

Slovenski NLTK označevalnik

Niko Colnerič Nejc Banič

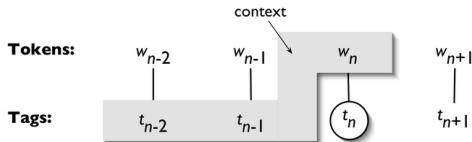
Fakulteta za Računalništvo in Informatiko
Univerza v Ljubljani

January 16, 2012

- NLTK - Natural Language Toolkit (Python 2.7)
- Označevalnik (tagger) besedam pripiše oblikoslovne oznake
- Težave z dvoumnostjo
- Uporabnost za nadaljno procesiranje teksta
- Tega so nas učili v šoli 😊

- **Trigram označevalnik**

a priori najbolj verjetno oznako v kontekstu dolžine 3



- **Brillov označevalnik**

ugane oznako vsako besede, nato popravi svoje napake tako, da uporabi seznam transformacijskih pravil

- označevalnik na podlagi **naivnega Bayes-ovega klasifikatorja**

predpostavi pogojno neodvisnost atributov pri danem razredu

- korpus = velika in strukturirana zbirka besedil
- Jezikoslovno Označevanje Slovenskega jezika (IJS)
- Zbirko besedil s ročno preverjenimi jezikoslovnimi oznakami
- *.XML* datoteka
- Uporabila sva *josiM* - 1 milijon označenih besed
- MULTEXT-East V4 specifikacija
- MSD (morphosyntactic descriptions) - definirajo kategorije oz. besedne vrste
- Namenjena spodbujanju razvoja jezikovnih tehnologij za slovenski jezik

- Uporabljen za treniranje NLTK objektov - označevalnikov
- *Train NLTK objects with zero code*
- Dobro dokumentiran
- Prosto dostopen na Github-u

Postopek treniranja označevalnikov

- 1 Transformacija *.XML* korpusa v *.pos*
- 2 Treniranje z skripto iz NLTK-trainer
- 3 Ocenjevanje točnosti
- 4 Ocenjevanje hitrosti

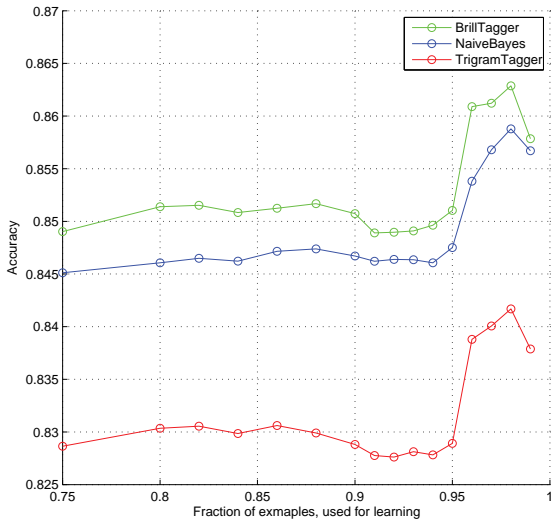
Lep je dan, vse diši že po pomladi!

- (*Lep* — *PPNMEIN*) - pridevnik
- (*je* — *GP-STE-N*) - glagol
- (*dan* — *SOMEI*) - samostalnik
- (*,* — *,*) - ni razlage
- (*vse* — *ZC-SEI*) - zaimek
- (*diši* — *GGNSTE*) - glagol
- (*že* — *L*) - členek
- (*po* — *DM*) - predlog
- (*pomladi* — *SOZEM*) - samostalnik
- (*!* — *!*) - ni razlage

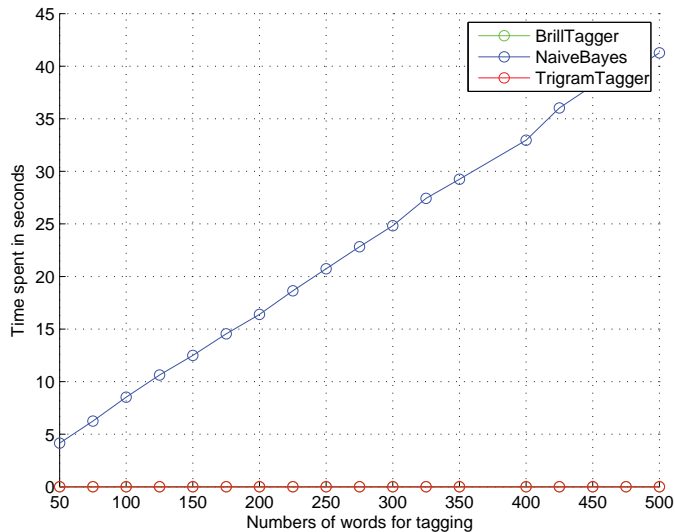
pridevnik

vrsta	splošni
spol	moški
število	ednina
sklon	imenovalnik
živost	0
vid	0
oblika	0
oseba	0
nikalnost	0
stopnja	nedoločeno
določnost	ne
število_svojine	0
spol_svojine	0
naslonskost	0
zapis	0

Rezultati natančnost



Rezultati hitrost



- Brill in Trigram sta hitra, NaiveBayes veliko počasnejši
- Najbolj natančen Brill, najmanj Trigram
- \Rightarrow Priporočama uporabo Brill označevalnika
- Vključitev v NLTK
- Dostopno na:
`https://github.com/nikicc/slovene-nltk-tagger`
- Z lahkoto bi zgradili tudi druge označevalnike, ki bi lahko bili boljši ali hitrejši
- Vključitev drugih korpusov, bi bila koristna