08 Imenske entitete

Imenske entitete (angl. named entities; NE) so samostalniki in samostalniške besedne zveze, ki identificirajo neko osebo, lokacijo, organizacijo ali drug edinstven objekt v realnem prostoru in času (poleg tega v širšem smislu tudi svojilne pridevnike, izpeljane iz osebnega lastnega imena, npr. DERIV-PER[Obamova] izvolitev). Na ortografski ravni so pogosto izražene z veliko začetnico (Slovenska tiskovna agencija) ali kratico (STA), vendar pa velika začetnica in kratica ne označujeta samo imenskih entitet (npr. BDP). Identifikacija imenskih entitet v besedilu je pomembna za luščenje informacij iz besedila, odkrivanje koreferenčnosti, analiziranje sentimenta, ekstrakcijo informacij, povezav in dogodkov ter drugih nalog, povezanih s procesiranjem naravnega jezika.

- Predstavitev oznak
- Označevalne smernice
- Reference in povezave

Predstavitev oznak

V tem poglavju so navedene in strnjeno predstavljene oznake za imenske entitete (angl. *named entities*; NE). Podrobnejšo predstavitev najdete v smernicah v poglavju Označevalne smernice.

Kategorija	Podkategorija	Primeri	Ne spada v to kategorijo
PER	Oseba (ime in/ali priimek)	Janez Novak, da Vinci, Ludvik XIV.	dr., gospa, sv.
	Ime domače živali	Fifi	
	Umetniško ime, psevdonim	Madonna, mati Terez(ij)a, Banksy	
	Imena fiktivnih oseb (iz filmov, knjig ipd.)	Ana Karenina, Rdeča kapica	
	Vzdevki	(Boštjan Gorenc -) Pižama, Zvezdica89	
	Poimenovane skupine ljudi (družina ali lokalno omejena skupnost)	Angleži, Nemec, Ljubljančan; Novakovi	
	Omembe oz. sklici na Twitterju	@pizama, @Nike	
DERIV-PER	Svojilni pridevnik iz osebnega imena	Novakov (pes)	Alzheimerjeva (bolezen)
ORG	Organizacije	EU, Nato, Rimskokatoliška cerkev	parlament, vlada
	Podjetja	Microsoft, Pasadena d.o.o.	
	Upravljalci letališč	Aerodrom Ljubljana	Letališče Jožeta Pučnika
	Izobraževalne ustanove	Filozofska fakulteta	
	Instituti	Institut “Jožef Stefan”	
	Muzeji, knjižnice	Prirodoslovni muzej	
	Gledališča, kinematografi ipd.	MGL, Kinodvor	
	Mediji (TV, radio, časopisi)	Dnevnik, Delo, Radio Center	
	Restavracije, hoteli, lokali	Kavarna Zvezda, [hH]otel Lev	
	Zdravstvene ustanove	[zZ]dravstveni dom Ribnica	

Kategorija	Podkategorija	Primeri	Ne spada v to kategorijo
	Glasbene skupine	U2, Beatli, [aA]nsambel Avsenik	
	Institucije	[oO]bčina Piran, NPK	
	Politične stranke in druga civilna združenja	DeSUS, Zveza potrošnikov Slovenije	
	Športni klubi, društva in združenja	(HDD SIJ) Acroni Jesenice, (FC) Barcelona	
	Kulturne organizacije (tudi amaterske)	[mM]ešani pevski zbor Divača	
LOC	Nebesna telesa (planeti ipd.)	Mars, Andromeda, Halleyjev komet	
	Celine	Južna Amerika	
	Države, dežele (pretekle in sedanje)	Slovenija, Združene države (Amerike)	EU
	Regije	Primorska, Valonija, Nova Anglija	
	Mesta in predeli mest, kraji in deli krajev	Ljubljana, Šiška, Vrhnika, Na klancu	
	Ulice, trgi	Jamova cesta 39	A2, gorenjska AC
	Nakupovalna središča	Citypark, Supernova	
	Letališča	Letališče Jožeta Pučnika	
	Cerkve (kot poimenovane stavbe)	[cC]erkev sv. Nikolaja	Rimskokatoliška cerkev
	Krajevne znamenitosti (kulturne, naravne)	Tromostovje, Triglavski narodni park	
	Druge poimenovane zgradbe (brez org. strukture)	[kK]ulturni dom Ljubno, WTC 2	Cankarjev dom (ima org. strukturo, npr. direktorja)
	Gore, jezera, reke in druge poimenovane georgrafske entitete	Triglav, Blejsko jezero, Sava, Logarska dolina	
MISC	Sistemi, programi, aplikacije	Windows 10, Word, Android 5.1 Lollipop	.docx, pdf, OCR
	Naslovi knjig, filmov, nanizank, slik in drugih umetniških del; naslovi dokumentov ipd. (tudi imena zakonov), oddaje	Vojna in mir, Ko jagenjčki obmolknejo, Sopranovi, Guernica; Uradni list RS	

Kategorija	Podkategorija	Primeri	Ne spada v to kategorijo
	Registrirana imena ali modeli naprav (avti, mobiteli, računalniki, igre ipd.) in drugi komercialni izdelki (znamke)	Galaxy Note 7, Nokia Lumia 950, Toyota RAV4, Minecraft, Človek ne jezi se	
	Imena prireditev in drugih dogodkov	Oskarji, Zlata lisica, 10. mednarodna konferenca Jezikovne tehnologije	shod nacifašistov
	Imena projektov	Obzorje 2020	
	Borzni indeksi	SBI20, Dow Jones, Nasdaq	Bonitetne ocene (AAA)

Označevalne smernice

V tem poglavju so zbrane označevalne smernice za imenske entitete.

Različica 1.1

projekt Janes - Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine
ZUPAN, Katja; LJUBEŠIĆ, Nikola in ERJAVEC, Tomaž, 2017: *Smernice Janes-NER za označevanje imenskih entitet v slovenskem jeziku*: Različica 1.1. [PDF]

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

MUC-6 Named Entity Task Definition

CONLL 2003

BSNLP 2017 shared task

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

Razvoj slovenščine v digitalnem okolju

Reference

Marc Reznicek: *Linguistische Annotation von Nichtstandardvarietäten / Guidelines und „Best Practices*" *Guidelines NER* (version 1.5).

https://www.linguistik.huberlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5

LDC - Linguistic Data Consortium: ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 6.6 2008.06.13, http://projects.ldc.upenn.edu/ace (Dostopano dne: 2. november 2020).