**Nyelvtechnológia alapjai**

**2. házi feladat**

**1. Feladat:**





Új szabályok:

 VP → AP NPb

S → VP NP VP

NP → NP AP

VP → VP NP

S → NP VP NP

A szabályok alkalmazásának jó sorrendje az, ha minél beszédesebb logikai egységeket igyekszünk képezni a mondat részeiből.

**2. Feladat:**

****

**3. Feladat:**

hypernym:

* (vizsla, puli) → (kutya, eb) → emlős
* fehér → világos

entailment:

* repül → mozog
* (mozog, lélegzik) → él

part meronym:

* (autó, motor, busz) → kerék

antonym:

* fekete ↔ fehér

has instance:

* (kutya, eb) → Frakk

**TnT tagger:**

A TnT a Trigrams’n’Tags rövidítése. Ez egy általános címkéző algoritmus ami a másodrendű Markov modellen alapszik. Sok programozási nyelv hoz rá open-source implementációt, amik egytől egyig hatékonyan taníthatók, és sokkal effektívebbek, mint a többi, más algoritmust használó címkézők.

A célja a minél hatékonyabb szövegelemzés. Különböző források szerint sokkal hatékonyabb is tud lenni nagy tanító korpuszok esetén, mint például a Maximum Entrópia keretrendszer. Ezt nagyrészt a kifinomult módszerének köszönheti amivel az ismeretlen szavakat kezeli.

Mint már említettem, ez a módszer a másodrendű Markov modellt használja fel a működéséhez. A modell állapotai reprezentálják a címkéket, míg a kimenete a szavakat. Mivel az átmeneti valószínűségek az állapotoktól vagyis lényegében a címkéktől függenek. A kimenet valószínűsége viszont csakis a legutóbbi kategóriától.

Az ismeretlen szavakra alkalmazott módszere pedig, hogy ezen szavak képzőit tekinti meghatározónak. Kimutatások szerint a Wall Street Journal-ban megjelent szavak 98%-áról helyesen megállapítható volt a szófajuk úgy, hogy csak a képzőiket vizsgálták.

Még használ egy úgynevezett simító algoritmust is ami segít kiszűrni a hibás kiugrásokat, kevés példány esetén.