Ekstrakcija likova iz kratkih priča

Gorana Levačić
Tomislav Horina

O problemu

- Bavili smo se ekstrakcijom likova iz kratkih priča
- Priče smo skinuli s Project Gutenberg
- Označavali smo ih ručno
- Imali smo oznake “O” i “C”
- Promatrali smo dali utječe na model veličina slova te postojanje interpunkcijskih znakova

Korišteni modeli

- Preoblikovali smo postojeći Stanford Ner da radi na našim oznakama
- Od modeli smo koristili
  - Skrivene Markovljeve modele(eng. Hidden Markov Model, HMM)
  - Uvjetna nasumična polja(eng. Conditional Random Fields, CRF)

- Marica konfuzije

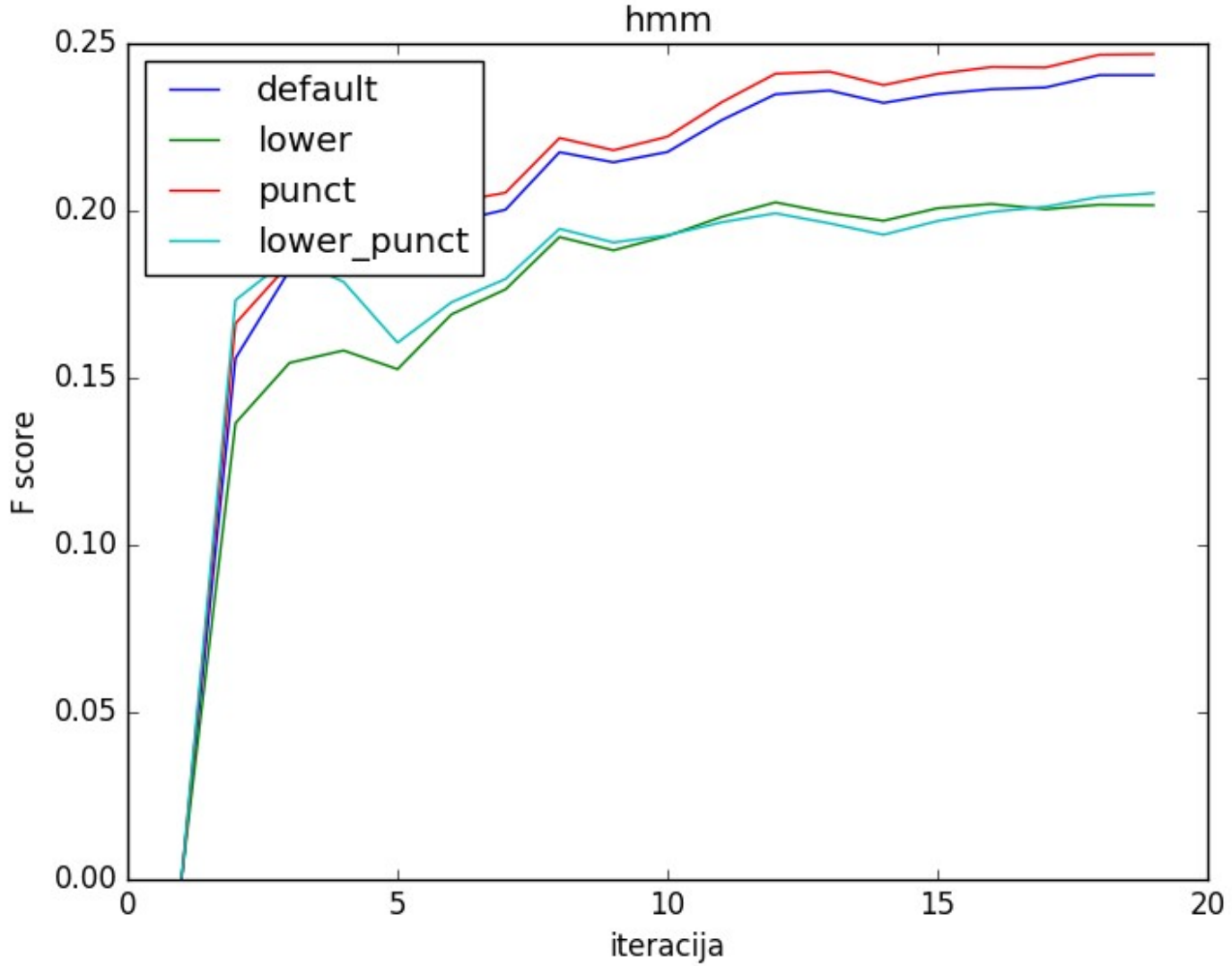
$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

  - TP – označio s “C” , trebao s “C”
  - FN – označio s “O”, trebao s “C”
  - FP – označio s “C”, trebao s “O”
  - TN – označio s “O”, trebao s “C”

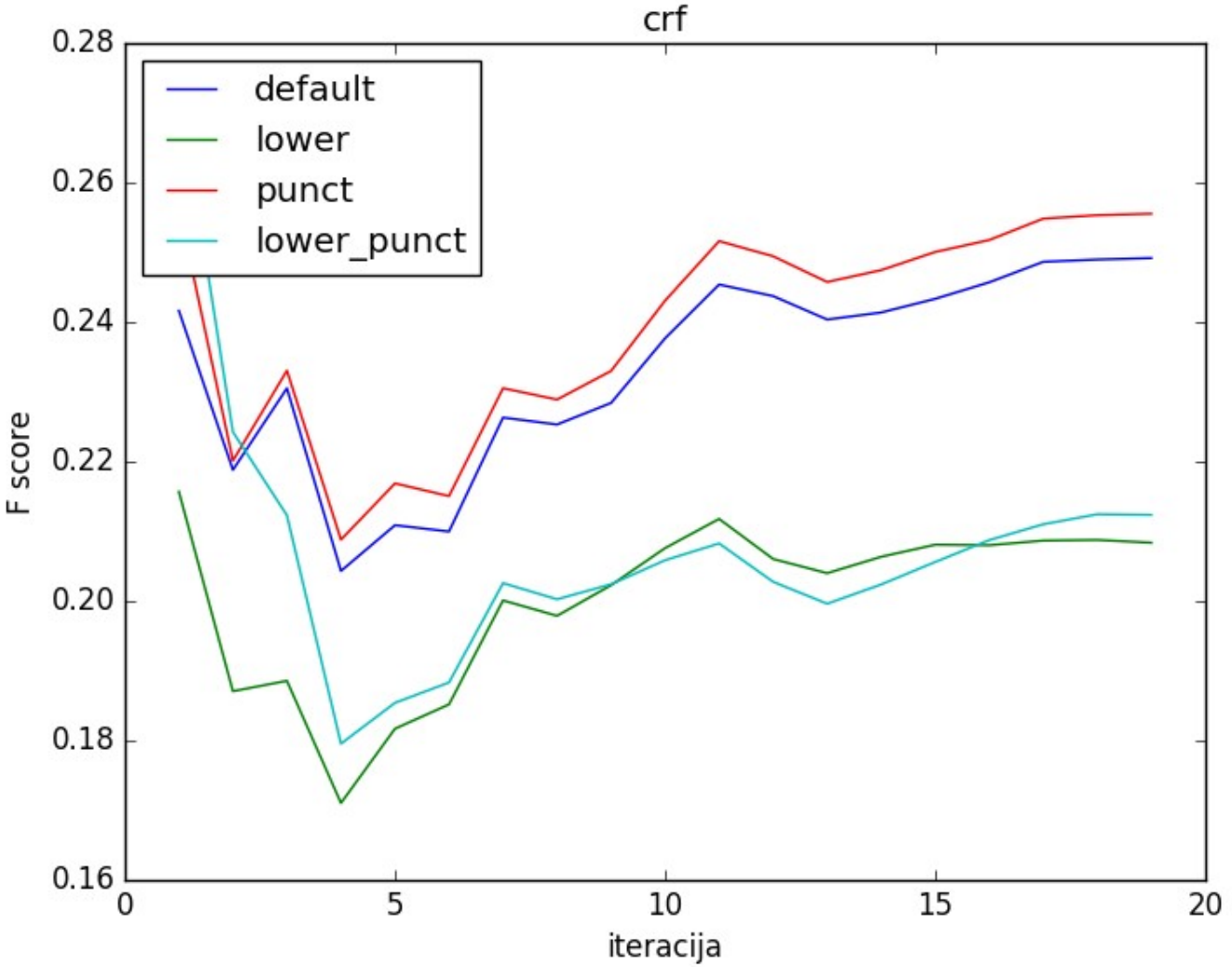
- Mjera koju smo koristi je F_2

$$F_2 = \frac{5 \times preciznost \times osjetljivost}{4 \times precizniost + osjetljivost}$$

Unakrsna validacija - HMM



Unakrsna validacija - CRF



Test - HMM

HMM svi modeli		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	0	86
	Nisu likovi	0	5366

Test – CRF

CRF default		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	32	54
	Nisu likovi	52	5314

$F_2 = 0.3738$

CRF punct		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	33	53
	Nisu likovi	54	5312

$F_2 = 0.3828$

CRF lower		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	25	61
	Nisu likovi	47	5319

$F_2 = 0.3005$

CRF lower punct		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	27	59
	Nisu likovi	36	5330

$F_2 = 0.3317$

Test – Stanford NER

Stanford NER default		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	55	55
	Nisu likovi	31	5311

$F_2 = 0.6057$

NER punct		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	48	38
	Nisu likovi	30	5326

$F_2 = 0.5556$

NER lower		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	41	45
	Nisu likovi	66	5300

$F_2 = 0.4545$

NER lower punct		Oznake modela	
		Likovi	**Nisu likovi**
Točne oznake	**Likovi**	30	56
	Nisu likovi	36	5300

$F_2 = 0.3659$

Zaključak

- Dobili smo relativno dobre rezultate za Stanford Ner
- Očekujemo da bismo dobili bolje rezultate uz veći skup podataka za učenje