Naredil sem program, ki prepozna jezik dokumenta. Uporablja n-grame (zaporedja znakov dolžine 1 do 5), ki jih izlušči iz besedila. Program ima dva dela:

- učenje (izdelava jezikovnih profilov),
- klasifikacija (ugotavljanje jezika novega dokumenta).

Izgradnja jezikovnih profilov:

Za vsak jezik (angleščina, slovenščina, nemščina, španščina, hrvaščina) sem iz korpusov naredil profil. Profil vsebuje 300 najpogostejših n-gramov.

Najprej sem implementiral funkcijo create_profile, ki je celotno datoteko korupsa prebrala na enkrat, razdelila na tokene, nato pa shranila in preštela vse ngrame za profil.

Drugje v kodi sem imel bug, ampak sem mislil, da je problem tukaj, zato sem naredil še funkcijo train_with_large_corpus, ki bere korpus po delih in je bolj primerna za večje datoteke.

Primer zagona:

python main.py train <JEZIK> --corpus <POT_DO_KORPUSA> --output <POT_DO_IZHODA>

Klasifikacija dokumenta:

Program prebere dokument, iz njega naredi profil in ga primerja z vsemi jezikovnimi profili. Uporabi »out-of-place« metodo. Ta preveri, koliko so n-grami v dokumentu odmaknjeni od tistih v jezikovnem profilu. Jezik z najmanjšo razliko je rezultat.

Primer zagona:

python main.py classify --text <POT_DO_VHODA > --profiles <POT_DO_DIR_PROFILOV >

Testiranje:

Iz wikipedije sem skopiral stran besedila v angleškem in nemškem jeziku. Nisem uporabil isto besedilo kot je bilo v korpusu!

Za posamezni jezik sem zagnal klasifikacijo in dobil naslednji izhod:

py.exe .\main.py classify --text .\german_test.txt --profiles .\Profiles\

Document language: german

Distances from each language profile:

german: 31719

english: 51828

slovenian: 63080

spanish: 63082

croatian: 64032

py.exe .\main.py classify --text .\english_test.txt --profiles .\Profiles\

Document language: english

Distances from each language profile:

english: 31328

spanish: 55060

german: 56895

croatian: 65789

slovenian: 66106