWHILLETORS WITAYSWNTI
RTORODOSHE OFAVIREDBORENSFAGASONINTYFECHEAKLLYAKTINPRANBOFEL

3.4.4 Vygenerované trigramy s pamäťou 2

Princíp je ten istý ako pri bigramoch s pamäťou 1 len s týmito rozdielmi:

1. na začiatku generujeme trigram
2. pamätáme si posledné dva znaky
3. generujeme na základe frekvencií pre trigramy

Výsledky Kolmogorovho-Smirnovho testu pre trigramy s pamäťou 2 sú dané v tabuľke 8 na strane 17. Trigramový model má síce výsledky Kolmogorovovho-Smirnovovho testu rovnaké ako bigramový test ale z počtu vzniknutých slov môžeme povedať, že je lepší ako predchádzajúce modely. Tento model sa využíva aj v praxi na riešenie niektorých problémov, napr. vo vyhľadávači Google sa (okrem iného) jazyk detekuje pomocou tochto modelu jazyka.

Slovenčina
ASLOCHLTERVEDMICSTASKEBOPOZDYTAMIFONMILAHADANATOTOPECLOSTAMU
SIKSAPECJEMOLISATENTITAKEDATNYKAVOSTAKOVACEROCIKAMANIESCHVIZ
ELNEODOLASPUJUCOSRELOSVELEHAPRUTYVELMIECHODPRIKOSPOPRETYPETL

Španielčina
ERCARINLADARITOABEROBROAGARMADENCIEYATADESTENTALAQUEDETEREPE
RONELTOYORIOSUOSARELSINVIOQUESDULOBLAMPROSTADONVIDOELACARDEI
NAVUERMASATRAJELPRONSALTALOSARECEDFATRICHACANUENRESPORNAPIEN

Angličtina
ASTTHEPLACQUENTOFHICHATEDTHENSHETHEEKNOTALITSWHEHATINEVEINGT
ONEDWINGTHISTMGIREWOLDSCULKEDALET HEGINIHONLOTHISHROUSCOSTORL
OVERYPASLIENMYESTRICETITSOFTHECAMONEREAUVASUPARKBEELOOHORKST

3.4.5 Vygenerované tetragramy s pamäťou 3

Výsledky Kolmogorovho-Smirnovho testu pre tetragramy s pamäťou 3 sú dané v tabuľke 9 na strane 17. Pri tomto modeli z výsledkov môžme povedať, že text generovaný tetragramovým modelom s pamäťou 3 je štatisticky neodlíšiteľný od normálneho textu. Tento model sa využíva pri riešení rôznych problémo v napr. v oblasti kryptoanalýzy pri detekovaní správneho (zmysluplného) textu.

Slovenčina
DOMKU ZA K UTEST VLAC HLAVU VY VELMI MAJU TOT PUSTI NEBA PRINA
ASPOSA SEDEL SA NIM LEBO A VOLA VONI NEJ STIE PANO AKOZIAC
HANOM NA MI VRCHNIE NUZ MAS POSKOCI VTOM TI PUSTI TROM

Španielčina
QUELO ME DE DE TRERON QUE FUESTRABALLEVER NO SI AL PAZONA LMA
NO UN HAGARA VIERRARDUGOBERTOS CAMARADO EL QUE ESCARN AMENSO
A DONDISCUCHAN RECUNSTICO ENTIO PASAJELES COMENO SENOS MENT

Angličtina
ER VICTED THEY LONE LAWYERLED LOUS HAD BECALMOST GRANCE WITH
HAD BLOOK FRAND PASTIFULLY AWAY IN HIS WITH ARTS BARONGS ONE
STIONMADED INNING FOR PENEAL STOUSING IN THE BIN THIS DEARS

4 Slová

Slovo je základná stavebná jednotka jazyka, nesie význam a pozostáva z jednej alebo viacerých morfém. Slová je možné kombinovať do slovných spojení, viet a súvetí. Slovo pozostávajúce z viac ako jednej morfémy (morféma je najmenšia vydeliteľná časť slova alebo slovného tvaru, ktorá je nositeľom významu alebo funkcie. Nositeľom významu je napr. koreňová morféma, príponová morféma, nositeľom funkcie je napr. spájacia morféma (cest-o-pis)) sa nazýva zložené slovo. Slovo je základným prostriedkom jazykovej komunikácie. Môže mať lexikálny a gramatický význam.

4.1 Generovanie slov

Slová a ich frekvencie sme zrátali z rovnakých zdrojov ako pre n-gramy. Generovali sme ich veľmi podobným spôsobom ako n-gramy.Výsledky Kolmogorovho-Smirnovho testu pre slová sú dané v tabuľke 10 na strane 18. Z výsledkov vidíme že tento model je rovnako dobrý ako tetragramový model ale je náročnejší na pamäť lebo potrebuje veľký súbor slov.

Slovenčina
VINA DAMY ON UTEKALA TI POKYN SKRATA AJ TEN PODMYVALA STROJE
PODVIHOL AJ SEDELI CELA VEZME BOLA PO JEDNEHO SVOJIM BOHATOU
NIC LEN VELITELOVI POZOBUDZAJ KRALOVNEJ VELMI NAJMENSIE
POZNAL CESTU A NEODNIESLI URADOVANY TELIATKA NOHACH ABY AKO
TOHO TO PODLHOVASTOU

Španielčina
LAS UN DEJO DE CAROLEA LO FAMOSO EN COMO EN GUSTOPARA YA Y
CONTAGIOSA CON POR ELLA EL METIERON CABALLO MUCHO TOMAR EN
Y A EN UNO LA DE EL DE CUAL ESTA DEVOTOLAS YO Y COSTUMBRES
Y EL EL NASCEHARA SE

Angličtina
THE CONDUCT BY HIM THE THE HIM WAS COTTAGE A IDEAS GORBEAU
SET WAYTHANKS FRENCHFORMATION GREW SECOND USES I MARRIAGE
CERTAIN PATRON IS OF EARNING AS RETURN AZUREDEPTHS LANGUAGE
RIGHT A ALL A ALL SPEAK ANOTHER ON VIRGINITY VALJEAN AN

5 Entropia

5.1 Definícia entropie

Entropia jazyka je štatistický ukazovateľ, ktorý v určitom zmysle hovorí o tom, aké množstvo informácie poskytuje priemerne jedno písmeno textu v tomto jazyku. Entropia H vyjadruje počet bitov informácie, ktorú získame z jedného znaku textu. Takže ak je text zakódovaný binárne, pomocou entropie vyjadrime minimálnu veľkosť kódu.

$$H = \lim_{N\to\infty} F_N$$

5.2 Výpočet entropie

Jednou z metód výpočtu entropie H je postupnosť aproximácií F_0, F_1, ..., ktoré postupne zohľadňujú viac a viac štatistických údajov o jazyku a v limite sa blížia k entropii H. Číslo F_n nazývame entropiou n-gramu a hovorí o tom, aké množstvo informácie o jazyku poskytuje N za sebou idúcich písmen textu.
Hodnota F_N je daná ako:

$$F_N = -\sum_{i,j} p(b_i, j) \log_2 p_{b_i}(j)$$

kde b_i je (N-1)-gram
j je písmeno nasledujúce b_i
$p(b_i, j)$ je pravdepodobnosť N-gramu, ktorý vznikne spojením
b_i a písmena j
$p_{b_i}(j)$ je podmienená pravdepodobnosť písmena j nasledujúceho
za (n-1)-gramom b_i a je daná vzťahom

$$p_{b_i}(j) = p(b_i, j)/p(b_i)$$

Keď máme dvadsaťsedem znakovú abecedu ktorá sa nazýva telegrafná abeceda (TS) čiže dvadsaťšesť písmen a medzera, tak z definície vyplýva pre F_0, že bude $\log_2 27$ čo je 4.75.

$$F_1 = -\sum_{i=1}^{27} p(i) \log_2 p(i) = 4.0999$$

pre bigrami aproximácia F_2 dáva výsledok

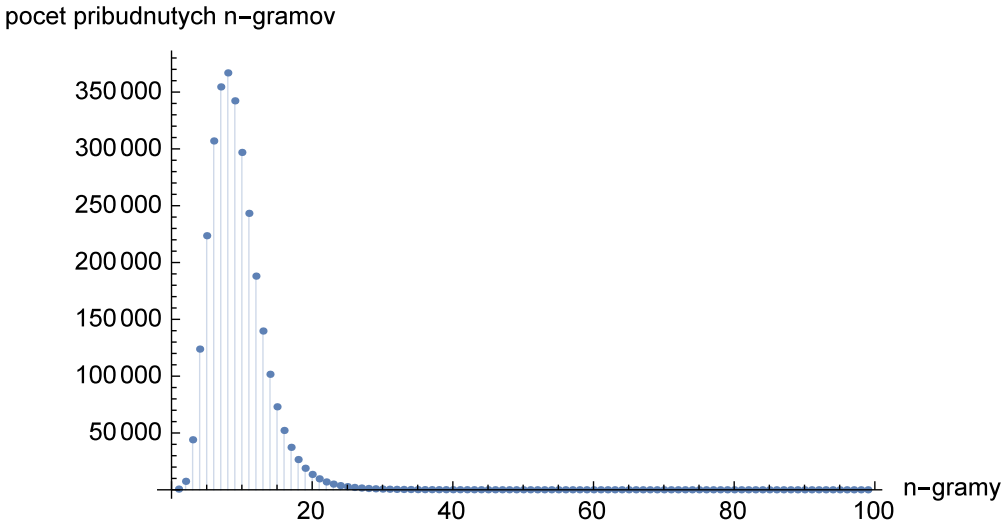
$$\begin{aligned} F_2 &= -\sum_{i,j} p(i, j) \log_2 p_i(j) \\ &= -\sum_{i,j} p(i, j) \log_2 p(i, j) + \sum_i p(i) \log_2 p(i) \\ &= 7.4862 - 4.0999 = 3.3863 \end{aligned}$$

V tabuľke 11 na strane 18 sú ukázané výsledky pre entropiu znakov, bigramov, trigramov a tetragramov.

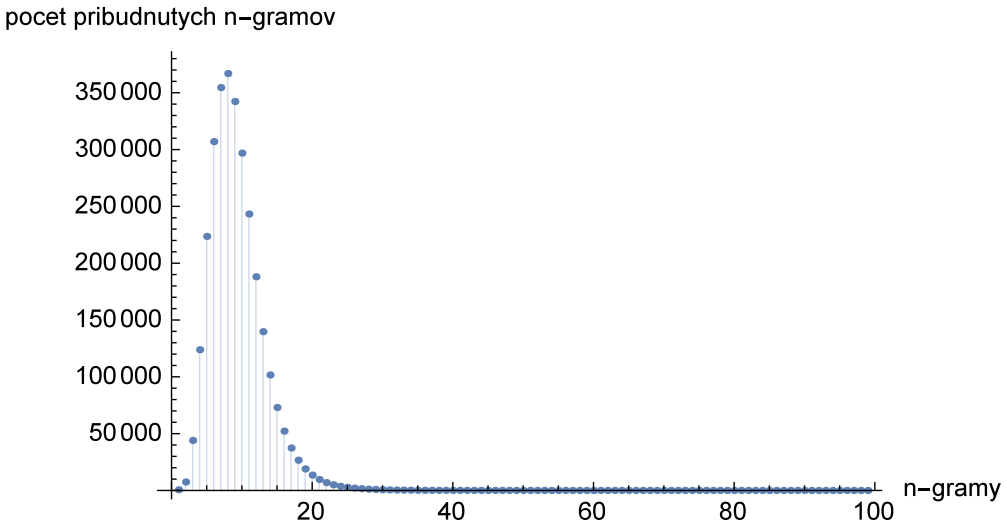
5.3 Závislosť a pokrytie

S entropiou úzko súvisí pokrytie a závislosť znakov. Pod pokrytím sme uvažovali že koľko rôznych n-gramov sa nachádza v skúmanom texte daného jazyka. V tabuľkách 12, 13 na strane 18 sú zobrazené počty rôznych n-gramov zo skúmaného textu. Z pokrytia vidíme že už aj pri malých n-gramoch sa vyskytuje malé percento z celkového možného počtu n-gramu 27^n. To poukazuje nato že aj modely ktoré sú vytvorené z malých n-gramov, sa dajú použi na rôzne úlohy.

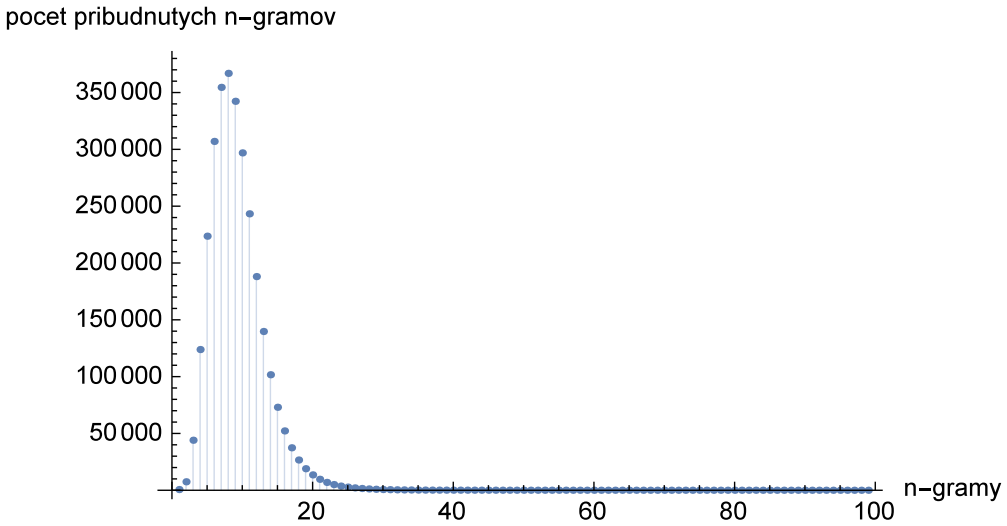
Závislosťou znakov sa nazýva jav keď posledný znak n-gramu nezávisí od predošlých znakov. Prejaví sa to tým že počet pribúdajúcich, rôznych $(n+1)$-gramov prudko klesne. Pre jednotlivé jazyky to môžme vidieť v týchto grafoch: 5, 6, 7 na strách 12, 12 a 13 ktoré zobrazujú rozdiel v počte rôznych n-gramov medzi n-gramom a $(n+1)$-gramom.



Obr. 5: Rozdiely v slovenčine



Obr. 6: Rozdiely v španielčine



Obr. 7: Rozdiely v Angličtine

6 Záver

V práci sme zistili frekvenciu n-gramov a na základe týchto frekvencií sme vytvorili štyri modely prirodzeného jazyka: znakový model, bigramový model, trigramový model, tetragramový model a slovný model. Následne sme zisťovali vhodnosť týchto modelov a či sú dostatočné na modelovanie prirodzeného jazyka. Ďalej sme v práci načrtli čo to je Shannonova entropia a ako sa vypočíta. Na záver sme skúmali závislosť n-gramov. Zostávajúcou úlohou na rozšírenie práce je presné zrátanie Shannonovej entropie a hlbšie preskúmanie závislosti n-gramov.

Frekvencie znakov			
písmeno	**Slovenčina**	**Španielčina**	**Angličtina**
A	0.119881	0.125076	0.081132
B	0.018808	0.015039	0.014673
C	0.031240	0.037024	0.025648
D	0.038491	0.050928	0.042891
E	0.088203	0.135834	0.130855
F	0.001861	0.005419	0.022286
G	0.001782	0.011163	0.019538
H	0.023987	0.011131	0.069538
I	0.067064	0.058149	0.068239
J	0.018504	0.005683	0.002070
K	0.037457	0.000026	0.006295
L	0.052807	0.055945	0.039629
M	0.034364	0.028116	0.024881
N	0.056399	0.069606	0.066811
O	0.094226	0.096936	0.073892
P	0.029914	0.022317	0.016694
Q	0.000007	0.016081	0.001011
R	0.044476	0.063243	0.057873
S	0.055851	0.076187	0.064009
T	0.054651	0.038796	0.092903
U	0.034515	0.046702	0.027707
V	0.040559	0.011210	0.010125
W	0.000489	0.000009	0.022640
X	0.000147	0.000650	0.001519
Y	0.023278	0.013903	0.016457
Z	0.031041	0.004829	0.000687

Tabuľka 1: Frekvencie znakov

Zoradené frekvencie znakov					
znak	**Slovenčina**	**znak**	**Španielčina**	**znak**	**Angličtina**
A	0.119881	E	0.135834	E	0.130855
O	0.094226	A	0.125076	T	0.092903
E	0.088203	O	0.096936	A	0.081132
I	0.067064	S	0.076187	O	0.073892
N	0.056399	N	0.069606	H	0.069538
S	0.055851	R	0.063243	I	0.068239
T	0.054651	I	0.058149	N	0.066811
L	0.052807	L	0.055945	S	0.064009
R	0.044476	D	0.050928	R	0.057873
V	0.040559	U	0.046702	D	0.042891
D	0.038491	T	0.038796	L	0.039629
K	0.037457	C	0.037024	U	0.027707
U	0.034515	M	0.028116	C	0.025648
M	0.034364	P	0.022317	M	0.024881
C	0.031240	Q	0.016081	W	0.022640
Z	0.031041	B	0.015039	F	0.022286
P	0.029914	Y	0.013903	G	0.019538
H	0.023987	V	0.011210	P	0.016694
Y	0.023278	G	0.011163	Y	0.016457
B	0.018808	H	0.011131	B	0.014673
J	0.018504	J	0.005683	V	0.010125
F	0.001861	F	0.005419	K	0.006295
G	0.001782	Z	0.004829	J	0.002070
W	0.000489	X	0.000650	X	0.001519
X	0.000147	K	0.000026	Q	0.001011
Q	0.000007	W	0.000009	Z	0.000687

Tabuľka 2: Zoradené frekvencie znakov

Frekvencie bigramov					
bigram	**Slovenčina**	**bigram**	**Španielčina**	**bigram**	**Angličtina**
AL	0.014828	EN	0.023966	TH	0.034998
NE	0.014189	ES	0.023592	HE	0.032187
ST	0.013710	DE	0.020725	ER	0.018141
NA	0.013615	UE	0.019791	IN	0.017740
LA	0.012570	OS	0.019529	AN	0.015313
RA	0.011227	EL	0.017917	RE	0.014484
OV	0.011210	AS	0.017291	ED	0.013178
PO	0.011170	ER	0.016569	ES	0.012709
AN	0.011101	LA	0.016234	HA	0.012146
TO	0.011045	QU	0.016063	EN	0.011657
AT	0.010385	AN	0.015530	ST	0.011218
TA	0.010096	RA	0.014979	AT	0.011132
EN	0.009933	ON	0.013591	ON	0.011098
AS	0.009899	AD	0.012898	NT	0.011048
SA	0.009494	AL	0.012532	EA	0.010865
KO	0.009253	RE	0.012319	HI	0.010853
CH	0.009245	DO	0.012264	TO	0.010589
ED	0.009233	AR	0.012215	ND	0.010496
IE	0.009158	SE	0.012077	IS	0.010258
AK	0.009036	NT	0.010763	OU	0.009901
PR	0.009021	CO	0.010439	ET	0.009285
NI	0.008826	OR	0.010005	AS	0.009122
OM	0.008802	NO	0.009912	IT	0.008539
HO	0.008705	TA	0.009534	OF	0.008392
ES	0.008377	ST	0.009530	SE	0.008357
OS	0.008336	TE	0.009385	AR	0.008338
LI	0.008088	LO	0.008653	NG	0.008333
VA	0.008049	IE	0.008610	TE	0.008215
OL	0.007957	SA	0.008572	OR	0.008083
AV	0.007931	RO	0.008497	TI	0.007503

Tabuľka 3: Frekvencie bigramov

Absolútne a percentuálne počty n-gramov			
Jazyk	**znaky**	**bigramy**	**trigramy**
Slovenčina	16 (61.5 %)	218 (32.2 %)	2 540 (14.5 %)
Španielčina	13 (50 %)	163 (24.1 %)	1 501 (8.5 %)
Angličtina	15 (57.7 %)	213 (31.5 %)	2 216 (12.6 %)

Tabuľka 4: Počet n-gramov pokrývajúcich 90 % z celkového počtu

Výsledky K-S testu pre znaky				
porovnanie frekvencií	znaky	bigramy	trigramy	tetragramy
Slovenčina	rovnaké	rozdielne	rozdielne	rozdielne
Španielčina	rovnaké	rozdielne	rozdielne	rozdielne
Angličtina	rovnaké	rozdielne	rozdielne	rozdielne

Tabuľka 5: Výsledky K-S testu pre znaky

Výsledky K-S testu pre bigramy				
porovnanie frekvencií	znaky	bigramy	trigramy	tetragramy
Slovenčina	rovnaké	rovnaké	rovnaké	rozdielne
Španielčina	rovnaké	rozdielne	rozdielne	rozdielne
Angličtina	rovnaké	rozdielne	rozdielne	rozdielne

Tabuľka 6: Výsledky K-S testu pre bigramy

Výsledky K-S testu pre bigramy s pamäťou 1				
porovnanie frekvencií	znaky	bigramy	trigramy	tetragramy
Slovenčina	rovnaké	rovnaké	rovnaké	rozdielne
Španielčina	rovnaké	rovnaké	rovnaké	rozdielne
Angličtina	rovnaké	rovnaké	rovnaké	rozdielne

Tabuľka 7: Výsledky K-S testu pre bigramy s pamäťou 1

Výsledky K-S testu pre trigramy s pamäťou 2				
porovnanie frekvencií	znaky	bigramy	trigramy	tetragramy
Slovenčina	rovnaké	rovnaké	rovnaké	rozdielne
Španielčina	rovnaké	rovnaké	rovnaké	rozdielne
Angličtina	rovnaké	rovnaké	rovnaké	rozdielne

Tabuľka 8: Výsledky K-S testu pre trigramy s pamäťou 2

Výsledky K-S testu pre tetragramy s pamäťou 3				
porovnanie frekvencií	znaky	bigramy	trigramy	tetragramy
Slovenčina	rovnaké	rovnaké	rovnaké	rovnaké
Španielčina	rovnaké	rovnaké	rovnaké	rovnaké
Angličtina	rovnaké	rovnaké	rovnaké	rovnaké

Tabuľka 9: Výsledky K-S testu pre tetragramy s pamäťou 3

Výsledky K-S testu pre slová				
porovnanie frekvencií	znaky	bigramy	trigramy	tetragramy
Slovenčina	rovnaké	rovnaké	rovnaké	rovnaké
Španielčina	rovnaké	rovnaké	rovnaké	rovnaké
Angličtina	rovnaké	rovnaké	rovnaké	rovnaké

Tabuľka 10: Výsledky K-S testu pre slová

Entropia				
jazyk	znaky	bigramy	trigramy	tetragramy
Slovenčina	4.1190	3.4028	2.9843	2.5228
Španielčina	3.9944	3.1558	2.7257	2.3316
Angličtina	4.0998	3.3863	2.8074	2.2964

Tabuľka 11: Entropia

Pokrytia						
jazyk	znaky	bigramy	trigramy	tetragramy	32-gramy	33-gramy
Slovenčina	27	568	6810	42812	2423727	2424695
Španielčina	27	512	5447	36011	2993592	2994791
Angličtina	27	652	8108	52134	2996461	2996846

Tabuľka 12: Pokrytia

Pokrytia 2						
jazyk	34-gramy	35gramy	38-gramy	39-gramy	98-gramy	99-gramy
Slovenčina	2425585	2426415	2428604	2429259	2443775	2443857
Španielčina	2995800	2996641	2998281	2998569	2999902	2999901
Angličtina	2997148	2997394	2997848	2997955	2999638	2999643

Tabuľka 13: Pokrytia 2

Zoznam obrázkov

Zoznam tabuliek

Literatúra

[1] Anděl, J.: Statistické metody, *MatfyzPress, Praha 2007*

[2] Gaines, H. F.: Cryptanalysis, a study of ciphers and their solution, *Dover Publication Inc., New York, 1939*

[3] Ganesan, R., Sherman, A.: Statistical Techniques for Language Recognition: An Introduction and Guide for Cryptanalysis, `http://web.cecs.pdx.edu/ bart/decrypter/`*, 1993*

[4] Jaglom, A. M., Jaglom, I. M..: Pravdepodobnosť a informácia, časť 4, kap. 2 a 3, str. 195–273, *Moskva 1973*

[5] Shannon, C. E.: A Mathematical Theory of Communication, *Bell System Technical Journal Vol. 27, No. 3, [379–423], 1948*

[6] Shannon, C. E.: Prediction and Entropy of Printed English, *Bell System Technical Journal Vol. 30, [50–64], 1951*