Poročilo za izbirni projekt 1B (Simulacija Markovovega vira slovenskega besedila), pri predmetu Informacija in Kodi

Gašper Šavle[1]

[1] *Univerza v Ljubljani, Fakulteta za Elektrotehniko*
E-pošta: gaspersavle@yahoo.com

Abstract

The aim of this project was to develop a program for analyzing Slovenian literature and simulating a Markov source model [1] using statistical text analysis. Key tasks included frequency analysis, examining the resemblance to Zipf's law [2], and calculating entropy to explore text complexity. A Markov simulation was implemented to generate random text reflecting the style of the original templates, with multiple sampling strategies and user-initiated text generation from three input words.

1 Uvod

Cilj naloge je bil razviti računalniški program za informacijsko analizo slovenskih besedil in simulacijo delovanja Markovovega vira [1], ki temelji na rezultatih statistične analize besedil. Glavni cilj je bil izvesti poglobljeno analizo besedil iz danega nabora s poudarkom na razumevanju frekvenčne porazdelitve besed, ocenah njihovih verjetnosti in pogojnih verjetnosti ter na njihovi osnovi simulirati oddajanje naključnega besedila v slogu originalnega gradiva.

V poročilu podajam rezultate frekvenčne analize, kjer so obravnavane najpogostejše, najmanj pogoste in besede z verjetnostjo, blizu geometrični sredini med skrajnimi vrednostmi. Preveril sem skladnost statistične distribucije besed z Zipfovim zakonom [2] in izračunal različne vrste entropije, kar omogoča vpogled v kompleksnost in strukturiranost besedil.

Poleg analitičnega dela sem izvedel simulacijo Markovovega vira, ki na osnovi ocenjenih lastnih in pogojnih verjetnosti generira naključno besedilo. Implementiral sem različne strategije vzorčenja, ki omogočajo različne izide generacije besedila. Program omogoča umetno generiranje besedil na podlagi poljubno naključno izbranih treh besed.

2 Metodologija

2.1 Priprava podatkov

Za analizo in simulacijo nam je bil podan arhiv slovenskih leposlovnih besedil. Vsaka datoteka v arhivu je predstavljala eno besedilo. Besedila so najprej bila razčlenjena na podlagi presledkov, tabulatorjev in znakov za naslednjo vrstico. Nato sem seznam besed še nadalje razčlenil glede na ločila, ki so jih vsebovali posamezni elementi originalne razčlembe. Izbral sem vsa ločila, ki so se pojavljala v besedilih, jih odbil elementu in novo pridobljena elementa dodal novemu seznamu dokončno razčlenjenih besed.

2.2 Statistična analiza posameznih besed

Na podlagi razčlenjenega besedila sem izvedel različne statistične analize

2.2.1 Frekvenčna analiza

Z uporabo Python orodja `Counter`, iz knjižnice `colections` sem preštel vse unikatne besede posameznega besedila. Prav tako sem pridobil skupno število besed posameznega besedila in s temi podatki izračunal verjetnost posamezne besede v obravnavanem besedilu. Izločil sem najpogostejših in najmanj pogostih 5 besed z uporabo metode `most_common()`, nato pa izračunal geometrično povprečje verjetnosti besed z metodo,

$$p_{rms} = \sqrt{max(verjetnosti)^2 - min(verjetnosti)^2}, \tag{1}$$

kjer $verjetnosti$ predstavljajo slovar verjetnosti vseh unikatnih besed izbranega besedila. Ko sem pridobil geometrijsko povprečje vseh verjetnosti, sem izbral besedo s prvo manjšo verjetnostjo od povprečne.

2.2.2 Preverjanje skladnosti z Zipfovim zakonom

Za preverjanje skladnosti z Zipfovim zakonom je bilo besede potrebno urediti po verjetnosti od največje, do najmanjše, za kar je prikladno že poskrbel objekt `Counter`. Besedam sem glede na mesto v slovarju priredil rang. Za vsak rang sem izračunal teoretično vrednost, ki jo narekuje Zipfov zakon, po formuli

$$zipf(rang) = \frac{max(n)}{rang}. \tag{2}$$

Nato sem na graf z logaritmično skalo izrisal dejanske frekvence unikatnih besed pri določenih rangih in teoretične Zipfove frekvence pri istih rangih. Na koncu sem izračunal tudi koeficient krizne korelacije med obema potekoma.

2.2.3 Izračun pogojnih verjetnosti

Sprehodil sem se skozi seznam unikatnih besed in po 2 zaporedni združil v nov seznam bigramov. Nato sem

preštel frekvenco unikatnih bigramov ter izračunal njihove verjetnosti

$$P(AB) = \frac{n(AB)}{n_{besed}}, \tag{3}$$

z uporabo novo-pridobljenih verjetnosti bigramov in prej izračunanih verjetnosti posameznih besed sem nato izračunal pogojne verjetnosti bigramov.

$$P(B|A) = \frac{P(AB)}{P(A)} \tag{4}$$

V naslednjih fazah naloge je bila potrebna tudi generacija besedila z Markovim virom z uporabo verjetnosti besede glede na prejšnje 3, tako sem moral izračunati tudi pogojne verjetnosti tri-gramov in štiri-gramov, kar sem storil z uporabo naslednjih formul

$$P(C|AB) = \frac{P(ABC)}{P(A) \cdot P(B|A)}, \tag{5}$$

$$P(D|ABC) = \frac{P(ABCD)}{P(A) \cdot P(B|A) \cdot P(C|AB)}. \tag{6}$$

2.3 Izračun entropije

Za boljše razumevanje jezikovne strukture in določitev kompleksnosti besedila je bilo potrebno izračunati entropije posameznega besedila, izračunati je bilo potrebno 3 entropije na besedilo, pri tem pa upoštevati naslednje kriterije:

2.3.1 Enake verjetnosti besed

Če privzamemo, da imajo vse besede enako verjetnost, to verjetnost izračunamo, kot kvocient

$$P_{eq} = \frac{1}{n_{besed}}. \tag{7}$$

Entropijo besedila s to predpostavko izračunamo po formuli

$$H = -\sum_{i=0}^{N} P_{eq} \cdot log_2(P_{eq}) \tag{8}$$

2.3.2 Dejanske verjetnosti besed

Za izračun entropije uporabimo dejanske verjetnosti posameznih besed besedila (2.2). Entropijo besedila s tako verjetnostjo distribucijo izračunamo po formuli

$$H = -\sum_{i=0}^{N} P_i \cdot log_2(P_i) \tag{9}$$

2.3.3 Pogojne verjetnosti bigramov

Za izračun entropije uporabimo pogojne verjetnosti posameznih bigramov (2.2.3). Entropijo besedila s tako verjetnostjo distribucijo izračunamo po formuli

$$H = -\sum_{i=0}^{N} P(B|A)_i \cdot log_2(P(B|A)_i) \tag{10}$$

2.4 Simulacija Markovega vira

Markov vir generira naključna besedila na osnovi ocenjenih verjetnosti besed, naloga je zahtevala, da to implementiramo s kontekstom 1 in 3 predhodnih besed. Generator s 3 predhodnimi besedami konteksta, pa je moral biti implementiran v 3 različnih izvedbah, vsako z drugačnim načinom vzorčenja

2.4.1 Ena predhodna beseda

Za generacijo teksta z eno predhodno besedo konteksta sem funkciji podal eno, naključno izbrano besedo iz besedila, nato sem iz slovarja verjetnosti bigramov izbral tiste, v katerih je nastopala te podana beseda. Za naslednjo besedo sem nato izbral besedo iz novega nabora, ki ima najvišjo verjetnost. Funkcija se ponovi tolikokrat, kot je dolg nabor unikatnih besed, pri tem pa se v vsaki iteraciji "nova beseda"preslika v začetno besedo naslednje iteracije.

Za generacijo teksta s 3 predhodnimi besedami konteksta sem funkciji podal 3 naključno izbrane besede, nato pa iz slovarja štiri-gramov izbral tiste, ki vsebujejo prve 3 podane besede. Te kandidate sem preslikal v nov slovar, kjer sem verjetnosti posameznih štiri-gramov normaliziral. Naslednjo besedo sem nato izbral na 3 različne načine:

2.4.2 Deterministično vzorčenje s 3 besedami konteksta

Pri deterističnem vzorčenju, je naslednja beseda izbrana, kot beseda iz nabora možnih naslednjih besed, z najvišjo verjetnostjo, tako, kot je bilo storjeno v prejšnjem razdelku (2.4.1).

$$i_{max} = argmax(P), \tag{11}$$

$$naslednja = kandidati(i_{max}). \tag{12}$$

2.4.3 Uteženo, naključno vzorčenje s 3 besedami konteksta

Pri uteženem, naključnem vzorčenju, je naslednja beseda izbrana, kot naključna izbira iz nabora možnih besed, kjer na izbiro vpliva verjetnostna distribucija besed v tem naboru. To sem izvedel z uporabo metode `random.choice()` iz knjižnice `numpy`, ki omogoča, da ji kot argument vnesemo verjetnostno porazdelitev kandidatov za izbor [3].

2.4.4 Naključno, uteženo vzorčenje s 3 besedami konteksta in parametrom temperature

Naključno vzorčenje s temperaturnim parametrom τ je metoda, ki omogoča prilagoditev ravnotežja med deterministično izbiro in psevdo-naključno izbiro. Za naključno, uteženo vzorčenje s parametrom temperature moramo verjetnosti posameznih besed iz nabora kandidatove preoblikovati z uporabo formule:

$$\tilde{p}_i = \frac{p_i^\tau}{\sum_i^N p_i^\tau} \tag{13}$$

Če parameter τ limitira proti ∞, porazdelitev postaja deterministična, torej ima samo beseda z največjo verjetnostjo, verjetnost 1, ostale pa 0. Če je vrednost parametra τ enaka 1, ostane verjetnostna distribucija nespremenjena. Če pa vrednost parametra τ limitira proti 0, se

porazdelitev približuje enakomerni verjetnostni porazdelitvi, kjer imajo vse besede enako verjetnost, torej izbira postane naključna

3 Rezultati

V sledečem poglavju sta predstavljeni analizo in simulacija Markovovega vira slovenskega besedila, izvedena na osnovi slovenski leposlovnih besedil. Izračunal sem frekvence, lastne in pogojne verjetnosti besed, preveril skladnost z Zipfovim zakonom ter ocenili entropijo besedila pod različnimi predpostavkami. Poleg tega sem razvili program, ki generira nova besedila z uporabo različnih strategij vzorčenja. Prikazana je tudi simulacija višjerednega Markovovega vira za bolj kompleksne jezikovne vzorce.

3.1 Frekvence in verjetnosti besed

V naslednjih tabelah so prikazani rezultati statistične analize posameznih besedil, ti rezultati zajemajo 5 najpogostejših in najmanj pogostih besed vsakega besedila ter besedo, z verjetnostjo, najbližjo geometričnemu povprečju verjetnosti posameznega besedila.

Tabela 1: Statistična analiza besedila 'Gospa Judit'

Gospa Judit		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	2898
	je	1713
	.	1073
	in	1044
	se	769
Najredkejše	usodi	1
	premišljeval	1
	življenjem	1
	javnim	1
	pisanim	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
je	0.05057	

Tabela 2: Statistična analiza besedila 'Hiša Marije Pomočnice'

Hiša Marije Pomočnice		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	3291
	je	3026
	in	1392
	.	1315
	se	1148
Najredkejše	pričakovani	1
	ljubljeni	1
	ženin	1
	spleteni	1
	žarkov	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
je	0.08502	

Tabela 3: Statistična analiza besedila 'Hlapci'

Hlapci		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	1311
	:	734
	.	636
	je	487
	in	410
Najredkejše	novo	1
	dekle	1
	Duša	1
	ogrnjena	1
	ZUNAJ	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
:	0.04096	

Tabela 4: Statistična analiza besedila 'Hlapec Jernej in njegova pravica'

Hlapec Jernej in njegova pravica		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	1951
	je	1474
	in	884
	.	715
	se	494
Najredkejše	grešnih	1
	ognja	1
	Jernejevi	1
	stopili	1
	plamena	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
je	0.06602	

Tabela 5: Statistična analiza besedila 'Križ na gori'

Križ na gori		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	3645
	je	3204
	se	1509
	.	1481
	in	1452
Najredkejše	slutnji	1
	radostni	1
	koprneč	1
	zlatordečem	1
	obsenčil	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
je	0.07045	

Tabela 6: Statistična analiza besedila 'Na klancu'

Na klancu		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	5899
	je	5699
	in	2941
	.	2211
	se	2101
Najredkejše	učiteljevo	1
	iskra	1
	plapolajoča	1
	vžigalo	1
	prižigale	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
je	0.08209	

Tabela 7: Statistična analiza besedila 'Podobe iz sanj'

Podobe iz sanj		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	3860
	je	1741
	in	1299
	.	1119
	se	866
Najredkejše	ozdravel	1
	odrešnica	1
	svetnica	1
	čaju	1
	bolezni	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
je	0.04666	

Tabela 8: Statistična analiza besedila 'Pohujšanje v dolini šentflorjanski'

Pohujšanje v dolini šentflorjanski		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	3860
	!	1741
	.	1299
	je	1119
	in	866
Najredkejše	uči	1
	mát'	1
	pobožno	1
	svetosti	1
	nauk	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
!	0.04677	

Tabela 9: Statistična analiza besedila 'Tujci'

Tujci		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	4530
	je	3152
	in	1761
	.	1692
	se	1417
Najredkejše	stražnik	1
	finančni	1
	zašumela	1
	Voda	1
	oprostil	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
je	0.06077	

Tabela 10: Statistična analiza besedila 'Za narodov blagor'

Za narodov blagor		
Pogostost	Beseda	Frekvenca
Najpogostejše	,	2103
	.	995
	...	859
	je	648
	-	469
Najredkejše	šipe	1
	razbil	1
	Ostavite	1
	(Odideta)	1
	Skrajni	1
Najbližje geometrični sredini		
Beseda	Verjetnost	
je	0.06077	

3.1.1 Najpogosteje ponavljajoče-se besede

Iz tabel je razvidno, da v praktično vseh besedilih najpogosteje nastopajo predvsem ločila (vejica, pika) in pomožne besede slovenskega jezika, kot so ”je”, ”in”, in še”. To kaže na osnovne slovnične značilnosti slovenskega jezika, kjer vezniki, pomožni glagoli in refleksivne oblike igrajo pomembno vlogo. Frekvenca ločil, kot so vejice, odraža kompleksen stavčni slog in strukturo besedil.

3.1.2 Najredkeje ponavljajoče-se besede

Besede, ki se pojavijo najredkeje (samo enkrat), so običajno bolj specifične za vsebino posameznega besedila. Pogosto vključujejo manj pogoste izraze, imena, lastnosti ali redke samostalnike, ki odsevajo unikatne teme ali slog pisanja avtorja, kažejo na specifičnost teh besedil. Ker je 1 najnižja možna frekvenca besede, so rezultati najmanj pogosto rabljenih besed, ki jih vrne program naključno izbrani iz nabora vseh besed s frekvenco 1. Zaradi te lastnosti programa, se lahko izhodni podatki spreminjajo med različnimi ponovitvami analize.

3.1.3 Beseda, najbližje geometrični sredini

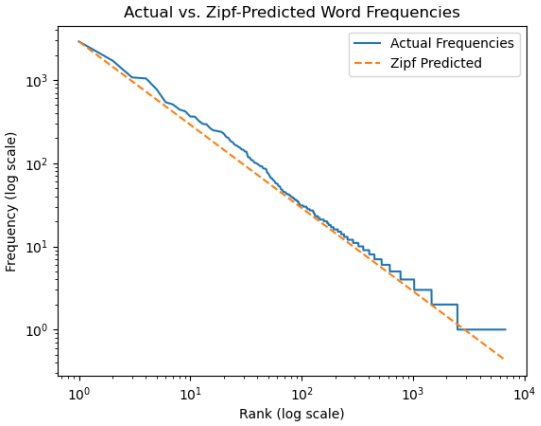
V skoraj vseh primerih, je beseda ”je”imela verjetnostjo najbližje geometrični sredini, kar kaže na njeno pogostost, ki je dovolj uravnotežena, da ne prevladuje, a je kljub temu ključna za slovnično strukturo. To nakazuje tudi vlogo glagola ”biti”kot enega najosnovnejših gradnikov jezika.

3.1.4 Zanimivosti

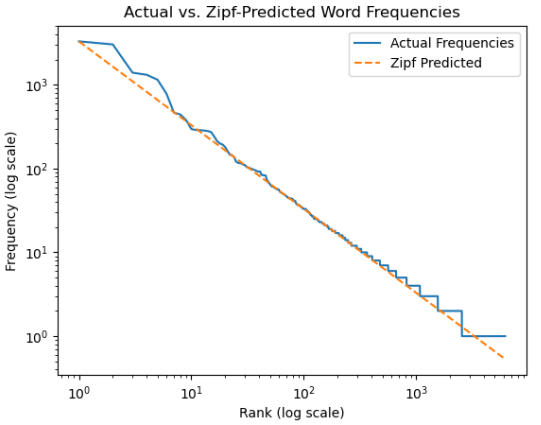
Besedila, kot je Za narodov blagor, izstopajo z uporabo posebnih ločil, kot je ”...”10, kar kaže na dramatiko ali dialog. V besedilu Pohujšanje v dolini šentflorjanski pa je razvidna visoka frekvenca klicaja (”!”)8, kar nakazuje na čustveno nabit slog.

3.2 Sledenje zipfovemu zakonu

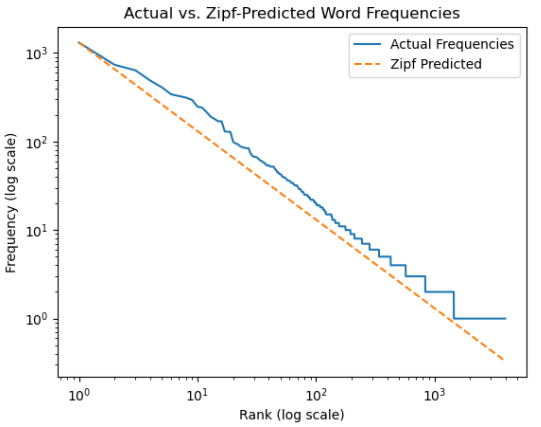
V naslednjih grafih in tabeli so prikazani rezultati analize skladnosti porazdelitve besed v posameznih besedilih z Zipfovim zakonom. Ti rezultati vključujejo primerjavo dejanske porazdelitve pogostosti besed v besedilih z idealno porazdelitvijo po Zipfovem zakonu ter oceno, kako dobro se besedila približajo predvidenemu jezikovnemu vzorcu.



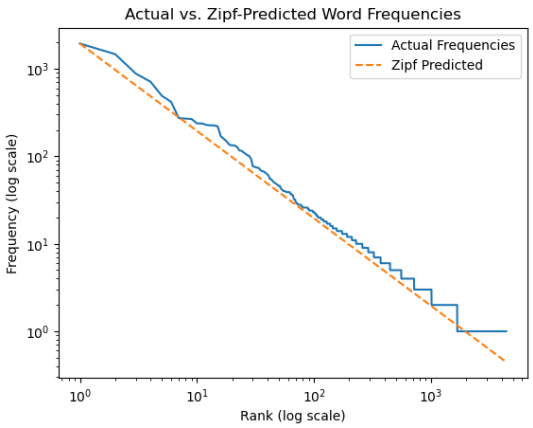
Slika 1: Prileganje verjetnostne porazdelitve besedila ’Gospa Judit’ z Zipfovim zakonom



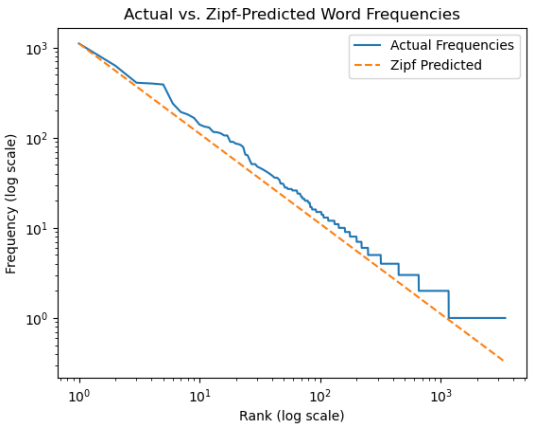
Slika 2: Prileganje verjetnostne porazdelitve besedila ’Hiša Marije Pomočnice’ z Zipfovim zakonom



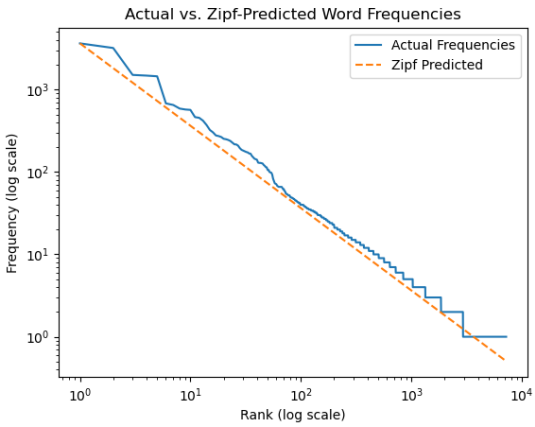
Slika 3: Prileganje verjetnostne porazdelitve besedila ’Hlapci’ z Zipfovim zakonom



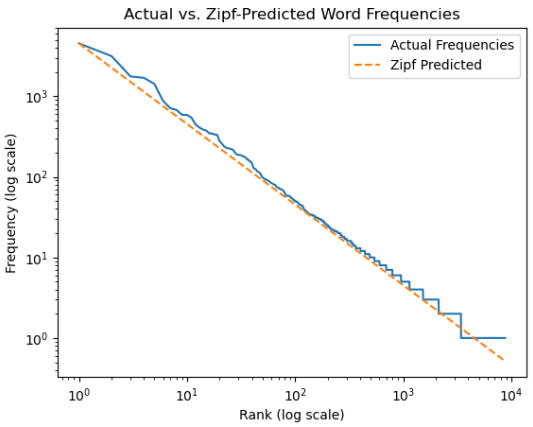
Slika 4: Prileganje verjetnostne porazdelitve besedila 'Hlapec Jernej in njegova pravica' z Zipfovim zakonom



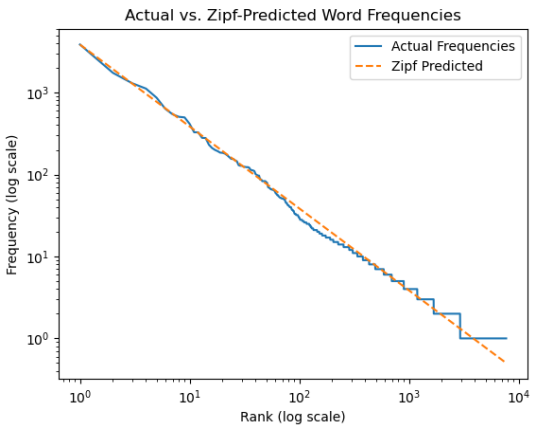
Slika 7: Prileganje verjetnostne porazdelitve besedila 'Pohujšanje v dolini šentflorjanski' z Zipfovim zakonom



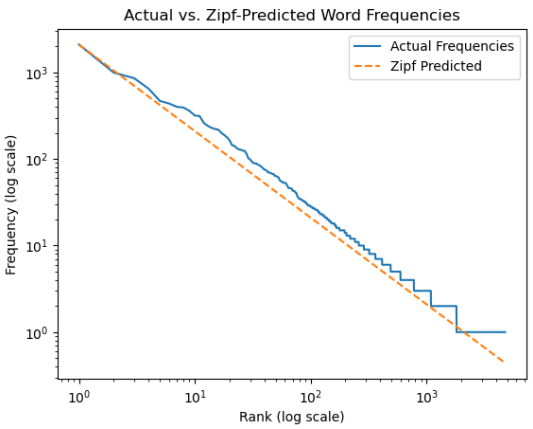
Slika 5: Prileganje verjetnostne porazdelitve besedila 'Križ na gori' z Zipfovim zakonom



Slika 8: Prileganje verjetnostne porazdelitve besedila 'Tujci' z Zipfovim zakonom



Slika 6: Prileganje verjetnostne porazdelitve besedila 'Podobe iz sanj' z Zipfovim zakonom



Slika 9: Prileganje verjetnostne porazdelitve besedila 'Za narodov blagor' z Zipfovim zakonom

V večini primerov besedila sledijo Zipfovemu zakonu, kar pomeni, da se pogostost besed zmanjšuje s povečevanjem njihovega ranga. Manjša odstopanja, predvsem pri red-

kejših besedah, kažejo na slogovne posebnosti in tematsko usmerjenost posameznih del. To potrjuje, da je Zipfov zakon močan model za analizo porazdelitve besed v naravnem jeziku, pri čemer unikatnosti posameznih besedil ponujajo vpogled v avtorjev slog in strukturo besedila.

Na grafu (2) je razvidno večje odstopanje pri nižjih rangih, če se sklicujemo na tabelo (2) lahko opazimo, da je med prvima dvema najpogostejšima besedama relativno majhna razlika, Zipfov zakon pa nam narekuje, da bi morala biti frekvenca druge najpogostejše besede približno enaka polovici frekvence najpogostejše besede.

Dobro ujemanje z Zipfovim zakonom lahko predvidimo tudi glede na besedo, najbližje geometrični sredini (3.1.3). Zipfov zakon lahko opišemo s padajočo geometrično vrsto, torej bi druga najpogostejša beseda morala biti približno podobna geometrijski sredini verjetnostne porazdelitve.

Korelacija z Zipfovim zakonom	
Besedilo	Koeficient korelacije
Gospa Judit	0.99239
Hiša Marije Pomočnice	0.96536
Hlapci	0.97455
Hlapec Jernej in njegova pravica	0.98338
Križ na gori	0.96818
Na klancu	0.96181
Podobe iz sanj	0.99822
Pohujšanje v dolini šentflorjanski	0.98825
Tujci	0.98773
Za narodov blagor	0.98921

Tabela 11: Ujemanje verjetnostne porazdelitve posameznih besedil z Zipfovim zakonom

3.3 Pogojne verjetnosti

Izračunane pogojne verjetnosti parov besed so predstavljene v `json` datotekah, ki jih vrne program. Za vsako besedilo se dotična datoteka z rezultati nahaja v mapi z imenom besedila. Zaradi ogromne količine podatkov, je nisem vključil v poročilo.

3.4 Entropije

V naslednji tabeli so prikazani rezultati analize entropije posameznih besedil pri različnih pogojih. Ti rezultati zajemajo entropijo ob predpostavki enakih verjetnosti vseh besed, entropijo na podlagi dejanske porazdelitve verjetnosti besed ter entropijo, izračunano z upoštevanjem pogojnih verjetnosti, glede na predhodno besedo v besedilu. S tem je omogočen vpogled v informacijsko kompleksnost in predvidljivost posameznega besedila.

Tabela 12: Entropije besedil

Entropije besedil		
Porazdelitev	Besedilo	Entropija
Enakomerna porazdelitev	Gospa Judit	15.05
	Hiša Marije Pomočnice	15.12
	Hlapci	14.12
	Hlapec Jernej in njegova pravica	14.44
	Križ na gori	15.47
	Na klancu	16.08
	Podobe iz sanj	15.18
	Pohujšanje v dolini šentflorjanski	13.71
	Tujci	15.66
	Za narodov blagor	14.61
Dejanske verjetnosti	Gospa Judit	9.40
	Hiša Marije Pomočnice	8.98
	Hlapci	9.10
	Hlapec Jernej in njegova pravica	8.94
	Križ na gori	9.18
	Na klancu	9.20
	Podobe iz sanj	9.51
	Pohujšanje v dolini šentflorjanski	9.09
	Tujci	9.42
	Za narodov blagor	9.16
Pogojne verjetnosti	Gospa Judit	4020.57
	Hiša Marije Pomočnice	3847.52
	Hlapci	2289.45
	Hlapec Jernej in njegova pravica	2635.54
	Križ na gori	4765.65
	Na klancu	6809.17
	Podobe iz sanj	4662.20
	Pohujšanje v dolini šentflorjanski	1733.85
	Tujci	5659.0
	Za narodov blagor	2855.73

Entropija pri enakomerni porazdelitvi predstavlja teoretično zgornjo mejo kompleksnosti besedila, kjer so vse besede enako verjetne. Višje vrednosti kažejo na večje število unikatnih besed v besedilu ali večjo dolžino besedila. Najvišjo entropijo ima Na klancu (16.08), kar nakazuje, da vsebuje največ raznolikih besed med vsemi analiziranimi besedili. Najnižjo entropijo ima Pohujšanje v dolini šentflorjanski (13.71), kar kaže na manj raznoliko besedišče in večjo koncentracijo na manjše število besed.

Entropija pri dejanskih verjetnostih upošteva dejansko pogostost besed v besedilu in kaže, koliko informacij povprečno prinese ena beseda. Nižje vrednosti pomenijo, da so določene besede zelo pogoste (večja predvidljivost). Najnižjo entropijo ima 'Hlapec Jernej in njegova pravica' (8.94), kar pomeni, da je uporaba besed bolj predvidljiva, ponavadi zaradi pogostega ponavljanja določenih izrazov ali slogovnih elementov. Najvišjo entropijo ima 'Podobe iz sanj' (9.51), kar kaže na bolj ra-

znoliko porazdelitev besed in manjšo predvidljivost. Razlike med besedili so majhne, saj so vse vrednosti blizu 9, kar kaže na relativno podobno jezikovno strukturo v vseh analiziranih delih. Za razliko od entropije pri enakomerni porazdelitvi ima dolžina besedila, na entropijo pri dejanskih verjetnostih besed, manjši vpliv.

Entropija pri pogojnih verjetnostih upošteva informacijo o predhodnih besedah in meri, kako informativno je besedilo glede na zaporedje besed. Višje vrednosti kažejo na večjo kompleksnost stavkov in bolj raznolik slog. Najvišjo entropijo ima Na klancu (6809.17), kar nakazuje na zelo kompleksno strukturo stavkov in raznoliko uporabo besed glede na kontekst. Najnižjo entropijo ima Pohujšanje v dolini šentflorjanski (1733.85), kar pomeni, da so zaporedne besede bolj predvidljive na podlagi prejšnjih, kar je lahko posledica preprostega ali ponavljajočega se sloga. Besedila, kot so Tujci (5659.0) in Križ na gori (4765.65), kažejo na bogatejšo sintaktično strukturo v primerjavi z enostavnejšimi deli, kot so Hlapci (2289.45) in Pohujšanje v dolini šentflorjanski.

3.5 Simulacija Markovega vira

Generacija besedila z Markovim virom je bila izvedena pri 3 različnih načinih vzorčenja naslednje besede. Generacija je bazirana na verjetnostni porazdelitvi unikatnih besed v dotičnem besedilu, zato je nabor generacije sestavljen le iz bese, ki sestavljajo izvorno besedilo. Generacija je bila inicializirana s 3 naključno izbranimi besedami iz izvornega besedila. Zaradi velike količine podatkov sem rezultate izpustil iz poročila, vendar so na voljo v obliki `.txt` dokumenta, ki se za vsako besedilo nahaja v njegovi dotični mapi z rezultati. V imenu dokumenta je vkljucena metoda vzorcenja, s katero je bil generiran. Vsaka mapa vsebuje 3 datoteke, ki so bile generirane s 3 začetnimi besedami in eno datoteko, ki je bila generirana z enobesednim kontekstom (deterministično).

Deterministična generacija besedil temelji na izbiri najbolj verjetne besede v danem kontekstu, kar zagotavlja doslednost in jasnost. Vendar, kot je razvidno iz deterministično generiranih izhodnih besedil, ta pristop pogosto vodi v ponavljajoče se fraze in monotonost, saj se model strogo drži najbolj očitnih možnosti. Posledično je slog besedila tog, predvidljiv in pogosto dolgočasen, brez naravne razgibanosti, ki je značilna za literarna dela. V vseh primerih deterministicne generacije besedila, se je izhod žaciklal", torej začel ponavljati iste stavke ali celo besede.

Temperaturna generacija ponuja bolj naraven in ustvarjalen slog besedila. S prilagajanjem temperature model doseže ravnovesje med predvidljivostjo in presenečenjem, kar omogoča večjo raznolikost besedišča in slogovnih elementov. Besedila so bogatejša, s podrobnejšimi opisi in bolj dinamično strukturo. Čeprav lahko pri previsoki temperaturi pride do rahle izgube smiselnosti, ta način generacije najbolj učinkovito združuje kreativnost in naravno berljivost.

Uteženo naključno generiranje besedil poudarja nenavadnost in nepredvidljivost, kar vodi v kaotične in pogosto nelogične kombinacije besed. Ta pristop ustvarja unikatne in presenetljive fraze, vendar pogosto izgubi smiselnost in povezavo med stavki. Medtem ko je primeren za eksperimentalne namene ali ustvarjanje nepričakovanih vzorcev, so taka besedila manj berljiva in slogovno manj skladna, kar jih omejuje za uporabo v bolj strukturiranih kontekstih.

Literatura

[1] Wikipedia contributors, "Markov information source — Wikipedia, the free encyclopedia," 2024. [Online; accessed 11-January-2025].

[2] Wikipedia contributors, "Zipf's law — Wikipedia, the free encyclopedia," 2024. [Online; accessed 11-January-2025].

[3] NumPy developers, "numpy.random.choice — numpy v2.1 manual," 2025. [Online; accessed: 11-January-2025].