Poročilo detekcije jezika na osnovi N-gramov

5. maj 2024

1 Korpus

Uporabljeni korpus, je na voljo na naslednji povezavi: https://www.clarin.si/repository/xmlui/handle/11356/1859#. Gre za večjezikovni korpus parlamentarnih razprav. V tabeli 1 lahko vidimo, koliko datotek je v vsakem jeziku in koliko jih je v učno in testno množico. Tako je celoten korpus sestavljen iz skupaj 7,159 datotek, od katerih jih je bilo 75% dodeljenih v učno množico, preostanek pa v testno množico.

Jezik	Učna množica	Testna množica	Skupaj
Slovenski	1,178	394	1,572
Angleški	1,656	553	2,209
Nemški	915	306	1,221
Španski	337	113	450
Hrvaški	1,280	427	1,707

Tabela 1: Število datotek za vsak jezik v korpusu.

Pred izdelavo profilov kategorij sem korpus predhodno obdelala. Ročno sem izločila samo datoteke .txt za vseh pet jezikov, in ker so bila imena datotek zapisana v vsaki vrstici vsake datoteke, sem jih odstranila z uporabo regexa. Z regexom sem prav tako odstranila številke in ločila, razen apostrofov. Te spremembe so vidne na slika 1.

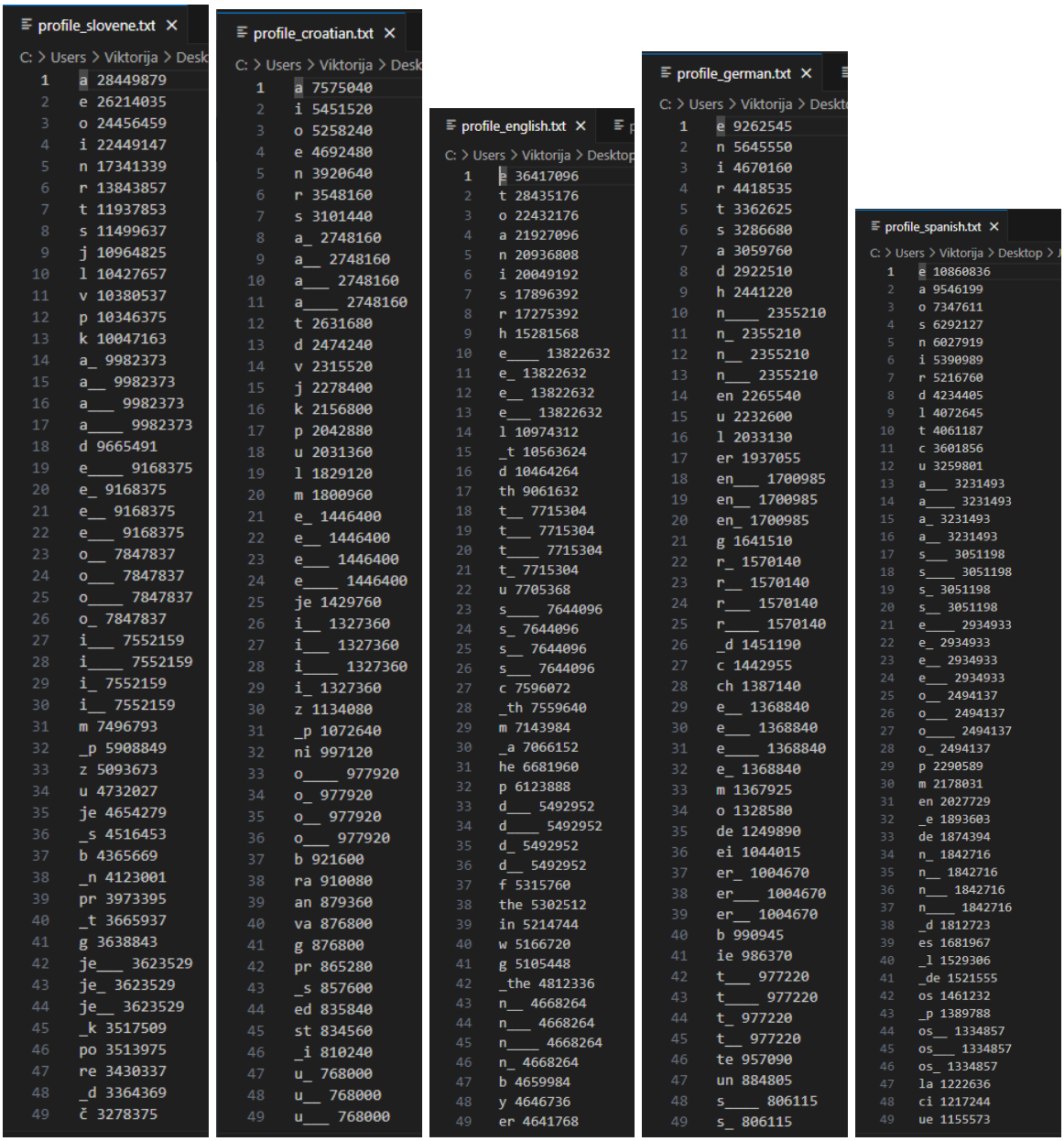
2 Izdelava profilov

Za izdelavo jezikovnih (kategorijskih) profilov sem uporabila datoteke iz učne množice. Iz vsake datoteke sem prebrala vsak žeton (besedo), ga dopolnila s pravilnim številom podčrtank in naredila N-gramov, pri čemer je N = 1..5. Preštela sem vsako ponovitev vsakega N-grama. Rezultate sem razvrstila po padajočem številu in za profil vsakega jezika vzela le 300 najvišjih. Dobljene profile so vidne na slika 2. Ustvarjanje profilov na mojem računalniku je trajalo približno 21 ur.

3 Testiranje klasifikacije za posamezne datoteke

Po izdelavi profila za vsak jezik, sem izbrala naključno datoteko iz testne množice vsakega jezika in zanjo izdelala profil. Vzela sem le prvih 300 žetonov, ki so se najpogosteje pojavljali v dokumentu, in preverila, kje je položaj posameznega žetona v profilih kategorij. Seštela sem razdalje med žetoni z enako vrednostjo v profilu trenutne datoteke in v profilu kategorije, in tako sem dobila oceno za vsako kategorijo. Če je bil v profilu trenutne datoteke žeton, ki ga ni bilo v profilu kategorije, sem k skupnemu rezultatu dodala največjo celoštevilsko vrednost INT_MAX. Tako najmanjša ocena kategorije pomeni, da je med profilom trenutne datoteke in profilom kategorije "najmanjša razdalja". To pomeni, da je jezik trenutne datoteke najbolj podoben tej kategoriji (jeziku). Na sliki 3 si lahko ogledamo rezultate testiranja klasifikacije na datotekah iz različnih kategorij.

Slika 1: Primer datoteke 1. pred predobdelavo; 2. po odstranitvi naslova iz datoteke; 3. po predobdelavi.



Slika 2: Izdelani profili za 1. slovenski, 2. hrvaški, 3. angleški, 4. nemški in 5. španski jezik.

```
PS C:\Users\Viktorija\Desktop\JT\RV3> .\categorization.exe
Please choose one of the following:
1. Build categories' profiles;
2. Classify document;
3. Calculate efficiency.
2
Please enter the path to the document: C:\Users\Viktorija\Desktop\JT\RV3\korpus3\output2\slovene\test\ParlaMint-SI_2017-02-13-SDZ7-Redna-27.txt
profile_croatian 369367193103
profile_english 423054285902
profile_german 397284482357
profile_slovene 328565004814
profile_spanish 405874417329
Most likely language is: profile_slovene
PS C:\Users\Viktorija\Desktop\JT\RV3>
```

```
PS C:\Users\Viktorija\Desktop\JT\RV3> .\categorization.exe
Please choose one of the following:
1. Build categories' profiles;
2. Classify document;
3. Calculate efficiency.
2
Please enter the path to the document: C:\Users\Viktorija\Desktop\JT\RV3\korpus3\output2\croatian\test\ParlaMint-HR_2018-12-07-0.txt
profile_croatian 356482291599
profile_english 425201769732
profile_german 412316867954
profile_slovene 335007455702
profile_spanish 405874417256
Most likely language is: profile_slovene
PS C:\Users\Viktorija\Desktop\JT\RV3>
```

```
PS C:\Users\Viktorija\Desktop\JT\RV3> .\categorization.exe
Please choose one of the following:
1. Build categories' profiles;
2. Classify document;
3. Calculate efficiency.
2
Please enter the path to the document: C:\Users\Viktorija\Desktop\JT\RV3\korpus3\output2\english\test\ParlaMint-GB_2020-10-23-commons.txt
profile_croatian 412316867862
profile_english 352187327170
profile_german 354334810596
profile_slovene 399431965737
profile_spanish 371514678870
Most likely language is: profile_english
PS C:\Users\Viktorija\Desktop\JT\RV3>
```

```
PS C:\Users\Viktorija\Desktop\JT\RV3> .\categorization.exe
Please choose one of the following:
1. Build categories' profiles;
2. Classify document;
3. Calculate efficiency.
2
Please enter the path to the document: C:\Users\Viktorija\Desktop\JT\RV3\korpus3\output2\german\test\ParlaMint-AT_2017-05-17-025-XXV-NRSITZ-00181.txt
profile_croatian 429496737427
profile_english 408021899829
profile_german 311385138189
profile_slovene 425201769545
profile_spanish 416611835527
Most likely language is: profile_german
PS C:\Users\Viktorija\Desktop\JT\RV3>
```

```
PS C:\Users\Viktorija\Desktop\JT\RV3> .\categorization.exe
Please choose one of the following:
1. Build categories' profiles;
2. Classify document;
3. Calculate efficiency.
2
Please enter the path to the document: C:\Users\Viktorija\Desktop\JT\RV3\korpus3\output2\spanish\test\ParlaMint-ES_2022-09-15-CD220915.txt
profile_croatian 418759318247
profile_english 392989516051
profile_german 386547064236
profile_slovene 390842031150
profile_spanish 341449908558
Most likely language is: profile_spanish
PS C:\Users\Viktorija\Desktop\JT\RV3>
```

Slika 3: Rezultati klasifikacije datotek, napisanih v 1. slovenskem, 2. hrvaškem, 3. angleškem, 4. nemškem in 5. španskem jeziku.

4 Testiranje klasifikacije s testnimi množicami

Da bi klasifikacijo testirala na celotni testni množici, sem v kodo dodala še en del, kjer se profili kategorij najprej preberejo iz datoteke in shranijo v program. Nato za vsako datoteko v vsaki testni množici izmerim najmanjšo razdaljo in tako poiščem najbolj podoben jezik za vsako datoteko. V tabeli 2 sem zbrala število pravilnih in nepravilnih klasifikacij.

Jezik	Pravilna klasifikacija	Nepravilna klasifikacija
Slovenski	394	0
Angleški	462	91
Nemški	305	1
Španski	113	0
Hrvaški	0	427

Tabela 2: Število pravilno in nepravilno klasificiranih datotek iz testnih množic.

5 Zaključek

Iz podatkov v tabeli 2 je razvidno, da je ta klasifikacijski model popolnoma neuspešen pri klasifikaciji datotek v hrvaškem jeziku. Vsaka datoteka iz testne množice je klasificirana kot slovenska datoteka, kar je verjetno posledica podobnosti med obema jezikoma. To je mogoče odpraviti tako, da namesto 300 najpogostejših N-gramov v profilih izberemo 400 najpogostejših N-gramov, za katere avtorji članka poročajo, da so jim dali skoraj perfekten rezultat.
V tabeli je treba omeniti še dve zanimivi stvari: datoteke, napisane v angleškem jeziku, so včasih nepravilno klasificirane kot napisane v nemškem jeziku, samo v enem primeru pa je bila nemška datoteka klasificirana kot hrvaška datoteka.