Tehnike predobdelave besedil v procesiranju naravnega jezika

Mladen Borovič, Jani Dugonik

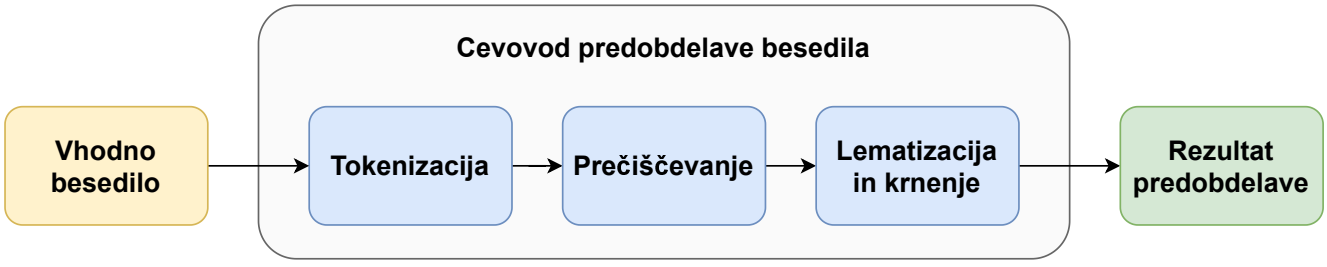
→ **Procesiranje naravnega jezika je v zadnjih nekaj letih postalo zelo prepoznavno področje računalništva. Naravni jezik je jezik, ki ga ljudje uporabljamo za komunikacijo. Pojavilo se je veliko novih metod za obdelavo naravnega jezika v različnih oblikah, kot so, recimo, besedila in zvočni posnetki. Te metode so prav tako postale del našega vsakdana s pomočjo pametnih storitev, ki jih dnevno uporabljamo. Danes lahko z govornimi ukazi izvajamo opravila na mobilnih napravah, s pametnimi iskalniki lahko na podlagi našega vhodnega besedila najdemo ustrezne vsebine, prav tako pa lahko brez večjih težav strojno prevajamo besedila v skoraj vse jezike. Na spletu obstaja kar nekaj prosto dostopnih spletnih prevajalnikov, kot so Google Translate [2], Microsoft Bing [3], Amebis Presis [1] in PONS [5].**

V tem prispevku se bomo osredotočili na sisteme, ki naravni jezik procesirajo v obliki besedila. Pri tem bomo podrobneje predstavili tehnike predobdelave besedil, ki navadno nastopijo kot prvi korak v takšnih sistemih. Predobdelava besedil je ključnega pomena za vse nadaljnje korake sistemov procesiranja besedil, saj neposredno vpliva na kakovost rezultatov.

Vhodno besedilo je potrebno najprej ustrezno preoblikovati v obliko, ki je primerna za nadaljnjo obdelavo. V ta namen uporabljamo cevovod predobdelave besedil *(angl. text preprocessing pipeline)*, ki ga sestavlja kombinacija različnih tehnik predobdelave. V nadaljevanju si poglejmo zasnovo takšnega cevovoda in nekaj najpogostejših tehnik predobdelave, ki se uporabljajo v praksi.

Cevovod predobdelave besedil

Predobdelava besedil je postopek, kjer v določenem zaporedju nad vhodnim besedilom izvajamo različne tehnike predobdelave besedil. Gre za transformacije vhodnega besedila, kjer po vsaki izvedeni teh-



SLIKA 1.
Primer cevovoda predobdelave besedil.

niki predobdelave besedil kot rezultat dobimo obdelano besedilo, ki je bolj primerno za nadaljnje postopke procesiranja besedil. Predobdelavo besedil si najlažje predstavljamo kot cevovod, kjer imamo na začetku neobdelano vhodno besedilo, na koncu pa dobimo obdelano izhodno besedilo. Slika 1 prikazuje strukturo takšnega cevovoda z zaporedjem nekaj napogostejših tehnik predobdelave besedil.

Tehnike, ki se v cevovodu predobdelave besedil najpogosteje uporabljajo, zajemajo (v tem vrstnem redu) tokenizacijo, prečiščevanje ter lematizacijo in krnjenje besedila. Vsaka izmed omenjenih tehnik predobdelave je samostojna enota znotraj cevovoda, ki nad besedilom na vhodu izvede ustrezno transformacijo in vrne izhodno obdelano besedilo. Zaporedje tehnik predobdelave se lahko v cevovodu tudi spremeni. To je največkrat pogojeno z jezikom, ki ga obdelujemo. V tem prispevku se bomo omejili na slovenski jezik.

Tokenizacija

Prvi korak predobdelave besedil je tokenizacija. To je proces delitve celotnega vhodnega besedila na manjše dele – žetone *(angl. token)*. Ponavadi govorimo o delitvi na besede, poznamo pa tudi druge delitve, kot sta npr. delitvi na besedne zveze in besedne n-grame. Pri delitvi na besedne zveze kot žeton uporabimo več besed. Delitev na besedne n-grame je podobna delitvi na besedne zveze, le da s številom n določimo, koliko besed ostane v besedni zvezi. Sicer n-grame uporabljamo tudi na nivoju besed, kjer ohranimo n črk v besedi. Slika 2 prikazuje razlike med različnimi delitvami na stavku Danes je lep dan.

Sam postopek delitve poteka na podlagi vnaprej določenih pravil za delitev. Za delitev na besede, ki je najpogosteje uporabljen način delitve, praviloma uporabimo pravilo deljenja s pomočjo znakov za presledke. Pri tem odstranimo tudi ponavljajoče zaporedne presledke, tabulatorje in ločila. Rezultat tokenizacije je seznam besed, ki se pojavijo v vhodnem besedilu.

Prečiščevanje

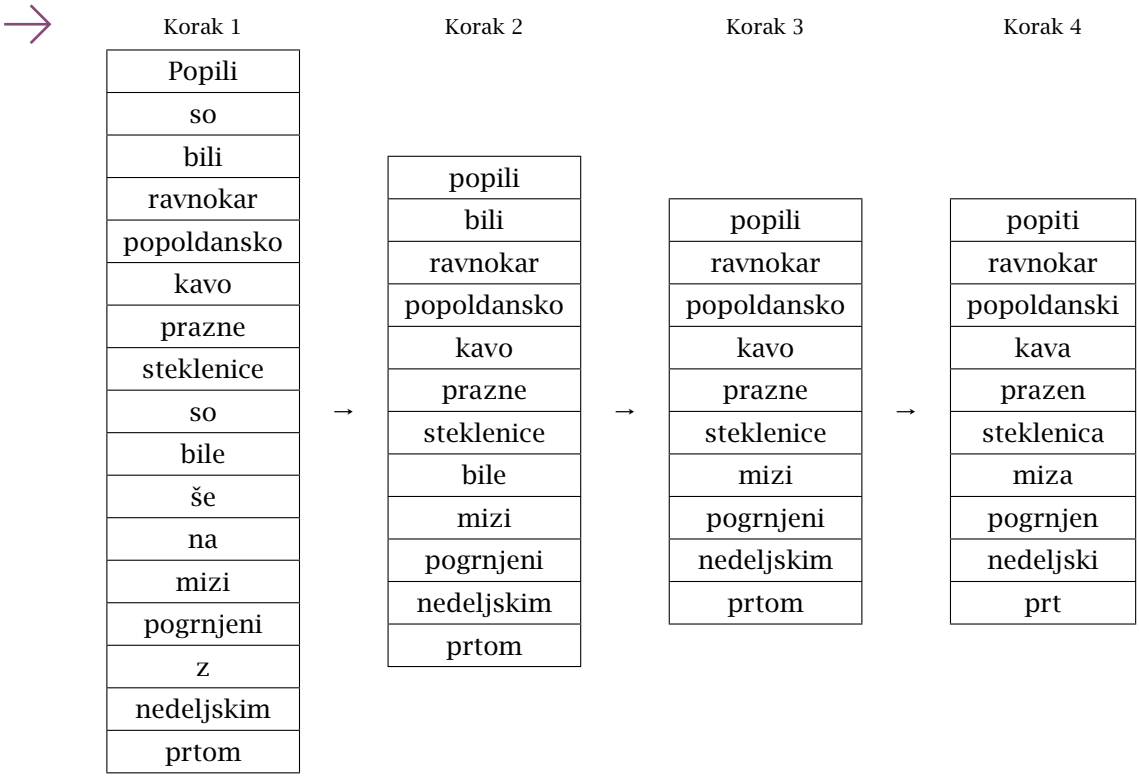
Naslednji korak predobdelave besedil je prečiščevanje. Besede iz seznama, pridobljenega s tokenizacijo, dodatno spreminjamo in odstranjujemo iz seznama. Najprej vse črke v besedah iz seznama pretvorimo v male črke. Takoj zatem ponavadi odstranimo tudi besede, ki so krajše oz. daljše od določenega števila znakov. Primer takšnega odstranjevanja je, recimo, odstranjevanje vseh besed krajših od treh znakov in daljših od petnajst znakov. Nazadnje zelo pogosto odstranimo najpogosteje uporabljene besede našega izbranega jezika. Gre za t. i. blokirane besede *(angl. stopwords)*, med katere spadajo vezniki, zaimki, prislovi, nekatere pridevniške besede in vse ostale besede, ki se v izbranem jeziku najpogosteje pojavijo. Z odstranjevanjem blokiranih besed odstranimo šum v besedilu, saj želimo ohraniti le tiste besede, ki vhodnemu besedilu dajejo največ vsebine. Rezultat je torej seznam prečiščenih besed, ki vsebujejo vsebino vhodnega besedila.

Lematizacija in krnjenje

V zadnjem koraku cevovoda predobdelave besedil nad seznamom prečiščenih besed izvedemo postopek lematizacije ali postopek krnjenja. Lematizacija

Danes je lep dan	
Tip delitve	**Rezultat delitve**
na besede	danes, je, lep, dan
na besedne zveze	danes je, danes je lep, danes je lep dan, je lep, je lep dan, lep dan
na n-grame ($n = 2$)	da, an, ne, es, s_, _j, je, e_, _l, le, ep, p_, _d, da, an
na besedne n-grame ($n = 2$)	danes je, je lep, lep dan

SLIKA 2.
Primer različnih delitev pri tokenizaciji na stavku Danes je lep dan. Podčrtaj pri delitvi na n-grame je uporabljen kot znak za presledek.



SLIKA 3. Potek delovanja cevovoda predobdelave besedil

in krnjenje sta podobna postopka, vendar je med njima zelo pomembna razlika. Lematizacija besedo preoblikuje v njeno osnovno obliko, krnjenje pa besedi zgolj odreže končnico. Tako z lematizacijo dobimo lemo, s krnjenjem pa krn besede. Dober primer razlike je beseda boljši. S krnjenjem bomo besedo pretvorili v besedo bolj, z lematizacijo pa v njeno osnovno obliko – dober. To je zelo pomembno pri morfološko bogatih jezikih, kot je slovenščina, zato pri takšnih jezikih v tem koraku pogosteje uporabljamo lematizacija. Slovenščina je oblikoslovno (morfološko) izjemno bogat jezik, saj se veliko besednih vrst pregiba (samostalniške in pridevniške besede, glagoli, zaimki, števniki). Tako s sklanjanjem, spreganjem in stopnjevanjem besede dobivajo različne končnice (morfeme). Krnjenje tako bolj uporabljamo za manj morfološko bogate jezike, kot je npr. angleški jezik. Z lematizacijo torej vse besede iz seznama prečiščenih besed pretvorimo v njihove osnovne oblike. Izhod cevovoda je spremenjen seznam besed in predstavlja rezultat celotne predobdelave vhodnega besedila.

Primer

Poglejmo si delovanje cevovoda predobdelave besedil na primeru (slika 3) dela besedila Cankarjevega romana Tujci:

> *Popili so bili ravnokar popoldansko kavo: prazne steklenice so bile še na mizi, pogrnjeni z nedeljskim prtom.*

Najprej izvedemo tokenizacijo, kjer bomo vhodno besedilo razdelili na besede, izhod pa bo seznam besed. Pri tem bomo odstranili tudi ločila (korak 1). Nato sledi prečiščevanje, kjer bomo vse velike črke pretvorili v male črke, iz seznama pa bomo odstranilo vse besede krajše od treh znakov, vse besede daljše od petnajst znakov (korak 2) in vse blokirane besede (korak 3). Seznami blokiranih besed za najpogostejše jezike najdemo v prosto dostopni Python knjižnici NLTK [4]. Nekatere besede, ki spadajo v seznam blokiranih besed za slovenščino, so: ali, ampak, bodisi, in, kajti, namreč, ne, niti, oziroma, pa.

Na spletu so še prosto dostopni seznami blokiranih besed za slovenščino na Wikiversity [8] in repozitoriju GitHub [7]. Na koncu izvedemo še lematizacijo, kjer vse besede iz seznama pretvorimo v osnovno obliko (korak 4). Rezultat koraka 4 je tudi izhod cevovoda. Kot je razvidno iz podanega primera, je izhodno besedilo krajše, hkrati pa ohranja vsebino originalnega besedila. S predobdelavo besedila smo uspeli izluščiti najpomembnejše besede oz. značilke, s katerimi lahko začnemo nadaljnje procesiranje.

V tem prispevku predstavljen cevovod predobdelave besedil je le eden izmed možnih načinov predobdelave besedil, saj lahko v cevovodu spreminjamo tehnike predobdelave besedil in njihovo zaporedje izvajanja. Postopek predobdelave je velikokrat pogojen s postopkom nadaljnjega procesiranja besedil. Nekateri izmed teh postopkov zahtevajo točno določeno obliko besedila na vhodu, za kar seveda poskrbimo že pri predobdelavi. Obstaja še nekaj tehnik predobdelave besedil, ki jih v tem prispevku nismo opisali, saj jih ponavadi uporabljamo v primeru specifičnega procesiranja besedil. Med te tehnike predobdelave besedil spadata, recimo, normalizacija besedila in preverjanje črkovanja. Pri normalizaciji besedila se določene besede razširijo v bolj pomenljivo obliko. Dober primer normalizacije so kratice, ki jih razširimo v obliko pred krajšavo (npr. STA v Slovenska tiskovna agencija). Pri preverjanju črkovanja vsako besedo preverimo za napake v črkovanju in jo nato ustrezno popravimo.

Čeprav je večina orodij za predobdelavo besedil v osnovi razvitih za angleški jezik, je podpora zelo dobra tudi za slovenski jezik. Na voljo je kar nekaj prostodostopnih orodij: za programski jezik Python je dobra izbira knjižnic Gensim [10], NLTK [4] in Classla [9], s katerimi lahko le v nekaj vrsticah kode sami sprogramiramo cevovod predobdelave besedil. Primer takšne implementacije v programskem jeziku Python je na voljo na javno dostopnem repozitoriju GitHub [6]. Kljub temu, da podpora za slovenski jezik v teh orodjih še ni optimalna, se iz leta v leto stanje izboljšuje, saj se vztrajno veča število ljudi, ki se v Sloveniji ukvarjajo s področjem procesiranja naravnega jezika.

Literatura

[1] *Amebis Presis*, dostopno na presis.amebis.si/prevajanje/, ogled 17. 1. 2022.

[2] *Google Translate*, dostopno na translate.google.com, ogled 17. 1. 2022.

[3] *Microsoft Bing*, dostopno na www.bing.com/translator, ogled 17. 1. 2022.

[4] *NLTK – Natural Language Toolkit*, dostopno na www.nltk.org/, ogled 17. 1. 2022.

[5] *PONS*, dostopno na sl.pons.com/prevod-besedisca, ogled 17. 1. 2022.

[6] *Procesiranje naravnega jezika – GitHub*, dostopno na github.com/procesiranje-naravnega-jezika, ogled 18. 1. 2022.

[7] *Seznam blokiranih besed za slovenščino – GitHub*, dostopno na github.com/stopwords-iso/stopwords-sl, ogled 17. 1. 2022.

[8] *Seznam blokiranih besed za slovenščino – Wikiversity*, dostopno na sl.wikiversity.org/wiki/Seznam_slovenskih_praznih_besed_za_izdelavo_besednega_oblaka, ogled 17. 1. 2022.

[9] N. Ljubešić in K. Dobrovoljc, *What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian*, Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Florence, Italy, 2019, Association for Computational Linguistics, 29–34.

[10] R. Řehůřek in P. Sojka, *Software Framework for Topic Modelling with Large Corpora*, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, 2010, ELRA, 45–50.

× × ×