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:

# Šumniki: top 300 ngrami

py.exe .\main.py classify --text .\test\_files\šumniki\_mali.txt --profiles .\Profiles\

Document language: croatian

Distances from each language profile:

croatian: 5700

english: 5700

german: 5700

slovenian: 5700

spanish: 5700

py.exe .\main.py classify --text .\test\_files\šumniki\_veliki.txt --profiles .\Profiles\

Document language: croatian

Distances from each language profile:

croatian: 5700

english: 5700

german: 5700

slovenian: 5700

spanish: 5700

py.exe .\main.py classify --text .\test\_files\šumniki\_oboje.txt --profiles .\Profiles\

Document language: croatian

Distances from each language profile:

croatian: 8400

english: 8400

german: 8400

slovenian: 8400

spanish: 8400

# Šumniki: top 500 ngrami

py.exe .\main.py classify --text .\test\_files\šumniki\_mali.txt --profiles .\Profiles2\

Document language: slovenian

Distances from each language profile:

slovenian: 8056

croatian: 8175

english: 9500

german: 9500

spanish: 9500

py.exe .\main.py classify --text .\test\_files\šumniki\_veliki.txt --profiles .\Profiles2\

Document language: slovenian

Distances from each language profile:

slovenian: 8056

croatian: 8175

english: 9500

german: 9500

spanish: 9500

py.exe .\main.py classify --text .\test\_files\šumniki\_oboje.txt --profiles .\Profiles2\

Document language: slovenian

Distances from each language profile:

slovenian: 12553

croatian: 12672

english: 14000

german: 14000

spanish: 14000

# klasifikacija testnih datotek – top 300 ngrami

datoteka klasificiran jezik

croatian\_test (1).txt - croatian

croatian\_test (10).txt - croatian

croatian\_test (2).txt - croatian

croatian\_test (3).txt - croatian

croatian\_test (4).txt - croatian

croatian\_test (5).txt - croatian

croatian\_test (6).txt - croatian

croatian\_test (7).txt - croatian

croatian\_test (8).txt - croatian

croatian\_test (9).txt - croatian

english\_test (1).txt - english

english\_test (10).txt - english

english\_test (2).txt - english

english\_test (3).txt - english

english\_test (4).txt - english

english\_test (5).txt - english

english\_test (6).txt - english

english\_test (7).txt - english

english\_test (8).txt - english

english\_test (9).txt - english

german\_test (1).txt - german

german\_test (10).txt - german

german\_test (2).txt - german

german\_test (3).txt - german

german\_test (4).txt - german

german\_test (5).txt - german

german\_test (6).txt - german

german\_test (7).txt - german

german\_test (8).txt - german

german\_test (9).txt - german

slovenian\_test (1).txt - slovenian

slovenian\_test (10).txt - slovenian

slovenian\_test (2).txt - slovenian

slovenian\_test (3).txt - slovenian

slovenian\_test (4).txt - slovenian

slovenian\_test (5).txt - slovenian

slovenian\_test (6).txt - slovenian

slovenian\_test (7).txt - slovenian

slovenian\_test (8).txt - slovenian

slovenian\_test (9).txt - slovenian

spanish\_test (1).txt - spanish

spanish\_test (10).txt - spanish

spanish\_test (2).txt - spanish

spanish\_test (3).txt - spanish

spanish\_test (4).txt - spanish

spanish\_test (5).txt - spanish

spanish\_test (6).txt - spanish

spanish\_test (7).txt - spanish

spanish\_test (8).txt - spanish

spanish\_test (9).txt - spanish