01 Tokenizacija

Tokenizacija je postopek deljenja besedila na posamezne pojavnice (besede, številke, ločila). Pri strojnem označevanju korpusov v slovenskem prostoru trenutno uporabljamo označevalnik CLASSLA-Stanza oz. vanj vključeni tokenizator Obeliks. Pravilom, na katerih je osnovan strojni označevalnik, sledi tudi ročni pregled.

- Označevalne smernice
- Reference in povezave

Označevalne smernice

V tem poglavju so predstavljene označevalne smernice oz. načela za tokenizacijo.

 Presledek je glavna ločnica med pojavnicami.

 Besede, ki jih lahko pišemo skupaj ali narazen, ne da bi spremenile pomen (npr. **kdorkoli**, **kdor koli**), se ravnajo po prvem načelu, tj. tvorijo eno ali dve pojavnici – odvisno od presledka.

 V procesu tokenizacije so vse pojavnice prepoznane kot alfanumerične ali pa kot znaki.

 Znaki so določeni s pomočjo vnaprej določenega seznama, na katerem so ločila, simboli in podobno (odvisno od označevalnega sistema, recimo UD ali JOS/MULTEXT-East). Ta seznam je vključen v tokenizator, sestavljen pa je le iz posameznih znakov. Zaporedja dveh ali več znakov (npr. **?!**) se obravnavajo kot zaporedja ločenih znakov.

 Če niz alfanumeričnih znakov med dvema presledkoma vsebuje znak, se običajno razdeli na več pojavnic (npr. **AC/DC** in **Micro$oft** sta razdeljena na tri pojavnice: 'AC' '/' 'DC' ter 'Micro' '$' 'oft').

 Vendar veljajo naslednje izjeme, pri katerih znak postane del alfanumerične pojavnice:

  apostrof postane del alfanumerične pojavnice, če je zapisan obojestransko stično (npr. **O'Brian**, **mor'va**).

  vejica in dvopičje postaneta del alfanumerične pojavnice, če sta zapisana obojestransko stično in pojavnico sestavljajo same številke (npr. **30:00**, **200,000,000**)

  pomišljaj postane del alfanumerične pojavnice, če je je zapisan obojestransko stično in če:

    • je levi del kratica (zapisana z velikimi črkami), ena sama črka ali številka

    • je desni del pripona ali pregibna končnica; končni seznam možnih pripon in končnic je vključen v tokenizator (npr. **OZN-ovski** "podoben Združenim narodom", **a-ju** "črki a", **15-i** "petnajsti"))

  pika postane del alfanumerične pojavnice, če je:

    • zapisana obojestransko stično in niz vsebuje samo številke (npr. **1.2**)

    • zapisana levo stično in je del kratice ali vrstnega števila (npr. **dr.**, **4.**, **IV.**); končni seznam možnih kratic je vključen v tokenizator.

  Vsi znaki postanejo del ene same alfanumerične pojavnice v nizih, ki so s pomočjo regularnega izraza prepoznani kot URL-ji ali naslovi.

Informacija o tem, da pojavnici ne sledi presledek (npr. **d.o.o.** proti **d. o. o.**), je navedena s SpaceAfter=No v stolpcu MISC.

Reference in povezave

V tem poglavju so zbrane relevantne reference in povezave na projekte, v katerih se je označevalni sistem razvijal ter uporabljal.

Projekti, na katerih se je označevalni sistem razvijal oz. uporabljal

JOS - Jezikoslovno označevanje slovenskega jezika: metode in viri

Sporazumevanje v slovenskem jeziku

Universal Dependencies

Janes - Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine

Razvoj slovenščine v digitalnem okolju

Orodje Obeliks za tokenizacijo in stavčno segmentacijo

https://github.com/clarinsi/obeliks

Reference

Krek, Simon; Erjavec, Tomaž; Dobrovoljc, Kaja; Gantar, Polona; Arhar Holdt, Špela; Čibej, Jaka in Brank, Janez. The ssj500k training corpus for Slovene language processing. V: Fišer, D. in Erjavec, T. Jezikovne tehnologije in digitalna humanistika: zbornik konference: 24.-25. september 2020, Ljubljana, Slovenija. Ljubljana: Inštitut za novejšo zgodovino, 2020. Str. 24–33.

http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene-Language-Processing.pdf [PDF]