Uvod. V sklopu druge domače naloge sem implementiral razvrščanje na osnovi medoidov oz. *k-medoid clustering*. Cilj naloge je razvrščanje jezikov z uporabo Splošne deklaracije človekovih pravic v različnih jezikih za merjenje razdalj med izbranimi jeziki. Uporabil sem kosinusno razdaljo 1 s primerjavo frekvenc trojk sosednjih črk, ki so se pojavile v prej omenjenih besedilih. Implementiral sem tudi postopek za napovedovanje jezika v katerem je napisano poljubno besedilo, prav tako na podlagi kosinusne razdalje. Razvrščanje jezikov sem za dodatno nalogo ponovil še z uporabo novic v različnih jezikih.

Izbrani jeziki. Na voljo je bilo 263 jezikov v datotekah v mapi *ready*, izmed katerih sem izbral 24 večinoma indo-evropskih jezikov. To so: slovenščina, bosanščina, srbščina, ruščina, ukrajinščina, poljščina, češčina, litvanščina, latvijščina, estonščina, kitajščina, madžarščina, grščina, francoščina, italijanščina, romunščina, španščina, portugalščina, angleščina, nemščina, nizozemščina, švedščina, norvežanščina in finščina. Med izbranimi jeziki so tudi taki, ki ne uporabljajo latinice, temveč imajo svojo abecedo (npr. grščina, ruščina, kitajščina). Da bi ocenil natančost razvrščanja sem vključil več jezikov iz različnih jezikovnih skupin in podskupin, nabor jezikov pa vsebuje tudi bolj samostojne ali drugačne jezike kot npr. madžarščina in kitajščina.

Ker so besedila v različnih jezikih in zato tudi različnih abecedah uporabimo transliteracije s knjižnico *unidecode*. Dobljeno besedilo obdelamo tako, da ga v celoti pretvorimo v male črke ter odstranimo znake za nove vrstice in odvečne presledke. Za potrebe algoritma k-medoidov podatke iz besedil shranimo v obliki slovarja trojk sosednjih črk in njihovih frekvenc za vsak posamezen jezik.

Rezultati razvrščanja. Rezultat razvrščanja (k=5) z *najboljšo* silhueto = 0.245 (medoid skupine je naveden prvi):

- latvijščina, litvanščina
- srbščina, češčina, poljščina, ruščina, slovenščina, bosanščina, ukrajinščina
- španščina, angleščina, francoščina, italjanščina, portugalščina, romunščina
- finščina, estonščina, grščina, madžarščina
- nizozemščina, kitajščina, nemščina, švedščina

Rezultat razvrščanja (k=5) z *najslabšo* silhueto = -0.016 (medoid skupine je naveden prvi):

- bosanščina

$$dist\,(X,Y) = 1 - \frac{X \cdot Y}{\|X\|\|Y\|} = 1 - \frac{\sum_{i=1}^{m} X_i \times Y_i}{\sqrt{\sum_{i=1}^{m} (X_i^2)} \times \sqrt{\sum_{i=1}^{m} (Y_i^2)}} \tag{1}$$

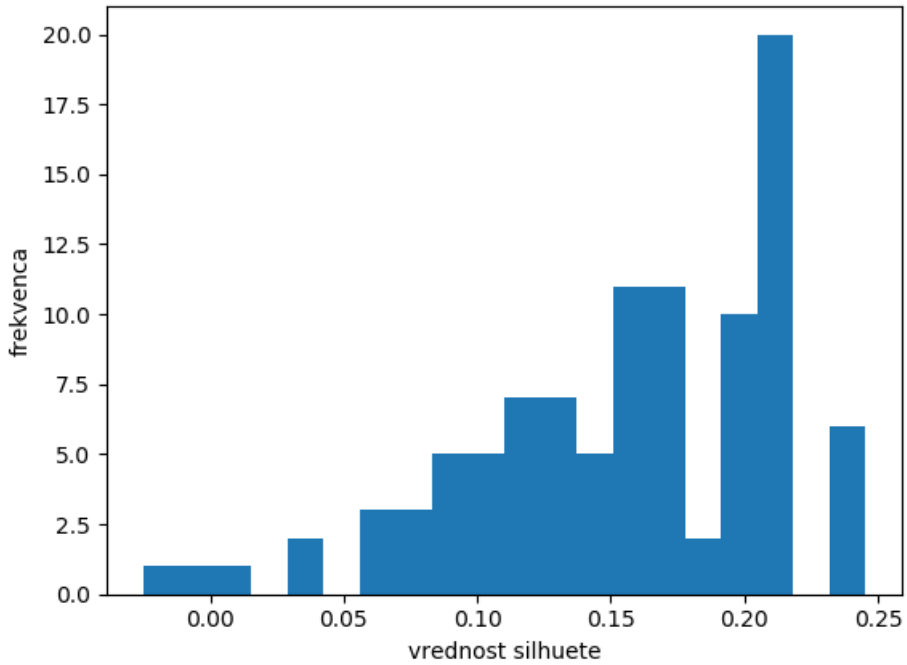
Slika 1: Formula za kosinusno razdaljo.

$$p_i = \frac{\frac{1}{dist_i^2}}{\sum_{i=1}^{m} \left(\frac{1}{dist_i^2} \right)} \tag{2}$$

Slika 2: Transformacija razdalje v verjetnost.

- švedščina, angleščina, estonščina, norveščina
- srbščina, češčina, litvanščina, poljščina, ruščina, slovenščina, ukrajinščina
- nizozemščina, kitajščina, nemščina, madžarščina
- španščina, francoščina, grščina, italjanščina, latvijščina portugalščina, romunščina

Rezultat razvrščanja z najboljšo silhueto določi smiselne skupine, v katerih lahko prepoznamo osnovne jezikovne skupine kot so slovanski, romanski in germanski jeziki. Prav tako so pravilno v isto skupino razvrščeni nekoliko bolj posebni jeziki kot sta latvijščina in litvanščina (baltska jezika), ter tudi finščina in estonščina (sorodna jezika znotraj skupine uralskih jezikov), ki sta grupirana skupaj z še enim ne-indoevropejskim jezikom, madžarščino. Manj smiselno pa je, da kitajščina ni v svoji skupini ampak se priključi skupini germanskih jezikov, angleščina pa je del skupine romanskih jezikov. Priložen je tudi histogram 3 porazdelitve vrednosti silhuet pri različno inicializiranih začetnih voditeljih.



Slika 3: Histogram porazdelitve vrednosti silhuet (bins = 20).

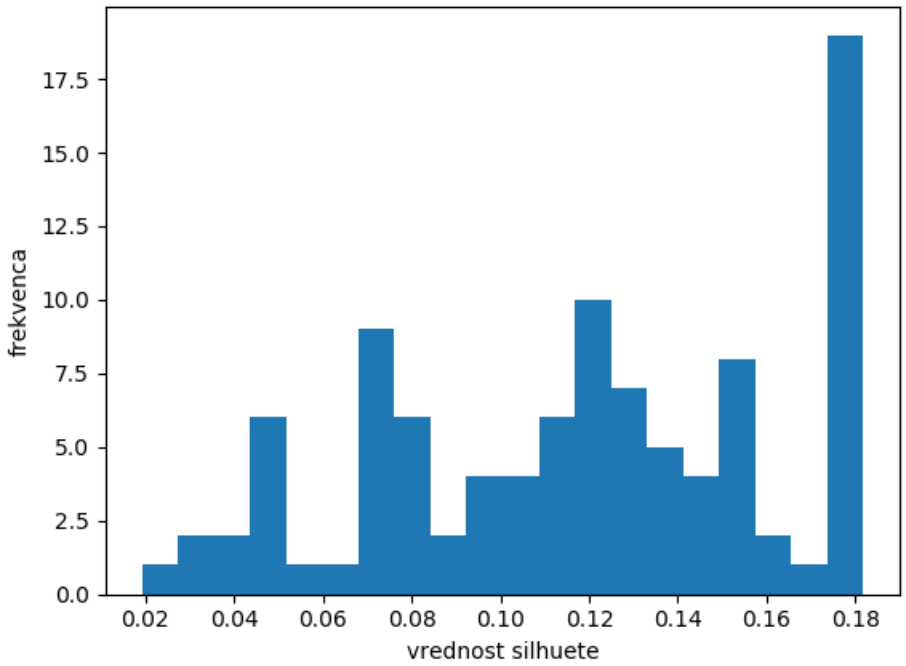
V sklopu dodatne naloge sem ponovil **razvrščanje na novicah** v različnih jezikih, pri tem

sem izbral podmnožico 20 držav izmed prvotnih štirindvajsetih saj sem si tako olajšal iskanje enakomerno dolgih člankov

Rezultati razvrščanja **novic** z *najboljšo* silhueto = 0.182:

- litvanščina, grščina, latvijščina
- finščina, estonščina
- norveščina, nizozemščina, angleščina, nemščina, madžarščina, švedščina
- srbščina, bosanščina, ruščina, slovenščina, ukrajinščina
- španščina, francoščina, italjanščina, portugalščina

Tudi na novicah metoda vrne smiselne rezultate, le z nekoliko nižjo silhueto (predvidevam zaradi večje raznolikosti besedil). Za razliko od razvrščanja na podlagi besedil Deklaracije človekovih pravic, je pri novicah angleščina pravilno v skupini germanskih jezikov. Spodaj prilagam še histogram 4 ocen silhuet za razvrščanje z novicami.



Slika 4: Histogram porazdelitve vrednosti silhuet (bins = 20) - novice.

Napovedovanje jezika. Napovedovanje jezika sem implementiral v metodi *recognize_language()* tako, da sem poljubno besedilo v enem izmed jezikov obdelal na enak način kot besedila iz deklaracij človekovih pravic in ga nato primerjal z vsemi vnosi za posamezen jezik v predpripravljenem slovarju frekvenc trojic črk. Uporabil sem kosinusno razdaljo in razdaljo do vseh jezikov

transformiral v verjetnosti, z uporabo tehnike *inverse distance weighting* (implementacija 2) - pri tem sem uporabil kvadrat razdalje zato, da sem dodatno utežil porazdelitev verjetnosti tako, da so bile verjetnosti za najbolj podobne jezike še višje od manj podobnih. Tako dobljene verjetnosti ustrezajo pogoju $0 \leq p_i \leq 1$ in $\Sigma \mathrm{p}_i = 1$. Da bi metoda vračala še bolj točne napovedi in verjetnosti bi lahko razširil obseg referenčnih besedil za posamezni jezik, toda metoda vrača dovolj točne napovedi že samo z uporabo besedil deklaracij človekovih pravic.

Spodaj so navedeni kratki besedilni odlomki iz zgodbe o Babilonskem stolpu (kratka svetopisemska zgodba o nastanku svetovnih jezikov) in rezultati metode za napovedovanje jezika. Besedila se nahajajo v mapi *test-recognize* in so pridobljena iz spletne strani Omniglot.

1. *slovenščina* - Vsa zemlja je imela en sam jezik in isto govorico. Ko so se ljudje odpravili od vzhoda, so našli ravnino v šinárski deželi in se tam naselili. Rekli so drug drugemu: »Dajmo, delajmo opeko in jo žgimo v ognju!« Opeko so uporabljali namesto kamna in zemeljsko smolo namesto malte. ...

    - slovenščina - 9.40%
    - bosanščina - 5.70%
    - srbščina - 5.65%

2. *ruščina (originalno besedilo je v cirilici)* - Na vsej zemlje Byl odin jazyk i odno narječje. Dvinuvšis' s vostoka, oni našli v zjemlje Sjennaar ravninu i poselilis' tam. I skazali drug drugu: nadjelajem kirpičjej i obožžem ognjem. I stali u inh kirpiči vmjesto kamnjej, a zjemljanaja smola vmjesto izvjesti. ...

    - ruščina - 7.49%
    - ukrajinščina - 5.97%
    - srbščina - 5.96%

3. *španščina* - En ese entonces se hablaba un solo idioma en toda la tierra. Al emigrar al oriente, la gente encontró una llanura en la región de Sinar, y allí se asentaron. Un día se dijeron unos a otros: «Vamos a hacer ladrillos, y a cocerlos al fuego.» Fue así como usaron ladrillos en vez de piedras, y asfalto en vez de mezcla. ...

    - španščina - 16.76%
    - portugalščina - 7.91%
    - francoščina - 6.07%

4. *francoščina* - La terre entière se servait de la même langue et des mêmes mots. Or en se déplaçant vers l'orient, les hommes découvrirent une plaine dans le pays de Shinéar et y habitèrent. Ils se dirent l'un à l'autre: «Allons! Moulons des briques et cuisons-les au four». Les briques leur servirent de pierre et le bitume leur servit de mortier. ...

    - francoščina - 21.88%

- španščina - 7.20%
- portugalščina - 4.81%

5. *nemščina* - Es hatte aber alle Welt einerlei Zunge und Sprache. Als sie nun nach Osten zogen, fanden sie eine Ebene im Lande Schinar und wohnten daselbst. Und sie sprachen untereinander: Wohlauf, laßt uns Ziegel streichen und brennen! - und nahmen Ziegel als Stein und Erdharz als Mörtel ...

   - nemščina - 21.69%
   - nizozemščina - 6.51%
   - norveščina - 4.90%

6. *angleščina* - Now the whole earth had one language and the same words. And as people migrated from the east, they found a plain in the land of Shinar and settled there. And they said to one another, Čome, let us make bricks, and burn them thoroughly."And they had brick for stone, and bitumen for mortar. ...

   - angleščina - 22.60%
   - grščina - 4.19%
   - norveščina - 4.01%

7. *grščina (originalno besedilo je v grški pisavi)* - Ke ito pasa i yi mias glossis ke mias fonis. Ke ote ekinisan apo tis anatolis, evron pediada en ti yi Sennaar; ke katokisan eki. Ke ipen o is pros ton allon, Elthete, as kamomen plinthus, ke as psisomen aftas en piri; ke exrisimefsen is aftus i men plinthos adi petras, i de asfaltos exrisimefsen is aftus adi pilu. ...

   - grščina - 20.64%
   - finščina - 4.87%
   - litvanščina - 4.72%

Vidimo, da metoda za napovedovanje jezika vrača smiselne rezultate in je v mojih testiranjih vsakič pravilno napovedala jezik. Verjetnost napovedi za najverjetnejši jezik je načeloma opazno večja, kot za drugi najbližji jezik - samo v primeru slovanskih jezikov (npr. slovenščina in ruščina) so te verjetnosti bolj enakomerno razporejene med najbolj ustreznimi jeziki saj so slovanski jeziki med seboj precej podobni.

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