2. domača naloga: podobnost jezikov

Domen Lušina (27172023)

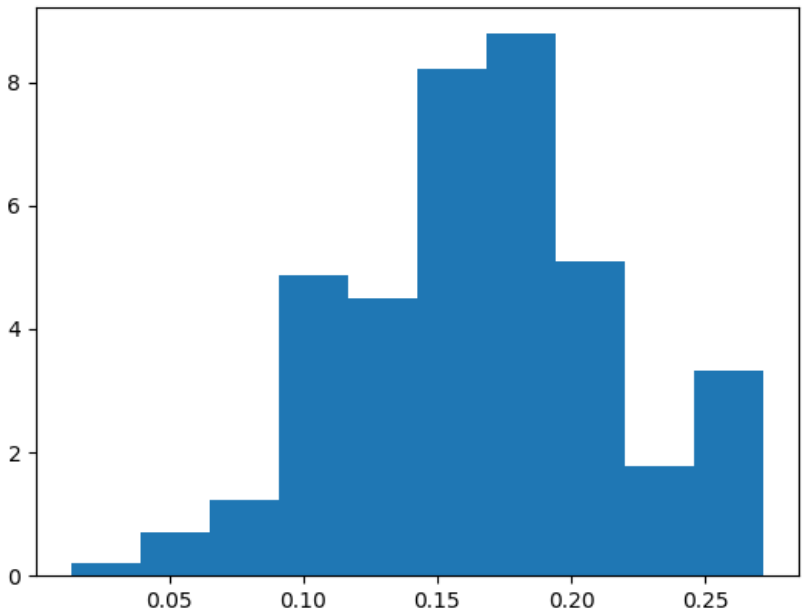
5. november 2017

1 Izbrani jeziki

Izbrani so bili naslednje jeziki : bolgarščina, češčina, nizozemščina, angleščina, finščina, francoščina, nemščina, grščina, italijanščina, japonščina, norveščina, portugalščina, romunščina, ruščina, slovaščina, slovenščina, španščina, srbo-hrvaščina, švedščina, ukrajinščina.

Pri predobdelavi besedila sem vse črke predstavil v mali tiskani obliki ter odstranil naslednje znake .!?, ;_ 0123456789().

2 Rezultati razvrščanja



Slika 1: Porazdelitev silhuet

Silhueta:	0.272
Skupina 1:	bolgarščina, češčina, ruščina, slovaščina, slovenščina, srbo-hrvaščina, ukrajinščina
Skupina 2:	grščina
Skupina 3:	angleščina, francoščina, italijanščina, portugalščina, romunščina, španščina
Skupina 4:	japonščina
Skupina 5:	nizozemščina, finščina, nemščina, norveščina, švedščina

Tabela 1: Največja silhueta

Silhueta:	0.01
Skupina 1:	bolgarščina, japonščina, ruščina, slovenščina, srbo-hrvaščina, ukrajinščina
Skupina 2:	angleščina, švedščina
Skupina 3:	češčina, slovaščina
Skupina 4:	nizozemščina, grščina, italijanščina, portugalščina, romunščina
Skupina 5:	finščina, francoščina, nemščina, norveščina, španščina

Tabela 2: Najmanjša silhueta

Ko opazujemo rezultate največje silhuete vidimo, da skupina 1 vsebuje slovanske jezike, skupina 5 germanske in skupina 3 romanske z izjemo angleščine. Razlog za to bi lahko pripisali podobnosti s francoščino (angleščina ima najmanjšo razdaljo do nje), saj imata Velika Britanija in Francija bogato skupno zgodovino. Rezultati z najmanjšo silhueto so manj smiselni.

3 Napovedovanje jezika

Naj bo D_i razdalja v seznamu razdalj z dolžino N. Vsako razdaljo v seznamu sem potenciral za faktor -10 ($K_i = D_i^{-10}$). Vsak element v novem seznamu, je predstavljal delež od 100% ($\sum_{i=1}^{N} K_i = 1$).

Ime besedila	Jezik	Napovedi
czc1.txt	češčina	češčina(19.9%), slovaščina(15.7%), ruščina(9.3%)
czc2.txt	češčina	grščina(26.2%), češčina(12.3%), slovaščina(10.0%)
eng1.txt	angleščina	angleščina(61.4%), francoščina(5.3%), španščina(4.7%)
eng2.txt	angleščina	angleščina(74.79%), norveščina(3.58%), češčina(2.37%)
fra.txt	francoščina	francoščina(86.1%), španščina(6.8%), portugalščina1.5%)
ger.txt	nemščina	nemščina(75.9%), nizozemščina(8.4%), norveščina(3.5%)
ita1.txt	italijanščina	italijanščina(29.7%), portugalščina(10.8%), bulgarščina(9.8%)
ita2.txt	italijanščina	italijanščina(47.0%), romunščina(10.9%), portugalščina(5.5%)
rus.txt	ruščina	ruščina(24.6%), ukrajinščina(15.3%), srbo-hrvaščina(10.0%)
spn.txt	španščina	španščina(31.1%), italijanščina(14.2%), portugalščina(10.2%)

Tabela 3: Napovedovanje besedil

Ime besedila	Odlomek
czc1.txt	"Když se Řehoř Samsa jednou ráno probudil z nepokojných snů, shledal, že se v posteli proměnil v jakýsi nestvůrný hmyz."
czc2.txt	"Uslyšel za sebou tiché kroky. To nevěštilo nic dobrého. Kdo by ho mohl sledovat tak pozdě v noci, v tomhle zapomenutém koutě města?"
eng1.txt	"One morning, when Gregor Samsa woke from troubled dreams, he found himself transformed in his bed into a horrible vermin."
eng2.txt	"Far far away, behind the word mountains, far from the countries Vokalia and Consonantia, there live the blind texts."
fra.txt	"Loin, très loin, au delà des monts Mots, à mille lieues des pays Voyellie et Consonnia, demeurent les Bolos Bolos."
ger.txt	"Jemand musste Josef K. verleumdet haben, denn ohne dass er etwas Böses getan hätte, wurde er eines Morgens verhaftet."
ita1.txt	"Gregorio Samsa, svegliandosi una mattina da sogni agitati, si trovò trasformato, nel suo letto, in un enorme insetto immondo. "
ita2.txt	"La mia anima è pervasa da una mirabile serenità, simile a queste belle mattinate di maggio che io godo con tutto il cuore."
rus.txt	"Проснувшись однажды утром после беспокойного сна, Грегор Замза обнаружил, что он у себя в постели превратился в страшное насекомое."
spn.txt	"Una mañana, tras un sueño intranquilo, Gregorio Samsa se despertó convertido en un monstruoso insecto."

Tabela 4: Odlomki iz besedil

4 Hierarično razvrščanje

Skupina 1:	bolgarščina, češčina, ruščina, slovaščina, slovenščina, srbo-hrvaščina, ukrajinščina
Skupina 2:	nizozemščina, finščina, nemščina, norveščina, švedščina
Skupina 3:	angleščina, francoščina, italijanščina, portugalščina, romunščina, španščina
Skupina 4:	grščina
Skupina 5:	japonščina

Tabela 5: Rezultati ob uporabi metode hierarhičnega razvrščanja.

Rezultati hierarhičnega razvrščanja se ujemajo z rezultati v Tabeli 5.

5 Rezultati na besedilih iz novičarskih strani

Silhueta:	0.223
Skupina 1:	finščina
Skupina 2:	francoščina, italijanščina, portugalščina, romunščina, španščina
Skupina 3:	nizozemščina,nemščina, norveščina, švedščina
Skupina 4:	angleščina
Skupina 5:	bolgarščina, češčina, grščina, japonščina, ruščina, slovaščina, slovenščina, srbo-hrvaščina, ukrajinščina

Tabela 6: Največja silhueta

Silhueta:	-0.01
Skupina 1:	angleščina, švedščina
Skupina 2:	nizozemščina, finščina, norveščina, portugalščina, romunščina, španščina
Skupina 3:	bolgarščina, italijanščina
Skupina 4:	francoščina, grščina, slovaščina, slovenščina, srbo-hrvaščina, ukrajinščina
Skupina 5:	češčina, nemščina, japonščina, ruščina

Tabela 7: Najmanjša silhueta

Rezultati z najboljšo silhueto so smiselni, vendar ne tako kot ob analizi besedil Splošne deklaracije človeških pravic. Skupina 1 predstavlja romanske jezike, skupina 3 germanske in skupina 5 slovanske jezike. Japonščina je bila uvrščena med slovanske jezike, najbolj bi bilo smiselno, če bi bila v svoji skupini. Zanimivo se je grščina uvrstila tudi med slovanske jezike. Zanimivo je tudi je angleščina osamelec.