Detektor jezika z uporabo n-gram profiliranja

1. Namen in opis programa

Cilj naloge je izdelati program, ki zna na podlagi kratkega besedila prepoznati jezik. Program uporablja metodo profiliranja n-gramov, pri kateri za vsak jezik sestavimo profil najpogostejših zaporedij znakov (n-gramov), nato pa s primerjanjem teh profilov z besedilom ugotovimo najverjetnejši jezik.

Program deluje v dveh glavnih fazah:

- **Učenje:** za vsak jezik iz učnega korpusa izračunamo najpogostejše n-grame.
- **Klasifikacija:** novo besedilo pretvorimo v njegov n-gram profil in izračunamo razdalje do vseh naučenih jezikovnih profilov. Najmanjša razdalja pomeni najbolj verjeten jezik.

2. Podprti jeziki

Model podpira naslednje jezike:

- angleščina (English)
- nemščina (German)
- španščina (Spanish)
- slovenščina (Slovenian)
- hrvaščina (Croatian)
- japonščina (Japanese) – dodana za preizkus z neevropskim jezikom

3. Tehnični opis

Program za detekcijo jezika temelji na metodi primerjanja jezikovnih profilov, ki jih sestavljajo najpogostejši **n-grami** (zaporedja znakov) posameznih jezikov. V tej implementaciji so uporabljeni **n-grami dolžin od 1 do 5 znakov**, saj tak razpon ponuja dobro ravnotežje med občutljivostjo in robustnostjo algoritma.

Generiranje jezikovnih profilov

Za vsak jezik je ustvarjen ločen jezikovni profil:

- Profil se tvori na osnovi velike količine učnega besedila v določenem jeziku.
- Z uporabo knjižnice collections.Counter se preštejejo pojavnosti vseh n-gramov v besedilu.
- Shrani se **400 najpogostejših n-gramov**, ki predstavljajo karakteristični podpis jezika.

Ta postopek omogoča, da se vsak jezik predstavi kot seznam značilnih zaporedij znakov, urejenih po pogostosti.

Predobdelava besedila

Pred generiranjem n-gramov se besedilo normalizira:

- Pretvori se v male črke.
- Odstranijo se vsi znaki, ki niso črke (vključno z ne-latinico, ker je podprta Unicode), številke ali apostrofi.
- Uporabljen je Unicode-zmožen **regularni izraz**: "[^\p{L}\p{N}']", ki s pomočjo knjižnice regex ohranja črke iz različnih pisav (npr. japonščina, cirilica), številke in apostrofe.
- Na začetku in koncu besedila se doda znak _, ki olajša zaznavo začetkov in koncev besed.

Po čiščenju se iz besedila generirajo vsi n-grami z drsečim oknom.

Detekcija jezika

Za novo vhodno besedilo (npr. posamezen stavek) se prav tako izdela njegov **profil** (top 400 n-gramov). Nato se ta profil primerja z vsemi vnaprej naučenimi jeziki.

Uporabljen je **rang-baziran pristop za merjenje razdalje**, kjer se razdalja med dvema profiloma izračuna kot:

- vsota razlik pozicij istih n-gramov v obeh profilih,
- če n-gram iz vhodnega besedila ni prisoten v jezikovnem profilu, se za ta n-gram k vsoti prišteje maksimalna možna vrednost (dolžina profila),
- manjša kot je razdalja, bolj verjetno je, da se besedilo ujema z določenim jezikom.

Prednosti tehnične zasnove

- Podpora za več pisav in jezikov (vključno z latinico, cirilico, japonščino).
- Ločitev **učne faze** in **faze klasifikacije** omogoča ponovno uporabo jezikovnih profilov brez ponovnega treniranja.
- Algoritem je učinkovit in enostaven za razširitev na nove jezike (z dodatnimi učnimi podatki).
- V primeru neznanega jezika bo algoritem še vedno izbral najbližji znani profil po minimalni razdalji.

4. Rezultati testiranja

Testiranje je bilo izvedeno na 25 primerih – za vsak jezik po 5 povedi. Spodaj je pregled rezultatov:

Skupno primerov	Pravilno prepoznanih	Natančnost
25	23	92.0 %

Napačno klasificirani primeri:

- **Primer 8:** *"Kdaj je sestanek?"* (slovenščina) → napačno prepoznan kot hrvaščina
- **Primer 20:** *"Računalnik je treba ponovno zagnati."* (slovenščina) → napačno prepoznan kot hrvaščina

V obeh primerih gre za podobnosti med slovenskim in hrvaškim jezikom, kar je glede na njuno sorodnost pričakovano.

#	Text	German	Japanese	Slovenian	English	Spanish	Croatian	Detected	Correct
1	Hello, how are you today?	35604	40771	35124	29955	34256	35918	english	✓
2	Guten Tag, wie geht es Ihnen?	32559	45008	39760	37594	38923	40374	german	✓
3	Buenos días, ¿cómo está?	34199	39289	33251	33566	29264	34362	spanish	✓
4	Dobro jutro, kako ste?	30070	34442	26575	30399	29674	26724	slovenian	✓
5	What time is the meeting?	31900	38737	33461	26227	34923	33716	english	✓
6	Wann ist das Meeting?	24776	34154	28160	25460	28647	28079	german	✓
7	¿A qué hora es la reunión?	36195	42220	36732	36235	30746	36870	spanish	✓
8	Kdaj je sestanek?	23694	27972	20400	24103	23453	19996	croatian	✗
9	I would like to order food	35392	43117	35529	32710	36812	36509	english	✓
10	Ich möchte etwas zu essen be…	40580	55910	47801	45994	46800	48876	german	✓
11	Me gustaría pedir comida	34681	39776	32324	32819	30953	33049	spanish	✓
12	Rad bi naročil hrano	28398	32693	25804	27583	27276	26031	slovenian	✓
13	It's a beautiful sunny day	36842	43381	38331	35458	36394	38030	english	✓
14	Es ist ein schöner sonniger …	35191	48477	41008	40617	41456	42383	german	✓
15	Es un hermoso día soleado	33916	40845	33467	32887	31984	34364	spanish	✓
16	Lep sončen dan je	22940	27658	21952	24137	23486	22254	slovenian	✓
17	The computer needs to be res…	44916	54391	44866	40341	43232	45902	english	✓
18	Der Computer muss neu gestar…	48172	60347	51164	49456	50414	51944	german	✓
19	La computadora necesita rein…	47699	56290	46281	47547	42802	46103	spanish	✓
20	Računalnik je treba ponovno …	51278	57407	43771	51000	49222	43742	croatian	✗
21	こんにちは、お元気ですか?	24801	21313	24801	24801	24801	24801	japanese	✓
22	会議は何時ですか?	17209	15428	17209	17209	17209	17209	japanese	✓
23	食べ物を注文したいです	21600	18583	21600	21600	21600	21600	japanese	✓
24	今日は晴れの美しい日です	23200	20073	23200	23200	23200	23200	japanese	✓
25	コンピュータを再起動する必要があります	37200	32994	37200	37200	37200	37200	japanese	✓

```
================================================================
                        FINAL SUMMARY
================================================================
Total test cases: 25
Correct detections: 23
Accuracy: 92.0%
================================================================
```

5. Zaključek

Program uspešno dosega visoko natančnost pri detekciji jezikov z uporabo preprostega, a učinkovitega modela n-gram profiliranja. Slabosti se kažejo pri jezikih, ki so si zelo

podobni (slovenščina in hrvaščina), kar bi lahko izboljšali z več podatki ali dodatnimi jezikovnimi značilnostmi (npr. funkcijske besede, morfološka analiza).