Elemzőláncok magyar nyelvre

Nyelv és informatika – Pécs, 2022/23 tavasz
8. óra

Simon Eszter – Vadász Noémi

2023. április 1.

Tartalom

emtsv

e-magyar/emtsv i

- moduláris, bármelyik lépésnél ki-be lehet szállni, új modulokat könnyen lehet integrálni
- lassabb: dependenciáig REST API-val: 310 token/s
- python, docker
- modulok:
  - tokenizálás, mondatrabontás
  - morfológiai elemzés és egyértelműsítés
  - szintaktikai elemzés: függőségi és összetevős
  - tulajdonnév-felismerés
  - főnévi frázisok felismerése
  - szótáralapú kifejezésfelismerő
  - zérónévmás-beszúró
  - konverterek
  - kiértékelő

e-magyar/emtsv ii

- nyílt forráskódú, szabadon felhasználható
- SOTA teljesítmény

https://github.com/dlt-rilmta/emtsv

https://e-magyar.hu/hu/

Stanza

Stanza

- a Stanford NLP Group Python-alapú szövegfeldolgozó eszköze
- a teljes pipeline neurális eszközökön alapul
- elemzési szintek:
  - tokenizálás, mondatrabontás
  - morfológiai elemzés és egyértelműsítés: csak UD POS tageket és morfológiai feature-öket ad ki, de nem lemmatizál
  - szintaktikai elemzés: függőségi
  - tulajdonnév-felismerés
- nyílt forráskódú, szabadon használható
- kevésbé jó teljesítmény

https://stanfordnlp.github.io/stanza/

UDPipe

UDPipe

- elemzési szintek:
  - tokenizálás, mondatrabontás
  - morfológiai elemzés és egyértelműsítés
  - szintaktikai elemzés: függőségi elemzés
- kevésbé jó teljesítmény
- gyors: dependenciáig: 3.300 token/s
- az egyes lépéseknél ki-be lehet szállni az egységes formátumnak köszönhetően, de új modulokat nem lehet integrálni
- nyílt forráskódú, szabadon használható
- neurális módszereken alapul minden modulja

http://ufal.mff.cuni.cz/udpipe

HuSpacy

HuSpacy

- elemzési szintek:
  - tokenizálás, mondatrabontás
  - morfológiai elemzés és egyértelműsítés
  - szintaktikai elemzés: függőségi
  - tulajdonnév-felismerés
  - szóvektorok
- a leggyorsabb: dependenciáig: 15.000 token/s
- nyílt forráskódú, szabadon használható
- python
- elég jó teljesítmény
- bármelyik lépésnél ki-be lehet szállni, új modulokat könnyen lehet integrálni

https://github.com/huspacy/huspacy

https://huggingface.co/spaces/huspacy/demo

magyarlánc

Magyarlánc

- a Szegedi Tudományegyetemen fejlesztett eszközlánc
- nem moduláris, csak egyben lehet futtatni az elejétől a végéig, új modulokat nem lehet integrálni
- gyorsabb: dependenciáig: 450
- java
- elemzési szintek:
  - tokenizálás, mondatrabontás
  - morfológiai elemzés és egyértelműsítés
  - szintaktikai elemzés: függőségi és összetevős
- letölthető, szabadon felhasználható
- SOTA teljesítmény

https://rgai.inf.u-szeged.hu/magyarlanc


"That's all Folks!"

Köszönjük a figyelmet!

simon.eszterke@gmail.com
vadasz.noemi@nytud.hu