Podobnost jezikov

Tilen Venko (63140280)

30. oktober 2016

1 Uvod

Prvi cilj naloge je bil, da zgradimo hierarhijo vsaj dvajsetih jezikov, ki vsebujejo vsaj 2 tuji pisavi in jo komentiramo. Drugi cilj pa je, da na podlagi besedil, ki smo jih obdelovali znamo iz poljubnega besedila v enem od teh jezikov ugotoviti, v kakšnem jeziku je to besedilo in podati, s kakšno verjetnostjo lahko to trdimo.

2 Podatki

Jezike smo pridobili iz spletne strani združenih narodov iz Splošne deklaracije človekovih pravic. Tako je besedilo, ki ga preverjami enako v vseh jezikih, kar naj bo dalo boljše rezultate pri gradnji hierarhije jezikov. Besedila se nahajajo v datoteki `/human_rights/ready/`. Za svojo nalogo sem si izbral 22 evropskih jezikov:

- angleščina
- beloruščina
- bosanščina v cirilici in latinici
- češčina
- danščina
- finščina
- francoščina
- grščina v grški pisavi
- italijanščina
- latinščina
- madžarščina
- makedonščina

- nizozemščina
- nemščina
- norveščina
- polščina
- portugalščina
- ruščina
- slovaščina
- slovenščina
- srbščina v cirilici in latinici
- španščina

Besedila nad katerimi sem preverjal ali program pravilno zazna jezik besedila se nahajajo v mapi `\test`. Besedila sem generiral s spletnim orodjem `http://randomtextgenerator.com/`. Tam sem generiral besedila v angleščini, francoščini, nemščini, grščini v grški pisavi, italijanščini, srbščini v cirilici. S portala MMC rtvslo `https://www.rtvslo.si/slovenija/anja-kopac-mrak-cerar-ni-zahteval-mojega-odstopa/406291` pa sem pridobil članek v slovenščini.

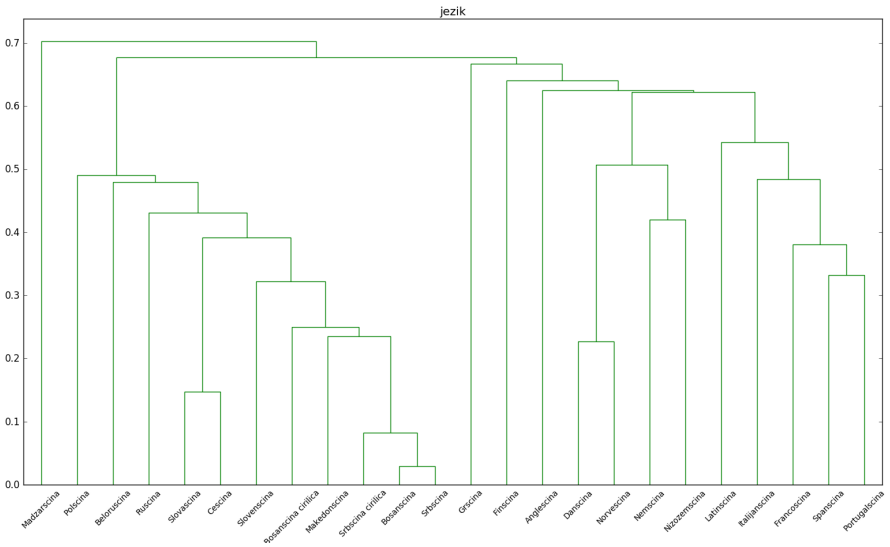
3 Metode

Iz besedil sem najprej poskušal odstranit nepotrebne podatke. Tako sem vse črke spremenil v male, odstranil prehode v novo vrstico in jih nadomestil s presledki. Besedilo sem tudi dekodiral iz unikode, zato da so vsa besedila v enaki pisavi, latinici. Brez tega bi si namreč bosanščina v latinici in cirilici ne bili nič podobni, prav tako pa ne bi mogel grščine primerjati z ostalimi jeziki.
Besedilo sem nato razbil na terke črk, ki so privzeto velikosti 3, in preštel kolikokrat se kakšna terka pojavi. Nad matriko teh terk sem nato klical funkcijo linkage, ki je za parameter "matric", prejela mojo funkcijo cosinus, v kateri sem implementiral metodo za izračun kosinusne razdalje. Nad podatki, ki sem jih dobil iz funkcije linkage sem nato klical še funkcijo dendrogram in izrisal dendrogram hierarhije jezikov.

Drugega dela naloge, kjer sem moral implementirati metodo, s katero sem ugotavljal v kakšnem jeziku je podano besedilo, sem se lotil na zelo podoben način, da sem najprej očistil besedilo enako, kot prej. Nato pa sem za vse jezike od prej in podanim besedilom izračunal kosinusno razdaljo med njimi izmed teh izbral tri z najmanjšo razdaljo in jih vrnil, kot tri najbolj verjetne jezike, v katerem je to besedilo, za najbolj verjeten jezik pa sem podal tudi verjetnost, da je iskano besedilo v tem jeziku.

4 Rezultati

4.1 Hierarhija jezikov



Dendrogram hierarhije jezikov

Podatki v dendrogramu so takšni, kot bi jih pričakovali. slovanski jeziki so blizu skupaj pri čemer lahko ločimo vzhodnoslovanske jezike (slovaščina in češčina) in jugoslovanske jezike, pri čemer pa je polščina izjema. Srbščina in bosanščina v cirilici in latinici nista čisto skupaj zato, ker ko dekodiramo iz unikode, se določeni znaki zapisejo malo drugače, kot pa so zapisano v latinici. Vidimo lahko tudi skupino romanskih in pa germanskih jezikov, ki pa nista tako očitno ločeni med sabo, kot sta od slovanske jezikovne skupine. Opazimo pa lahko tudi, da grščina, finščina in madžarščina ne spadajo v nobeno od teh jezikovnih skupin, pri čemer najbolj izstopa madžarščina.

4.2 Prepoznava jezika iz besedila

Če programu podam generirano besedilo v angleščini mi vrne:

> Besedilo je v jeziku Angleščina z verjetnostjo 38.426% lahko pa je tudi v jezikih Francoščina ali Danščina.

Program nam zna dobro povedati v kakšnem jeziku je besedilo, vendar to lahko trdi z manjšo gotovostjo.

5 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.