

Součástí úkolu jsou Python3 skripty `assignment3_1.py` a `assignment3_2.py`.

Task 1

Použili sme Brillův tagger z knihovny NLTK. Ako počiatkový tagger sme použili jednoduchý unigramový tagger, ktorý si zapamätá najčastejšie značku pre každé slovo tréningových dát a pre neznáme slová vráti špeciálnu značku “None”. Použili sme počiatkovú množinu 24 vzorov¹ (template), ktoré pôvodne použil Brill. Pre angličtinu sme počet pravidiel obmedzili² na 247, pre češtinu na 500. Tieto konkrétne hodnoty sme dostali skúšaním niekoľkých hodnôt. Všeobecne platí, že väčšie množstvo pravidiel môže mierne zlepšiť presnosť, ale výrazne spomaľuje.

Výsledky

Angličtina	Přesnost	Čeština	Přesnost
1. stupeň	0.8909342163585638	1. stupeň	0.7738987193793506
2. stupeň	0.8890974014126724	2. stupeň	0.7888304373239811
3. stupeň	0.8747591522157996	3. stupeň	0.7535799207397622
4. stupeň	0.8935028395769291	4. stupeň	0.7922554742597533
5. stupeň	0.8925100950892275	5. stupeň	0.7880596223212392
Průměr	0.888160740931	Průměr	0.779324834805
σ	0.00686479392893	σ	0.0124922171647
Průměr triv.	0.865374314735	Průměr triv.	0.738888267001
σ triv.	0.0143254860843	σ triv.	0.0134321134672

Tabulka 1: Přesnost (accuracy) pro Brillův tagger pro oba jazyky.

Task 2

Supervizovaný HMM

Angličtina	Přesnost
1. stupeň	0.9280369201887727
2. stupeň	
3. stupeň	
4. stupeň	
5. stupeň	
Průměr	
σ	

¹NLTK má tieto vzory ako súčasť knihovny

²Limit sa pri týchto počtoch vždy naplní, čiže limit je rovnaký ako počet pravidiel.

Nesupervizovaný HMM — Baum-Welch

Pre túto časť sme sa rozhodli použiť ako stavy iba bigramy, ktoré sme videli v trénovacích dátach, pretože inak by bol počet stavov bol veľký.

Navyše sme sa rozhodli použiť pruningovú metódu, aby sme tréning urýchlili.