Podobnost jezikov

Matea Lenček (63140392)

5. november 2017

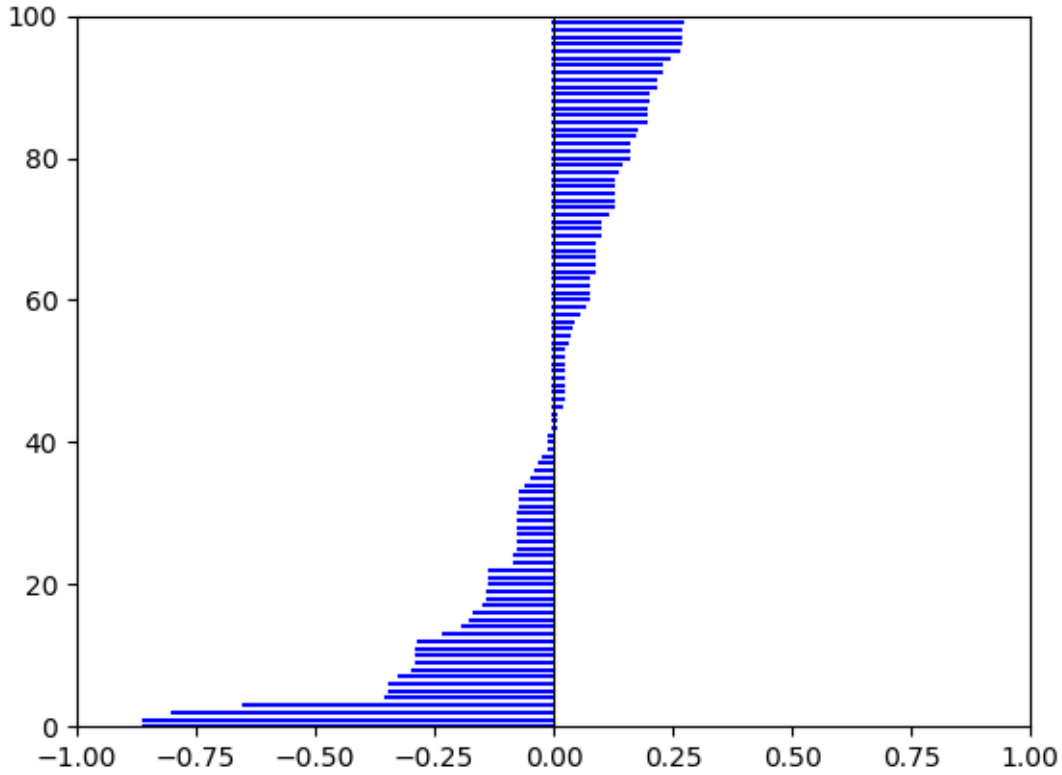
1 Izbrani jeziki

Izbrala sem naslednje jezike: bosanščina (latinica), bolgarščina, češčina, danščina, angleščina, španščina, estonščina, finščina, francoščina, nemščina, grščina, madžarščina, italijanščina, makedonščina, norveščina, poljščina, portugalščina, romunščina, srbščina (latinica), srbščina (cirilica), slovaščina in slovenščina. V datotekah sem nadomestila nove vrstice s presledki, nato sem dvojne presledke (ki so se pojavili pri nekaterih od teh zamenjav) zamenjala z enojnimi presledkomi in vse črke pretvorila v male črke. Besedilo sem potem še transliterirala z uporabo knjižnice unidecode. Iz obdelanih besedil sem nato tvorila trojke sosednjih črk.

2 Rezultati razvrščanja

Na sliki 1 je prikazana porazdelitev vrednosti silhuet.
Rezultati razvrščanja z najboljšo in najslabšo silhueto:

- Najboljša silhueta (0.272):
  - srbščina (latinica), makedonščina, srbščina (cirilica), bosanščina (latinica), češčina, slovaščina, poljščina, bolgarščina, slovenščina
  - španščina, italijanščina, portugalščina, romunščina, francoščina
  - finščina, estonščina, grščina
  - danščina, nemščina, angleščina, norveščina
  - madžarščina
- Najslabša silhueta (-0.859):
  - srbščina (latinica), srbščina (cirilica), češčina, slovaščina, slovenščina
  - španščina, grščina, italijanščina, portugalščina, danščina, nemščina, angleščina, norveščina, romunščina, francoščina, madžarščina
  - makedonščina, finščina, estonščina, bulgarščina
  - poljščina
  - bosanščina (latinica)



Slika 1: Porazdelitev vrednosti silhuet.

Smiselnost rezultatov je pri tej metodi razvrščanja v skupine je seveda zelo odvisna od naključnega izbora začetnih medoidov, zato silhuete tako močno variirajo. Rezultati z najboljšo silhueto se razdelijo v znane skupine jezikov (slovanski, romanski, germanski). Pri rezultatih z najslabšo silhueto pa temu ni tako in razlog za to lahko hitro najdemo v izboru začetnih medoidov, ki so bile v tem primeru datoteke v naslednjih jezikih: poljščina, bulgarščina, srbščina (latinica), bosanščina (latinica) in angleščina.

3 Napovedovanje jezika

Datoteko sem predobdelala na enak način kot sem navedla v prvem razdelku. Za dano besedilo sem napovedala jezik tako, da sem najprej izračunala kosinusne podobnosti besedila z vsemi primeri jezikov. Izbrala sem tri jezike s katerimi je imelo besedilo največjo podobnost in nato izračunala kakšna pripadnost vsakemu izmed teh treh jezikov, tako da sem kosinusno podobnost vsakega od teh treh delila z vsoto vseh treh podobnosti. Zaradi preglednosti poročila odlomkov nisem vključila v tabelo 1, sem jih pa priložila v mapo *sample*, v tabeli pa sem navedla ime datoteke in rezultate napovedovanja.

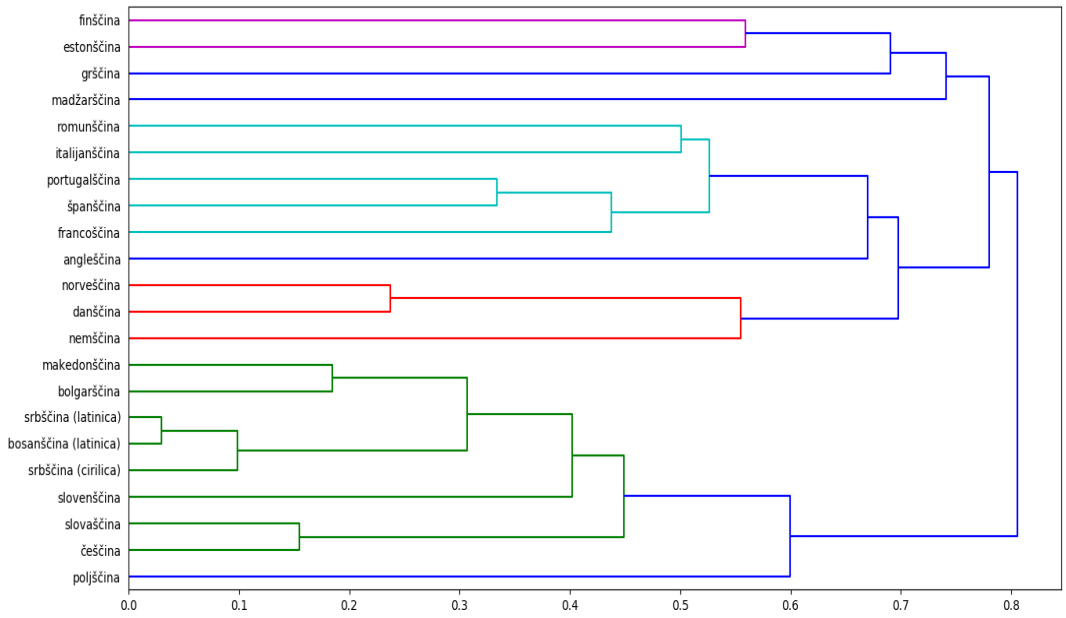
Tabela 1: Rezultati napovedovanja pripadnosti jezika.

napovedovan jezik	ime datoteke z odlomkom	najbolj verjetni jeziki		
španščina	spn.txt	španščina: 0.386	francoščina: 0.311	portugalščina: 0.302
slovenščina	slv.txt	slovenščina: 0.336	srbščina (latinica): 0.332	bosanščina (latinica): 0.332
francoščina	frn.txt	francoščina: 0.428	španščina: 0.328	romunščina: 0.244
nemščina	ger.txt	nemščina: 0.477	danščina: 0.276	norveščina: 0.247
italijanščina	itn.txt	italijanščina: 0.427	romunščina: 0.307	francoščina: 0.266
portugalščina	por.txt	portugalščina: 0.43	španščina: 0.343	francoščina: 0.227
bosanščina (latinica)	src1.txt	bosanščina (latinica): 0.352	srbščina (latinica): 0.34	srbščina (cirilica): 0.308
češčina	czc.txt	češčina: 0.35	slovaščina: 0.333	slovenščina: 0.317
angleščina	eng.txt	angleščina: 0.531	grščina: 0.253	nemščina: 0.215
estonščina	est.txt	estonščina: 0.5	finščina: 0.252	grščina: 0.248

.

4 Hierarhično razvrščanje

Rezultati hierarhičnega razvrščanja so vidni na sliki 2.



Slika 2: Rezultati hierarhičnega razvrščanja.

Prav tako, kot pri rezultatih z najboljšo silhueto razvrščanja s k-medoidi, so tudi pri hierarhičnem razvrščanju vidne znane skupine jezikov (slovanski, germanski, romanski, uralski).

5 Analiza novic

Pri analizi novičarskih strani sem dobila precej podobne rezultate kot pri analizi Splošnih deklaracij o človekovih pravicah. Tudi pri tej analizi so jeziki razvrščeni v znane jezikovne skupine. Pri razvrščanju z najboljšo silhueto, sem dobila naslednje skupine:

- srbščina (latinica), makedonščina, srbščina (cirilica), bosanščina (latinica), češčina, slovaščina, poljščina, bolgarščina, slovenščina
- španščina, italijanščina, portugalščina, romunščina, francoščina
- finščina, estonščina
- danščina, nemščina, angleščina, norveščina
- madžarščina, grščina