POROČILO

Jezikovne Tehnologije – vaja 3

DOLOČITEV JEZIKA BESEDILA Z UPORABO KLASIFIKACIJE TEMELJEČE NA ZNAKOVNIH *N*-GRAMIH

V tem poročilu je predstavljeno implementacija in delovanje programa ter prikaz rezultatov klasifikacije besedila med različnimi jeziki z uporabo znakovnih *n*-gramov.

OPIS NALOGE IN DELOVANJE PROGRAMA

Klasificikacija podpira 5 jezikov, in sicer **angleščino, nemščino, slovenščino, španščino in hrvaščino**.

Program je implementiran v programskem jeziku Python in nekaj osnovnimi knjižnicami (typing, enum, re, sys).

Na spletu sem najprej poiskal korpuse primernih velikosti za vsakega od 5 jezikov in jih shranil v tekstovne (.txt) datoteke na svoji napravi.
Korpusi, ki sem jih uporabil, so dostopni na tej povezavi.
Sam se uporabil najnovejše 30,000 vrstične korpuse iz sekcije *Wikipedia*.

Pred začetkom učenja modela, sem moral učne korpuse najprej predobdelati, pri čemer je bilo potrebno odstraniti posebne znake, števila in ločila (pri tem ohranimo jezikovno specifične posebne znake kot so *ä*, *ö* in *ü* za nemščino). Nato je bilo korpuse potrebno razdeliti v polja, kjer je vsak znak predstavljal svoj element – dobimo polje leksikalnih simbolov ali leksemov. Vsakemu leksemu dodamo še presledek na začetku in na koncu.

Iz dobljenih polj pravilno predobdelanih leksemov nato zgradimo *n*-grame, kjer je *n* celo število od 1 do 5 – zgradimo torej vse možne unigrame, bigrame, trigrame, kvadrigrame (tetragrame) in pentagrame, katere zatem združimo, uredimo padajoče po njihovih pojavnostih, ter ohranimo le 300 najbolj frekvenčnih *n*-gramov, saj le-ti navadno dovolj dobro opisujejo posamezen jezik.

Vse zgrajene *n*-grame (zbirki le-teh bomo v nadaljnje rekli jezikovnih profili) za prihranitev časa shranimo v tekstovne datoteke (vsak *n*-gram s svojo pojavnostjo v svojo vrstico datoteke).

Ob nadaljnjih uporabah že obstoječih jezikovnih profilov, lahko le-te preprosto prebremo iz datoteke, kar nam prihrani izjemno veliko časa napram ponovni izgradnji.

Zgoraj opisano je predstavljajo postopek učenja modela, v nadaljevanju pa naučeni model testiramo z uporabo testne množice oz. testnega dokumenta.

Testni dokument preberemo iz tekstovne datoteke, ter ga s pomočjo naučenih jezikovnih profilov klasificiramo, v kateri jezik spada. Klasifikacijo izvedemo tako, da najprej zgradimo znakovne *n*-grame za testni dokument, nato pa z izračunom razdalje urejanja med znakovnimi *n*-grami naučenih jezikovnih profilov in znakovnih *n*-gramov testnega dokumenta dobimo skupno razdaljo

za vsak jezikovni model. Jezikovni model, ki ima najmanjšo razdaljo do testnega dokumenta, je jezik testnega dokumenta.

Razdalja urejanja je v tem primeru implementirana na način, da se izračunavajo in seštevajo razdalje med ujemajočimi znakovnimi *n*-grami tako, da se gleda razlika v frekvenčnosti oz. pojavitvi (torej računamo razliko med indeksom *n*-grama v jezikovnem profilu in indeksom *n*-grama testnega dokumenta), v primeru neujemanja, se ta razdalja nastavi na vrednost 300 (kar je seveda maksimalna možna razdalja).

REZULTATI

Kot že rečeno, sem program učil s 5 korpusi (vsak za enega od petih jezikov), pri čemer je vsak korpus bil vsebinsko izjemno izčrpen in dolg 30,000 vrstic.

Model sem nato testiral nad 20 tekstovnimi datotekami, pri čemer je vsak imel drugačno vsebino pripadajočega jezika in bil dolg 400 vrstic. Prepričal sem se, da se testni podatki razlikujejo od učnih.

Klasifikacija je bila 100% uspešna pri vseh 5 jezikih, torej je modelu uspelo uspešno klasificirati vseh 20 testnih dokumentov.

Spodaj je prikazana tabela vrednosti razdalje urejanja za prvih 5 testnih dokumentov, za voljo časa in prostora.

Tabela 1: klasifikacija angleških testnih dokumentov

datoteka\profil	ENGLISH	GERMAN	SLOVENE	SPANISH	CROATIAN
0.txt	15859	49781	56519	49509	56652
1.txt	14326	50332	56231	48570	55908
2.txt	9893	49110	56542	47661	56474
3.txt	11789	46686	55664	47193	55339
4.txt	11848	48229	56693	47738	56507

Tabela 2: klasifikacija nemških testnih dokumentov

datoteka\profil	ENGLISH	GERMAN	SLOVENE	SPANISH	CROATIAN
0.txt	50196	17335	59534	56690	59340
1.txt	50273	16271	60125	58671	60424
2.txt	50015	13619	60291	57919	60045
3.txt	48049	10433	59107	55775	58646
4.txt	49270	14241	59829	56862	59361

Tabela 3: klasifikacija slovenskih testnih dokumentov

datoteka\profil	ENGLISH	GERMAN	SLOVENE	SPANISH	CROATIAN
0.txt	52857	56582	11070	49745	24565
1.txt	53927	58216	11136	51028	24618
2.txt	54416	57198	10075	50627	24135
3.txt	53045	56448	10123	50482	23220
4.txt	54397	57345	10895	51131	24766

Tabela 4: klasifikacija španskih testnih dokumentov

datoteka\profil	ENGLISH	GERMAN	SLOVENE	SPANISH	CROATIAN
0.txt	48534	58865	55262	13207	55902
1.txt	47536	57564	53389	10414	55149
2.txt	47000	57483	53231	8005	53773
3.txt	47848	58327	54415	12114	55027
4.txt	49822	59788	55375	15799	55947

Tabela 5: klasifikacija hrvaških testnih dokumentov

datoteka\profil	ENGLISH	GERMAN	SLOVENE	SPANISH	CROATIAN
0.txt	54159	58362	25192	51595	12115
1.txt	54842	59768	24534	52013	9399
2.txt	55852	59837	24756	53533	10609
3.txt	55467	58670	25665	52436	14573
4.txt	54624	58836	26959	52438	13733

Iz zgornjih 5 tabel je hitro jasno, da je klasifikacija bila pravilna zaradi znatno nižjega števila razdalje urejanja, le-ta stolpec je označen z zeleno barvo.

Poleg tega pa zgornji podatki dajejo tudi zanimiv vpogled v podobnost (angl. similarity) med različnimi jeziki. Stolpce jezikov, ki so najbolj podobni ciljnemu jeziku, sem označil z oranžno barvo. Jezike, ki so ciljnemu jeziku najbolj tuji, pa sem označil z rdečo barvo.

Kot pričakovano, sta si slovenščina in hrvaščina najbolj podobna med vsemi 5 jeziki, najbolj tuja pa je obema nemščina.

Angleščini kljub skupni germanski jezikovni skupini, nemščina ni najbolj podobna, temveč jo španščina prekaša za zelo majhno vrednost.

Nemščini je seveda najbolj podobna angleščina, saj je to edini drugi (vsaj delno) germanski jezik. Najbolj tuj jezik nemščini pa je ravno slovenščina, čeprav le za malo v primerjavi s hrvaščino.

Španščini je najbolj podobna angleščina in najbolj tuja nemščina.

Izkaže se tudi, da je nemščina najbolj tuj jezik vsem ostalim, saj so vrednosti njenega najbolj podobnega jezika (t.j. angleščina) najvišje med vsemi jeziki.

Jakob Oprešnik