

Szöveg- és Webbányászat házi feladat megoldási terv

Távközlési és Médiainformatikai Tanszék

|  |  |
| --- | --- |
| Készítette: | **Szántó Tamás, Benda Krisztián** |
| Neptun-kód: | **ET7D8H, J1CEI3** |
| Ágazat: | **Adat- és Médiainformatika** |
| E-mail cím: | [**tmas.szanto@gmail.com**](mailto:tmas.szanto@gmail.com)**,** [**krisztianbenda@gmail.com**](mailto:krisztianbenda@gmail.com) |
| Konzulens(ek): | **Dr. Szűcs Gábor** |
| E-mail címe(ik): | **szucs@tmit.bme.hu** |

Téma címe: Névelem felismerés

Feladat

A névelem-felismerést (named entity recognition) segítségével kinyerhetők egy adott korpuszon belül előforduló névelemek, s ezen belül a tulajdonnevek (személynevek, helyek, szervezetek és egyéb tulajdonnevek). A feladat angol nyelvű szövegben 7 típusú entitásnak a felismerése, melyek a következők:

event = esemény; geo = földrajzi entitás; gpe = geopolitikai entitás; obj = objektum, műtárgy; org = szervezet; per = személy; time = idő

A felismerendő entitások állhatnak 1 vagy akár több szóból is. Minden esetben az entitás első szavát külön detektálni kell, ennek jelzésére a B betű használandó (beginning): így B-event, B-geo, stb. címkékkel kell a megfelelő szavakat ellátni. Ha az entitások több szóból áll, akkor az összes többi I-vel jelölendő (inside), azaz I-event, I-geo, stb. címkék; így összesen az egyéb (O) címkével együtt 15 osztálycímke adódik.

Példa:

Indian border security forces are accusing their Pakistani counterparts of lobbing at least four rockets into northern Punjab state.

Indian: B-gpe, Pakistani: B-gpe, Punjab: B-geo, a többi pedig O címkéjű.

**2018/2019. 1. félév**

Vállalt részfeladatok:

1. Létező megoldások vizsgálata és kipróbálása
2. Órán tanult módszerek áttanulmányozása
3. A tapasztalatok alapján prototípus elkészítése
4. Az elkészült megoldás javítása, továbbfejlesztése
5. Bemutató elkészítése és előadása

Megoldási ötletek:

* Az általánosabb feldolgozási folyamat a következő:
  + Tokenizálás, normalizálás/szótövezés, névelem detektálás, névelem normalizálás
* Névelem detektálására az alábbi megközelítéseket ismerjük
  + Szótár alapú
    - Összes entitás összes formáját össze kell gyűjteni
    - Nagy tudásbázis vagy annotált korpusz szükséges
  + Szabály alapú
    - Mintákat kell írni az entitások illesztéséhez
    - Téma specifikus tudás szükséges
  + Statisztikai modell alapú
    - Valószínűségek hozzárendelése a szövegrészekhez
    - Sok tanuló példány szükséges
    - Előny: téma független tudás
* Ezen detekciók használatának előnyeit fogjuk felmérni és ezalapján a legmegfelelőbbet kiválasztani.

Létező eszközök, módszerek:

* [spaCy](https://spacy.io): python library, statisztikai modell alapú NER, sokféle kategória támogatott, továbbtanítható saját kategóriákkal
* [Stanford NER is a Named Entity Recognizer](https://nlp.stanford.edu/software/CRF-NER.shtml): Java library, kevés alapból támogatott kategória
* [Named-Entity-Recognition-BLSTM-CNN-CoNLL](https://github.com/mxhofer/Named-Entity-Recognition-BidirectionalLSTM-CNN-CoNLL): implementáció [ehhez a cikkhez](https://arxiv.org/abs/1511.08308), Keras

Használni tervezett technológiák:

* Elsősorban Python 3-at szeretnénk használni
* Kisebb részfeladatok/algoritmusok kipróbálásához opcionálisan RapidMinder-t is igénybe vennénk
* A párhuzamos munkavégzést a GitHub segítségével oldanánk meg