Miért van szükség a nyelvtechnológiára?

Ács Judit

BME AUT

KSZK40, 2016. szeptember 23.

természetes nyelvfeldolgozás

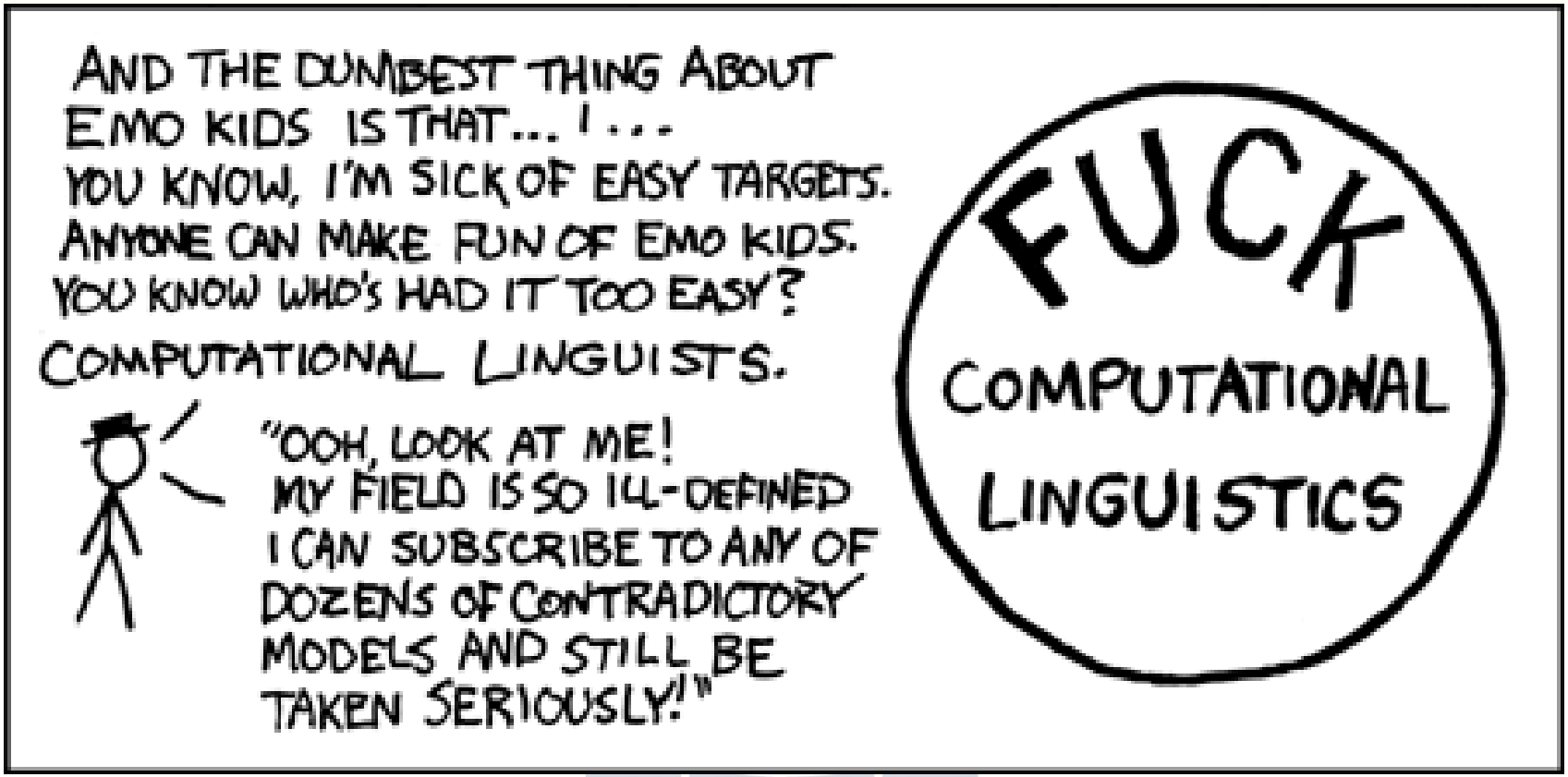
natural language processing

számítógépes nyelvészet

NLP

nyelvtechnológia

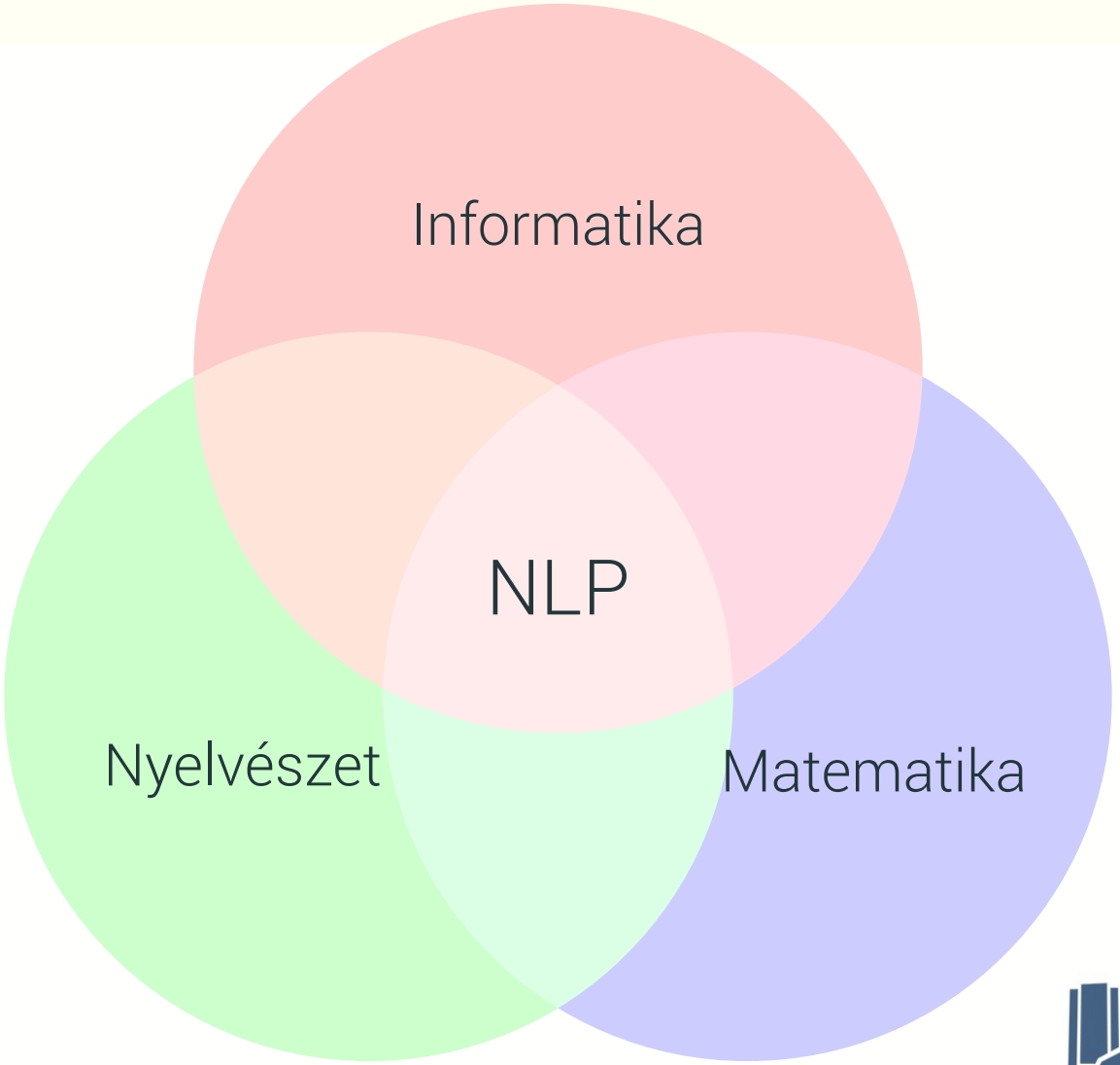
Schönherz



https://xkcd.com/114/

Schönherz

Interdiszciplináris



Schönherz

Definíció?

Schönherz

Definíció?

- az input és/vagy az output természetes nyelv

Schönherz

Definíció?

- az input és/vagy az output természetes nyelv
- formája sokféle lehet: írott, beszélt, jelnyelv, Braille stb.

Schönherz

Fordítás

Schönherz

Fordítás

- Fordítsuk le az alábbi mondatokat magyarra!

He went back to New York and left Mary behind.
She was devastated.

Schönherz

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Schönherz

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

Schönherz

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

- a ***back*** szó kétértelmű (vissza, hát),

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

- a ***back*** szó kétértelmű (vissza, hát),
  - *left, devastated* is többértelműek

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

- a ***back*** szó kétértelmű (vissza, hát),
  - ***left, devastated*** is többértelműek
- ***New York*** egy tulajdonnév,

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

- a ***back*** szó kétértelmű (vissza, hát),
  - ***left, devastated*** is többértelműek
- ***New York*** egy tulajdonnév,
- ***She*** kire vonatkozik? Angolul egyértelmű, magyarul nem,

Schönherz

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

- a ***back*** szó kétértelmű (vissza, hát),
    - ***left, devastated*** is többértelműek
- ***New York*** egy tulajdonnév,
- ***She*** kire vonatkozik? Angolul egyértelmű, magyarul nem,
- ***went back*** együtt fordítandó
- szórend, toldalékok

Schönherz

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

- a ***back*** szó kétértelmű (vissza, hát), word sense disambiguation
  - ***left, devastated*** is többértelműek
- ***New York*** egy tulajdonnév,
- ***She*** kire vonatkozik? Angolul egyértelmű, magyarul nem,
- ***went back*** együtt fordítandó
- szórend, toldalékok

Schönherz

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

- a ***back*** szó kétértelmű (vissza, hát), word sense disambiguation
  - ***left, devastated*** is többértelműek
- ***New York*** egy tulajdonnév, named entity recognition
- ***She*** kire vonatkozik? Angolul egyértelmű, magyarul nem,
- ***went back*** együtt fordítandó
- szórend, toldalékok

Schönherz

Fordítás szavanként

He went back to New York and left Mary behind. She was devastated.

Ő ment hát nak/nek Új York és van hátra Mary mögött. Ő volt elpusztított.

- a ***back*** szó kétértelmű (vissza, hát), word sense disambiguation
  - ***left, devastated*** is többértelműek
- ***New York*** egy tulajdonnév, named entity recognition
- ***She*** kire vonatkozik? Angolul egyértelmű, magyarul nem, anaphora resolution
- ***went back*** együtt fordítandó
- szórend, toldalékok

Schönherz

Google Translate

He went back to New York and left Mary behind. She was devastated.

Google Translate

He went back to New York and left Mary behind. She was devastated.

Visszament New York-i és a bal Mary mögött. Ő volt elpusztított.

Google Translate

He went back to New York and left Mary behind. She was devastated.

Visszament New York-i és a bal Mary mögött. Ő volt elpusztított.

- *New York-i* lett a *to New York*ból,

Schönherz

Google Translate

He went back to New York and left Mary behind. She was devastated.

Visszament New York-i és a bal Mary mögött. Ő volt elpusztított.

- *New York-i* lett a *to New York*ból,
- *left* rossz jelentését fordítja,

Schönherz

Google Translate

He went back to New York and left Mary behind. She was devastated.

Visszament New York-i és a bal Mary mögött. Ő volt elpusztított.

- ***New York-i*** lett a ***to New York***ból,
- ***left*** rossz jelentését fordítja,
- ***left behind***-ot szó szerint értelmezi?

Schönherz

Google Translate

He went back to New York and left Mary behind. She was devastated.

Visszament New York-i és a bal Mary mögött. Ő volt elpusztított.

- ***New York-i*** lett a ***to New York***ból,
- ***left*** rossz jelentését fordítja,
- ***left behind***-ot szó szerint értelmezi?
- a 2. mondatról inkább ne beszéljünk.

Schönherz

NLP alapfeladatok

Schönherz

NLP alapfeladatok

tokenizálás szöveg kisebb egységekre bontása (mondat, szó).

Schönherz

NLP alapfeladatok

tokenizálás szöveg kisebb egységekre bontása (mondat, szó).

- *New York-i* hány token?
- elválasztás
- dátumok kezelése stb.

Schönherz

NLP alapfeladatok

tokenizálás szöveg kisebb egységekre bontása (mondat, szó).

- *New York-i* hány token?
- elválasztás
- dátumok kezelése stb.

szótövezés toldalékok (képző, rag) eltávolítása. ***házaink - ház***

NLP alapfeladatok

tokenizálás szöveg kisebb egységekre bontása (mondat, szó).

- *New York-i* hány token?
- elválasztás
- dátumok kezelése stb.

szótövezés toldalékok (képző, rag) eltávolítása. ***házaink - ház***

- ***kályhák***, ***ettem***

Schönherz

NLP alapfeladatok

tokenizálás szöveg kisebb egységekre bontása (mondat, szó).

- *New York-i* hány token?
- elválasztás
- dátumok kezelése stb.

szótövezés toldalékok (képző, rag) eltávolítása. *házaink - ház*

- *kályhák*, *ettem*

morfológiai elemzés

- *zúzalékával*

Schönherz

NLP alapfeladatok

tokenizálás szöveg kisebb egységekre bontása (mondat, szó).

- *New York-i* hány token?
- elválasztás
- dátumok kezelése stb.

szótövezés toldalékok (képző, rag) eltávolítása. *házaink - ház*

- *kályhák*, *ettem*

morfológiai elemzés

- *zúzalékával*

névelemazonosítás named entity recognition

Schönherz

NLP alapfeladatok

tokenizálás szöveg kisebb egységekre bontása (mondat, szó).

- *New York-i* hány token?
- elválasztás
- dátumok kezelése stb.

szótövezés toldalékok (képző, rag) eltávolítása. ***házaink - ház***

- ***kályhák***, ***ettem***

morfológiai elemzés

- ***zúzalékával***

névelemazonosítás named entity recognition

- *Catholic* angolul, ***katolikus*** magyarul

Schönherz

NLP alapfeladatok II.

Schönherz

NLP alapfeladatok II.

szófaji elemzés part-of-speech tagging

Schönherz

NLP alapfeladatok II.

szófaji elemzés part-of-speech tagging

„nyelvtani elemzés" nyelvtani szerkezet feltárása. Különböző nyelvészeti elméletek különböző elemzéseket adnak.

Schönherz

NLP alapfeladatok II.

szófaji elemzés part-of-speech tagging

„nyelvtani elemzés" nyelvtani szerkezet feltárása. Különböző nyelvészeti elméletek különböző elemzéseket adnak.

egyértelműsítés word sense disambiguation

Schönherz

Magasabb szintű NLP feladatok

Schönherz

Magasabb szintű NLP feladatok

gépi fordítás rule-based MT, statistical MT, neural MT

Schönherz

Magasabb szintű NLP feladatok

gépi fordítás rule-based MT, statistical MT, neural MT

koreferencia feloldás mik utalnak ugyanarra az entitásra? *New York – Big Apple*

Schönherz

Magasabb szintű NLP feladatok

gépi fordítás rule-based MT, statistical MT, neural MT

koreferencia feloldás mik utalnak ugyanarra az entitásra? *New York – Big Apple*

szentiment analízis vélemény, érzelem kinyerése

Schönherz

Magasabb szintű NLP feladatok

gépi fordítás rule-based MT, statistical MT, neural MT

koreferencia feloldás mik utalnak ugyanarra az entitásra? *New York – Big Apple*

szentiment analízis vélemény, érzelem kinyerése

kapcsolatkinyerés relation extraction, ki kinek a rokona?

Schönherz

Magasabb szintű NLP feladatok

gépi fordítás rule-based MT, statistical MT, neural MT

koreferencia feloldás mik utalnak ugyanarra az entitásra? *New York – Big Apple*

szentiment analízis vélemény, érzelem kinyerése

kapcsolatkinyerés relation extraction, ki kinek a rokona?

Schönherz

Magasabb szintű NLP feladatok

gépi fordítás rule-based MT, statistical MT, neural MT

koreferencia feloldás mik utalnak ugyanarra az entitásra? *New York – Big Apple*

szentiment analízis vélemény, érzelem kinyerése

kapcsolatkinyerés relation extraction, ki kinek a rokona?

kérdésmegválaszolás

szövegösszegzés text summarization

Magasabb szintű NLP feladatok

gépi fordítás rule-based MT, statistical MT, neural MT

koreferencia feloldás mik utalnak ugyanarra az entitásra? *New York – Big Apple*

szentiment analízis vélemény, érzelem kinyerése

kapcsolatkinyerés relation extraction, ki kinek a rokona?

kérdésmegválaszolás

szövegösszegzés text summarization

helyesírás-ellenőrzés

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
    - kézzel írt szabályokat keresünk

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik
   - jó pontosság, de milyen a fedés?

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik
   - jó pontosság, de milyen a fedés?
2. Gépi tanulás

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik
   - jó pontosság, de milyen a fedés?
2. Gépi tanulás
   - felügyelt vs. felügyeletlen

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik
   - jó pontosság, de milyen a fedés?
2. Gépi tanulás
   - felügyelt vs. felügyeletlen
   - a tanítóadat a legtöbb feladathoz nagyon költséges

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik
   - jó pontosság, de milyen a fedés?
2. Gépi tanulás
   - felügyelt vs. felügyeletlen
   - a tanítóadat a legtöbb feladathoz nagyon költséges
   - nem felügyelt módszerekhez sok adatunk van

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik
   - jó pontosság, de milyen a fedés?
2. Gépi tanulás
   - felügyelt vs. felügyeletlen
   - a tanítóadat a legtöbb feladathoz nagyon költséges
   - nem felügyelt módszerekhez sok adatunk van
   - *zaj!*

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik
   - jó pontosság, de milyen a fedés?
2. Gépi tanulás
   - felügyelt vs. felügyeletlen
   - a tanítóadat a legtöbb feladathoz nagyon költséges
   - nem felügyelt módszerekhez sok adatunk van
   - *zaj!*
   - itt pezseg jobban az élet

Schönherz

Hogyan oldjuk meg ezeket a feladatokat?

1. Szabályalapú
   - kézzel írt szabályokat keresünk
   - nagyon időigényes
   - nem skálázódik
   - jó pontosság, de milyen a fedés?
2. Gépi tanulás
   - felügyelt vs. felügyeletlen
   - a tanítóadat a legtöbb feladathoz nagyon költséges
   - nem felügyelt módszerekhez sok adatunk van
   - *zaj!*
   - itt pezseg jobban az élet
   - deep learning

Schönherz

Adatok

Schönherz

Adatok

- címkézett

Schönherz

Adatok

- címkézett
  - névelemekre:

Schönherz

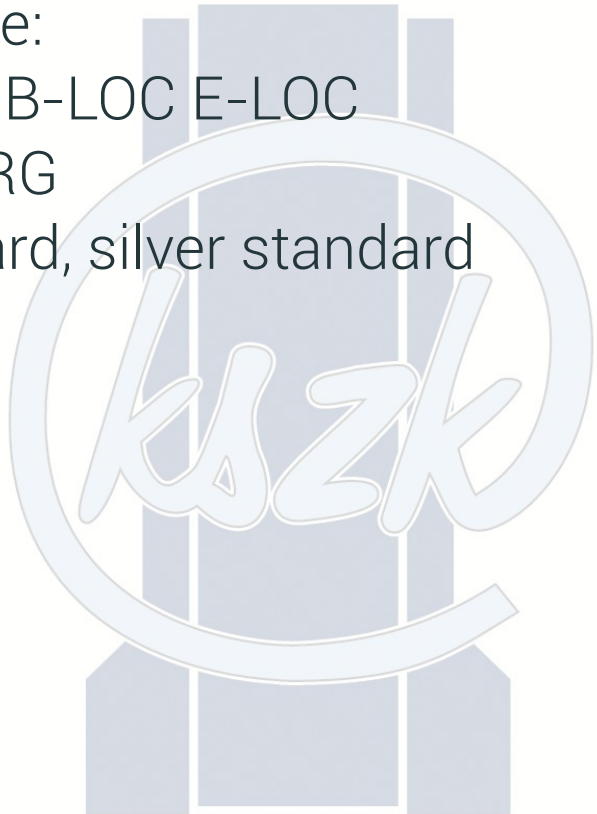
Adatok

- címkézett
  - névelemekre:
  - New York - B-LOC E-LOC
  - Google - ORG

Schönherz

Adatok

- címkézett
  - névelemekre:
  - New York - B-LOC E-LOC
  - Google - ORG
  - gold standard, silver standard



Schönherz

Adatok

- címkézett
  - névelemekre:
  - New York - B-LOC E-LOC
  - Google - ORG
  - gold standard, silver standard
- címkézetlen

Adatok

- címkézett
  - névelemekre:
  - New York - B-LOC E-LOC
  - Google - ORG
  - gold standard, silver standard
- címkézetlen
  - nyers szöveg
  - kontextusból nyerhetünk információt

Schönherz

Adatok

- címkézett
  - névelemekre:
  - New York - B-LOC E-LOC
  - Google - ORG
  - gold standard, silver standard
- címkézetlen
  - nyers szöveg
  - kontextusból nyerhetünk információt
- rengeteg standard adat

Schönherz

Adatok

- címkézett
  - névelemekre:
  - New York - B-LOC E-LOC
  - Google - ORG
  - gold standard, silver standard
- címkézetlen
  - nyers szöveg
  - kontextusból nyerhetünk információt
- rengeteg standard adat
- de sose elég, 6000+ nyelv :(

Schönherz

Technológiai háttér

Schönherz

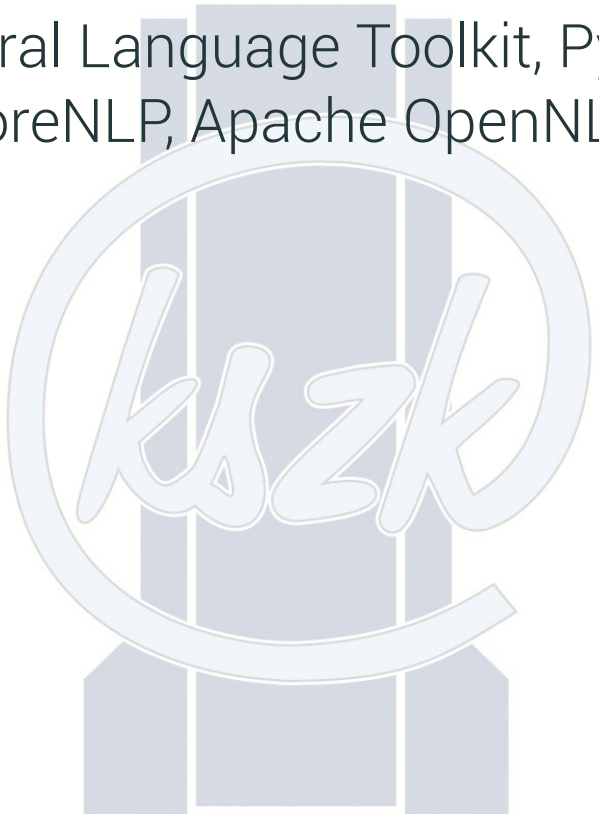
Technológiai háttér

- Python, Java

Schönherz

Technológiai háttér

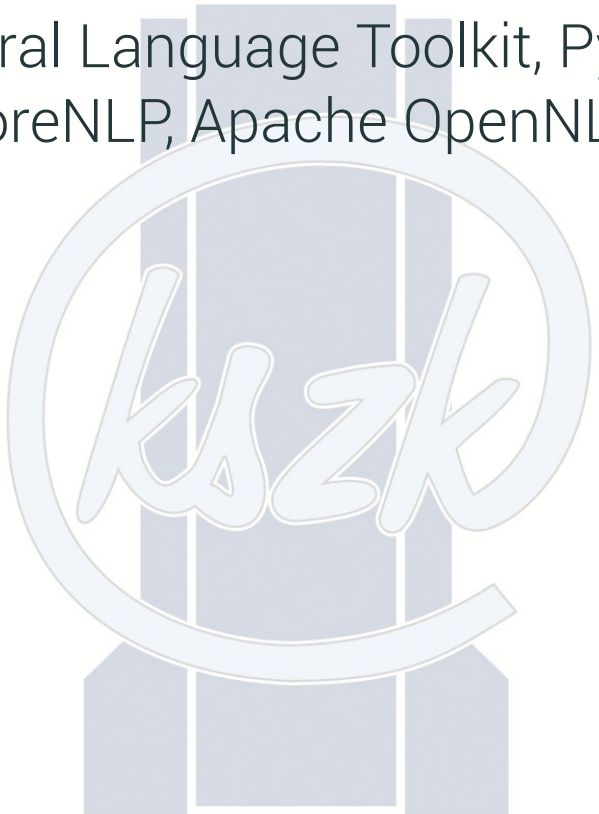
- Python, Java
  - NLTK (Natural Language Toolkit, Python)
  - Stanford CoreNLP, Apache OpenNLP (Java)



Schönherz

Technológiai háttér

- Python, Java
  - NLTK (Natural Language Toolkit, Python)
  - Stanford CoreNLP, Apache OpenNLP (Java)
- Linux (OSX)



Schönherz

Technológiai háttér

- Python, Java
  - NLTK (Natural Language Toolkit, Python)
  - Stanford CoreNLP, Apache OpenNLP (Java)
- Linux (OSX)
- gépi tanuláshoz: scikit-learn (Python), Weka (Java)

Schönherz

Technológiai háttér

- Python, Java
  - NLTK (Natural Language Toolkit, Python)
  - Stanford CoreNLP, Apache OpenNLP (Java)
- Linux (OSX)
- gépi tanuláshoz: scikit-learn (Python), Weka (Java)
  - deep learning: tensorflow, theano, torch, keras stb.

Schönherz

Technológiai háttér

- Python, Java
  - NLTK (Natural Language Toolkit, Python)
  - Stanford CoreNLP, Apache OpenNLP (Java)
- Linux (OSX)
- gépi tanuláshoz: scikit-learn (Python), Weka (Java)
  - deep learning: tensorflow, theano, torch, keras stb.
- plain text, TSV, esetleg XML

Schönherz

Technológiai háttér

- Python, Java
  - NLTK (Natural Language Toolkit, Python)
  - Stanford CoreNLP, Apache OpenNLP (Java)
- Linux (OSX)
- gépi tanuláshoz: scikit-learn (Python), Weka (Java)
  - deep learning: tensorflow, theano, torch, keras stb.
- plain text, TSV, esetleg XML
  - Linux CL toolokkal is elvégezhető sok egyszerű feladat (awk, sed)

Schönherz

Hogyan készítenénk helyesírásellenőrzőt?

Schönherz

Hogyan készítenénk helyesírásellenőrzőt?

Hogyan sorolnánk fel az összes magyar szót?

Schönherz

Hogyan készítenénk helyesírásellenőrzőt?

Hogyan sorolnánk fel az összes magyar szót?

Demo

Schönherz

Köszönöm a figyelmet

judit@aut.bme.hu

Demo: `https://gist.github.com/juditacs/4435129e6f79015ba98fba13f1736b84`

Önlab, szakdolgozat, diplomaterv, TDK lehetőség

Schönherz