



Určovanie Slovných Druhov v Slovenčine

DALIBOR MÉSZÁROS
VEDÚCI: MÁRIUS ŠAJGALÍK

Rekapitulácia

- ▶ Určovanie Slovných Druhov
 - ▶ Part-of-Speech (POS) Tagging
- ▶ Podmienené Náhodné Polia
 - ▶ Conditional Random Fields (CRF)
- ▶ Črty/Predlohy Črt
 - ▶ Features/Features Template

Rozšířené črty

- ▶ Predpona a prípona slova pevnej dĺžky
- ▶ Slová prekonvertované na malé písmená
- ▶ Binárna informácia, či slovo obsahuje veľké písmená
- ▶ Binárna informácia, či slovo obsahuje čísla, alebo interpunkčné znamienka
- ▶ Dĺžka slova
- ▶ Pozícia slova vo vete
- ▶ Vektorová reprezentácia slov

Predlohy črt

tag1_nrm_case_spcl_len_pos_fix1234

O	1685	0	0	0	0	1	4	4	1	16	168	1685	5	85	685	1685
R	si	0	0	0	0	0	2	5	s	si	si	si	i	si	si	si
S	Turci	1	0	1	0	0	5	6	T	Tu	Tur	Turc	i	ci	rci	urci
V	podrobili	0	0	0	0	0	9	7	p	po	pod	podr	i	li	ili	bili
T	aj	0	0	0	0	0	2	8	a	aj	aj	aj	j	aj	aj	aj
A	Nové	1	0	1	0	0	4	9	N	No	Nov	Nové	é	vé	ové	Nové
S	Zámky	1	0	1	0	0	5	10	Z	Zá	Zám	Zámk	y	ky	mky	ámky
Z	.	0	0	0	1	0	1	11	.	.	.	.	.	.	.	.

tag1_low_case_fix123_vec5

O	1685	0	0	0	1	16	168	5	85	685	2880	3075	2868	3034	2852
R	si	0	0	0	s	si	si	i	si	si	1111303	841276	750664	25483	805898
S	turci	1	0	1	t	tu	tur	i	ci	rci	539382	345400	523945	174675	155165
V	podrobili	0	0	0	p	po	pod	i	li	ili	994531	994527	994574	994591	994529
T	aj	0	0	0	a	aj	aj	j	aj	aj	613672	748803	1164532	923567	852789
A	nové	1	0	1	n	no	nov	é	vé	ové	359868	303887	221070	223986	520610
S	zámky	1	0	1	z	zá	zám	y	ky	mky	593801	180179	1247497	271204	519872
Z	.	0	0	0	.	.	.	.	.	.	-1	-1	-1	-1	-1

Výsledky

Trénovacia množina: 515 624 slov

Validačná množina: 141 939 slov

Neznámich: 20 791 slov

Veľkosť písmen	Informácie o veľkosti písmen	Obsahuje číslo alebo interp. zn.	Dĺžka slova	Pozícia vo vete	Dĺžka predpony a prípony	Presnosť	Presnosť pre neznáme slová
Lowercase	Áno	Nie	Nie	Nie	3	95,60%	84,34%
Lowercase	Áno	Nie	Nie	Nie	4	94,97%	81,04%
Lowercase	Áno	Nie	Nie	Nie	1,2,3	96,04%	86,44%
Lowercase	Áno	Nie	Nie	Nie	1,2,3,4	96,41%	87,16%
Lowercase	Áno	Áno	Áno	Áno	1,2,3	96,25%	86,54%
Lowercase	Áno	Áno	Áno	Áno	1,2,3,4	96,63%	87,34%
Nezmenená	Nie	Nie	Nie	Nie	3	94,94%	81,53%
Nezmenená	Nie	Nie	Nie	Nie	4	94,31%	78,34%
Nezmenená	Nie	Nie	Nie	Nie	1,2,3	95,84%	85,51%
Nezmenená	Nie	Nie	Nie	Nie	1,2,3,4	96,27%	86,49%
Nezmenená	Áno	Áno	Áno	Áno	1,2,3	96,11%	85,95%

Veľkosť písmen	Informácie o veľkosti písmen	Obsahuje číslo alebo interp. zn.	Dĺžka slova	Pozícia vo vete	Dĺžka predpony a prípony	Počet vektorov	Presnosť	Presnosť pre neznáme slová
Lowercase	Áno	Áno	Áno	Áno	1,2,3,4	10	96,36%	86,24%
Lowercase	Áno	Áno	Áno	Áno	1,2,3,4	20	96,68%	86,70%