Jana Novotná & Vladan Glončák
Statistical NLP II
Assignemnt 3

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

Task 1

Použili sme Brillov tagger z knižnice NLTK. Ako počiatočný tagger sme použili jednoduchý unigramový tagger, ktorý si zapamätá najčastejšie značku pre každé slovo trénovacích dat. Použili sme počiatočnú množinu 18 pravidel, rovnaké ako v ukázke v knihe[1]. Počet pravidiel je maximálne 200 (prednastavená hodnota NLTK).

Výsledky

Angličtina	**Přesnost**
1. stupeň	0.8878314603811958
2. stupeň	0.8877681340735528
3. stupeň	0.8591886684372233
4. stupeň	0.8916271557338613
5. stupeň	0.8927185098345708
Průměr	0.88382678569208084
σ	0.01247811683956135

Čeština	**Přesnost**
1. stupeň	0.7611722195653329
2. stupeň	0.7784951378925554
3. stupeň	0.7446763540290621
4. stupeň	0.7794976095512296
5. stupeň	0.7772650116698494
Průměr	0.76822126654160594
σ	0.01355396354317235

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 použiť nuly "natvrdo" zabralo by to príšerne veľa času pre angličtinu, pre češtinu by to bolo jednoducho nerealizovateľné, časovo ani pamäťovo.

[1] Volne dostupné na `http://www.nltk.org/book/`