

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. Použili sme počiatkovú množinu 24 vzorov¹ (template), ktoré pôvodne použil Brill. Pre angličtinu sme početpravidiel 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	0.8910645842567726	1. stupeň	0.7738987193793506
2. stupeň	0.88881069669247	2. stupeň	0.7888304373239811
3. stupeň	0.8602301723689006	3. stupeň	0.7535799207397622
4. stupeň	0.8931381232741104	4. stupeň	0.7922554742597533
5. stupeň	0.8927185098345708	5. stupeň	0.7880596223212392
Průměr	0.885192417285	Průměr	0.779324834805
σ	0.0125734981194	σ	0.0143254860843

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

Task 2

Nesupervizovaný HMM — Baum-Welch

Pre túto časť sme sa rozhodli použiť iba bigramové a unigramové štatistiky, pretože ak by sme nechceli použiť nuly “natvrdo” zabralo by to prílišne veľa času pre angličtinu, pre češtinu by to bolo jednoducho nerealizovateľné, časovo ani pamäťovo.

¹NLTK má tieto vzory ako súčasť knihovny

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