Izbrani jeziki. Jeziki, ki sem jih v nalogi uporabil so:

- **slovanski:**
  slovenski (slv), srbski (src3), bosanski (src1), ruski (rus), makedonski (mkj)
- **germanski:**
  nemški (ger), angleški (eng), nizozemski (dut), norveški (nrr)
- **romanski:**
  francoski (frn), španski (spn), italjanski (itn), romunski (rum), portugalski (por)
- **ostali:**
  madžarski (hng), grški (grk), nigerijski (pcm), kitajski (chn), japonski (jpn), svahili (swa).

Splošne deklaracije človekovih pravic v navedenih jezikih najdemo v direktoriju "drzave".

Datoteke sem obdelal tako, da sem vse vrstice združil v eno in jih ločil s presledki. Tekst sem nato spremenil v same male črke, saj se lahko beseda pojavi na začetku stavka, čeprav se je načeloma ne piše z veliko začetnico. Vse skupaj sem shranil v slovar, katerega ključi so imena datotek (brez končnic).

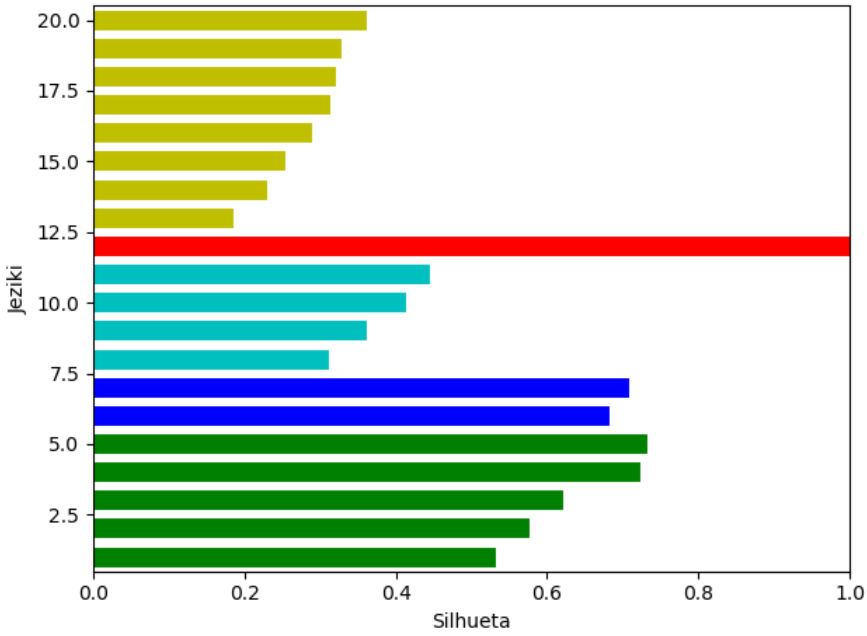
```
corpus = {}
for file_name in glob.glob("drzave/*"):
    text = "␣".join([line.strip() for line in open(file_name, "rt",
                     encoding="utf8").readlines()])
    text = text.lower()
    name = os.path.splitext(os.path.basename(file_name))[0]
    corpus[name] = unidecode(text)
```

Tekst sem nato "razrezal"na vse možne trojke in s pomočjo funkcije Counter ugotovil kolikokrat se vsaka trojka v besedilu pojavi. To sem kasneje rabil, da sem izračunal kosinusno razdaljo med jeziki.

Rezultati razvrščanja. Pri razvrščanju z najboljšo silhueto 1 so države razvrščene po skupinah, ki jih lahko razločimo iz histograma:

- **rumeno:**
  portugalski, španski, romunski, francoski, angleški, italjanski, grški, nigerijski
- **rdečo:**
  svahili
- **svetlo modro:**
  nizozemski, nemški, norveški, madžarski
- **temno modro:**
  kitajski, japonski

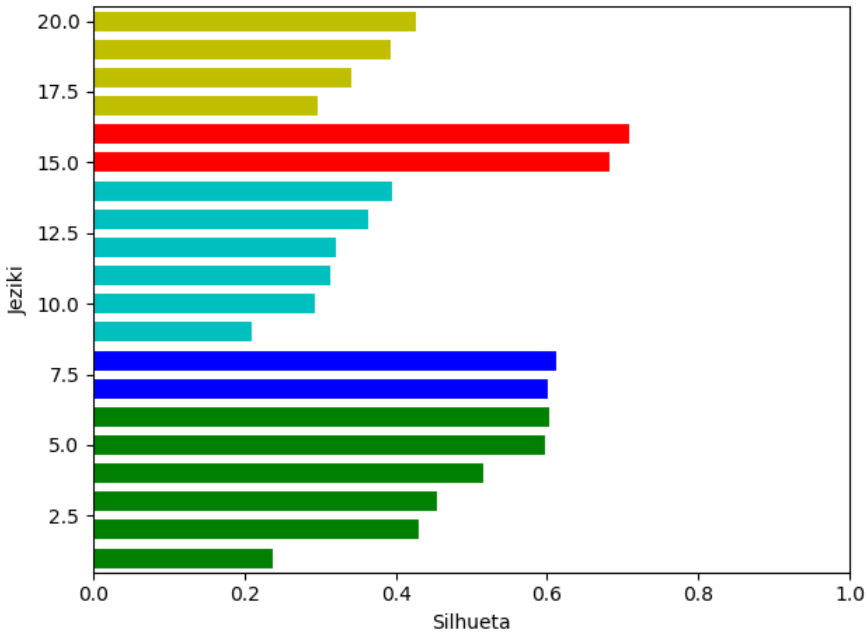
- **zeleno:**
  makedonski, ruski, slovenski, bosanski, srbski



Slika 1: Razvrščanje z najboljšo silhueto.

Pri razvrščanju z najslabšo silhueto 2 pa so države razvrščene po sledečih skupinah:

- **rumeno:**
  nizozemski, nemški, norveški, madžarski
- **rdečo:**
  kitajski, japonski
- **svetlo modro:**
  portugalski, španski, romunski, francoski, italjanski, grški
- **temno modro:**
  angleški, nigerijski
- **zeleno:**
  makedonski, ruski, slovenski, bosanski, srbski, svahili



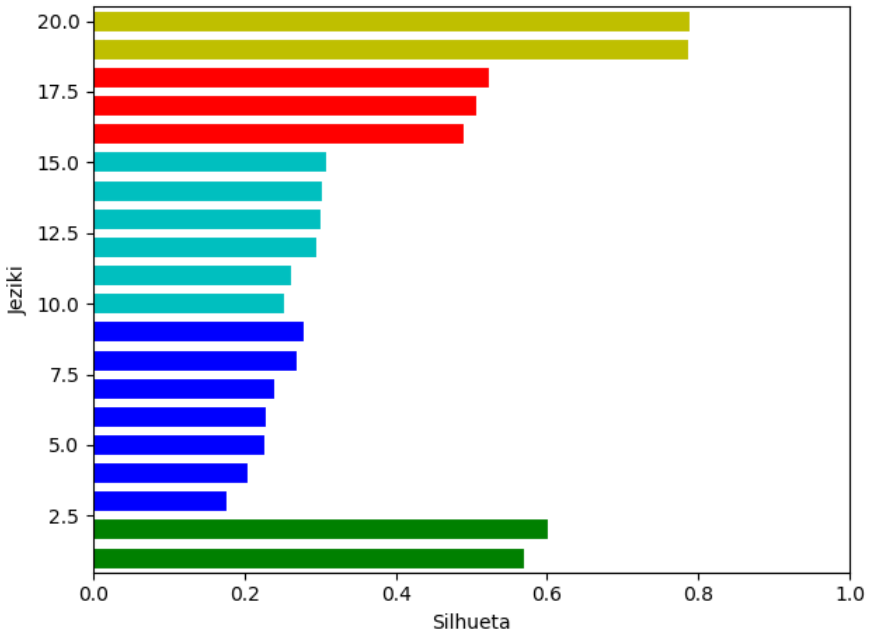
Slika 2: Razvrščanje z najslabšo silhueto.

Moj nabor jezikov je precej raznolik, kljub temu pa so nekateri jeziki v obeh primerih skupaj v skupini, kot naprimer vsi slovanski, romanski in oba azijska jezika. Glavna razlika se pokaže pri jezkih, ki so precej drugačni od drugih (svahili) in pri jezikih, ki so si podobni le z enim (nigerijski z angleškim). Tako v prvem primeru svahili tvori svojo skupino, v drugem primeru pa se pridruži slovanskim, medtem ko v tem primeru svojo skupino tvorita angleški in nigerijski. Rezultati so bili na splošno precej smiselni.

Dodatno (+10%):

Analizo sem ponovil na novičarskih straneh z istim naborom jezikov. Pri najboljši silhueti 3 so bili rezultati precej smiselni:

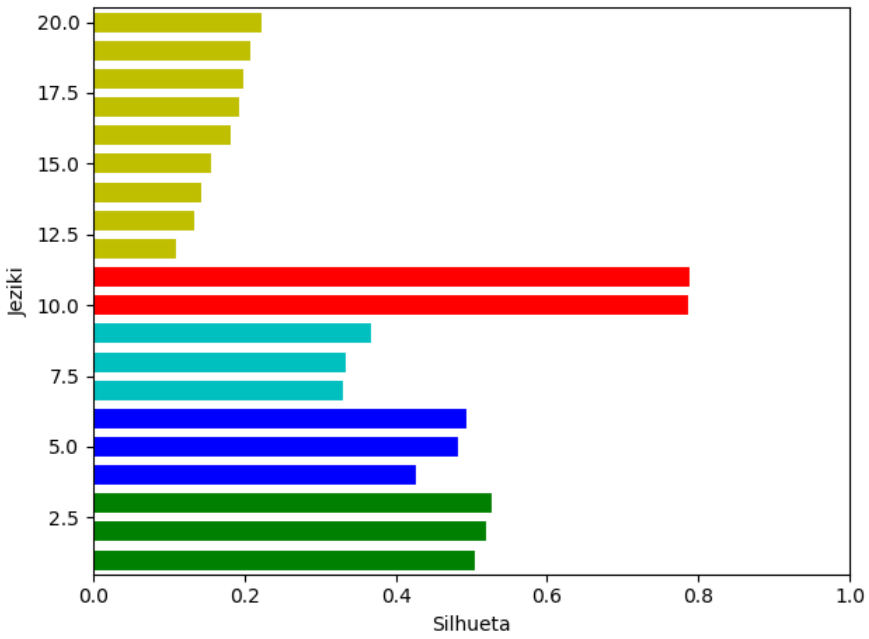
- še vedno sta bili skupaj kitajska in japonska (rumena),
- skupaj so bili tudi nekateri germanski jeziki - nizozemski, nemški in norveški (rdeča),
- slovanski in svahili (svetlo modra),
- vsi romanski, grški in madžarski (temno modra) in
- angleški in nigerijski jezik (zelena).



Slika 3: Razvrščanje z najboljšo silhueto (dodatna naloga).

Pri najslabši silhueti 4, pa so bili rezultati manj smisleni:

- vsi romanski, nigerijski, angleški in madžarski (rumena),
- kitajski in japonski (rdeča),
- ruski, makedonski in svahili (svetlo modra),
- bosanksi, srbski in slovenski (temno modra) in
- nizozemski, nemški in norveški (zelena).



Slika 4: Razvrščanje z najslabšo silhueto (dodatna naloga).

Rezultati niso bili tako smiselni kot pri analizi človekovih pravic, saj je bil tukaj kontekst precej drugačen (nanj nisem bil preveč pozoren, le kopiral sem novice iz interneta), to se je opazilo predvsem v slabši silhueti, kjer so bile jezikovne skupine bolj razdrobljene.

Napovedovanje jezika. Na začetku sem vsak odlomek obdelal na enak način kot v prvem delu naloge (razdelil tekst na trojke, male črke...). Ko sem imel vsak odlomek obdelan na tak način, da sem ga primerjal z vsakim jezikom s pomočjo kosinusne razdalje. Shranil sem najbližje tri jezike in jih izpisal v datoteko "rez_odstavki.txt". Funkcija, ki izpiše rezultate se imenuje "ugotoviJezik", odlomki, ki sem jih uporabljal pa so na voljo v mapi "odstavki". Rezultati so izpisani v tabeli 2. Vsi jeziki so bili prailno napovedani, le v bosanskem odlomku (src1) je za 1% bolj verjeten srbski jezik (src3).

Tabela 1: Tabela prikazuje napovedovanje jezika različnih odlomkov

Jezik odlomka	Odlomek	Napoved verjetnosti
Angleški (eng)	World Mental Health Day encourages us to be more aware of both our own mental health and other people's.	eng z verjetnostjo 0.28 pcm z verjetnostjo 0.18 nrr z verjetnostjo 0.14
Francoski (frn)	a annoncé dans une dans une vidéo diffusée sur les réseaux sociaux qu'il ne prendrait pas parti	frn z verjetnostjo 0.30 itn z verjetnostjo 0.21 por z verjetnostjo 0.18
Nemški (ger)	Zudem muss ein rechtsgerichtetes Internetportal mit antisemitischen Inhalten offenbar offline gehen.	ger z verjetnostjo 0.34 dut z verjetnostjo 0.25 eng z verjetnostjo 0.17
Madžarski (hng)	A tárcavezető elmondta, hogy a még hatékonyabb kiközvetítés érdekében a kormány külön programot	hng z verjetnostjo 0.27 grk z verjetnostjo 0.15 por z verjetnostjo 0.14
Norveški (nrr)	For ham har Begbyåsen nærmest blitt babyen. Han har fulgt prosjektet helt fra den spede begynnelsen.	nrr z verjetnostjo 0.28 dut z verjetnostjo 0.24 ger z verjetnostjo 0.18
Portugalski (por)	Pelo mundo são inúmeros movimentos separatistas que existem e no sul do Brasil destaca-se o movimento	por z verjetnostjo 0.33 spn z verjetnostjo 0.23 itn z verjetnostjo 0.17
Slovenksi (slv)	Zahtevajo tudi razumljive plačilne sisteme in višja denarna nadomestila za delo v težkih pogojih.	slv z verjetnostjo 0.26 itn z verjetnostjo 0.20 src3 z verjetnostjo 0.19
Španski (spn)	Ese aparente distanciamiento de las mujeres respecto al Partido Republicano se refleja también en la	spn z verjetnostjo 0.38 por z verjetnostjo 0.32 frn z verjetnostjo 0.27
Bosanksi (src1)	Nakon što je grupa migranata ušla u Hrvatsku, žene su zapele na nepristupačnom terenu, te se u akciju morao uključiti i HGSS,	src3 z verjetnostjo 0.27 src1 z verjetnostjo 0.26 mkj z verjetnostjo 0.23
Svahili (swa)	Ushindi huo wa Sulle ulipokewa kwa furaha kubwa kwa wadau wa riadha na viongozi wa mchezo huo kwa sababu ni muda mrefu wanariadha	swa z verjetnostjo 0.44 itn z verjetnostjo 0.10 mkj z verjetnostjo 0.09

Izjava o izdelavi domače naloge. Domačo nalogo in pripadajoče programe sem izdelal sam.