BIOLAB AND COLLABORATORS

UVOD V RUDARJENJE BES

BIOLAB

PUBLISHED BY BIOLAB

TUFTE-LATEX.GOOGLECODE.COM

First printing, October 2021

Contents

Delotoki v Orangeu 5

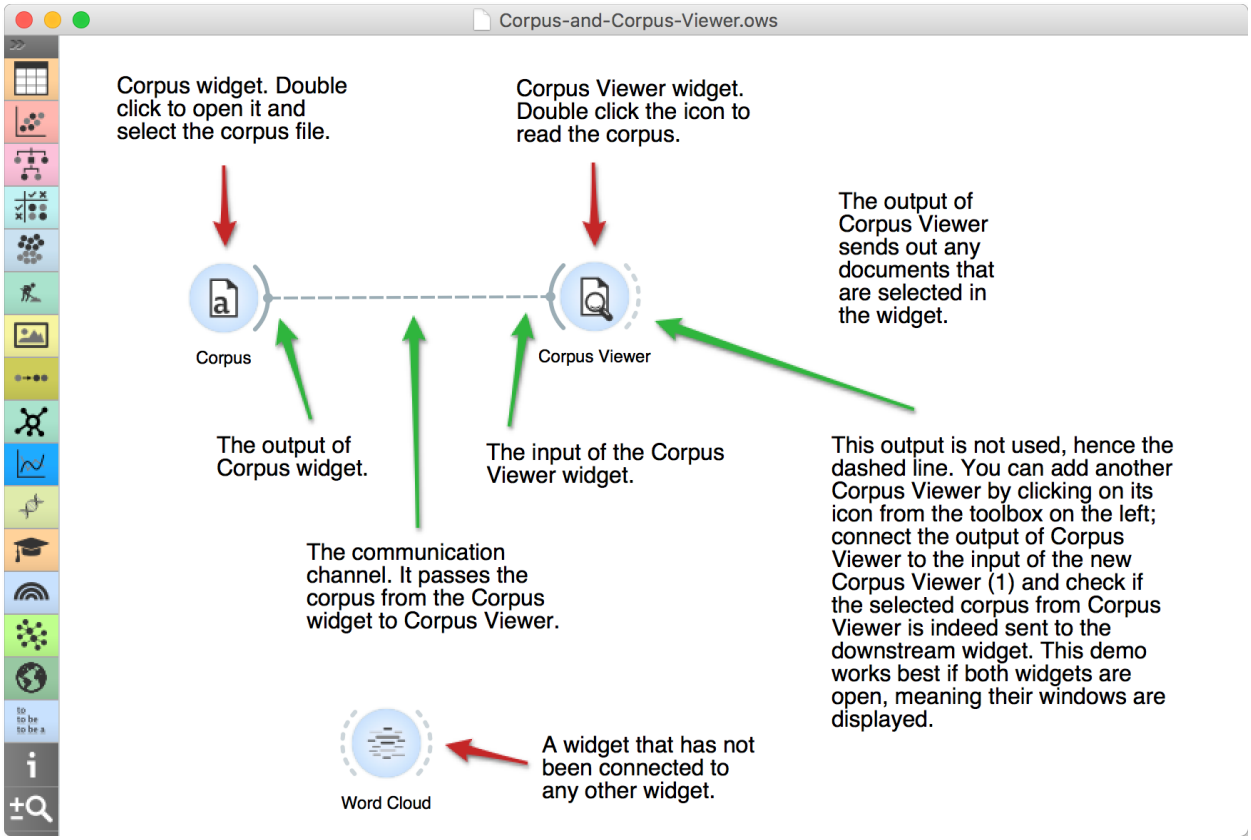
Priprava besedil 8

Bibliography 11

Index 12

Delotoki v Orangeu

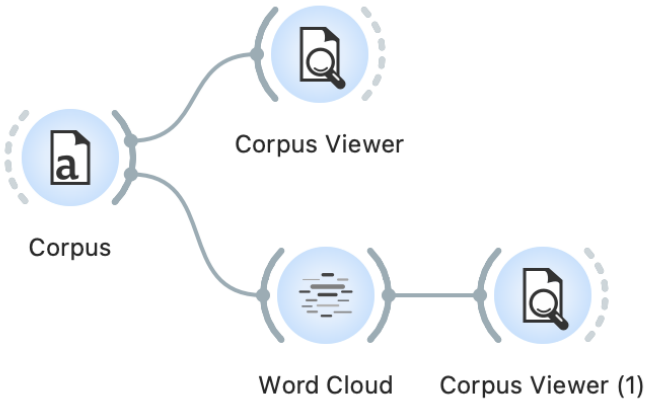
Delotoki v Orangeu so sestavljeni iz komponent, ki berejo, procesirajo in prikazujejo podatke. Te komponente imenujemo gradniki oz gradniki. Na desni je prazen prostor, t.i. platno. Nanj polagamo gradnike. Gradniki v Orangeu komunicirajo preko komunikacijskih kanalov. Izhod iz enega gradnika je uporabljen kot vhod za drug gradnik.



Slika zgoraj kaže preprost delotok z dvema povezanima gradnikoma in enim gradnikom brez povezav. Izhodi gradnika so na desni strani, vhodi pa na levi.

Delotoke sestavljamo tako, da polagamo gradnike na platno in jih povezujemo. Povezavo ustvarimo tako, da potegnemo črto od izhodnega v vhodni gradnik. Izhodi gradnika so na desni, vhodi pa na levi strani. V zgornjem delotoku gradnik *Corpus* pošilja podatke v gradnik *Corpus Viewer*.

Pričnite z gradnjo delotoka, ki vsebuje gradnik *Corpus*, dva gradnika *Corpus Viewer* in gradnika *Word Cloud*:

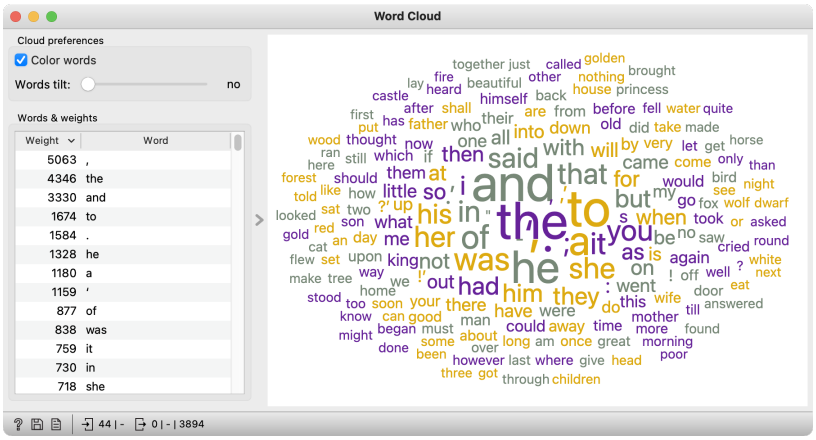


Delotok z gradnikom Corpus bere podatke iz računalnika in jih pošlje v gradnika Corpus Viewer in Word Cloud. Corpus Viewer prikaže besedila v iskalniku, Word Cloud pa izriše najpogostejše besede. Dokumenti, ki vsebujejo izbrano besedo iz gradnika Word Cloud, so prikazani v gradniku Corpus Viewer (1).

Gradnik *Corpus* bere podatke iz lokalnega diska. Odprite Corpus tako, da dvakrat kliknete na ikono. Dodatek Text že vsebuje nekaj prednaloženih korpusov. Iz teh ("Browse") izberite *Grimms-tales-selected.tab*, korpus z izbranimi Grimmovimi pravljicami.

Orangevi delotoki se pogosto pričnejo z gradnikoma File ali Corpus. Korpus Grimmovih pravljic vsebuje 44 dokumentov. Polje "Used text features"na levi pove, katere stolpce bomo smatrali kot del besedila, medtem ko polje na desni vsebuje dodatne informacije (naslov, povzetek, itd.).

Odprite *Word Cloud*. Word Cloud (oblak besed) prikaže pogostost besed v dokumentih, kjer so pogosteješe besede prikazane sorazmerno večje. Izberite besedo v oblaku in jo pošiljte v Corpus Viewer (1). Sedaj lahko pregledate samo dokumente, ki vsebujejo izbrano besedo.

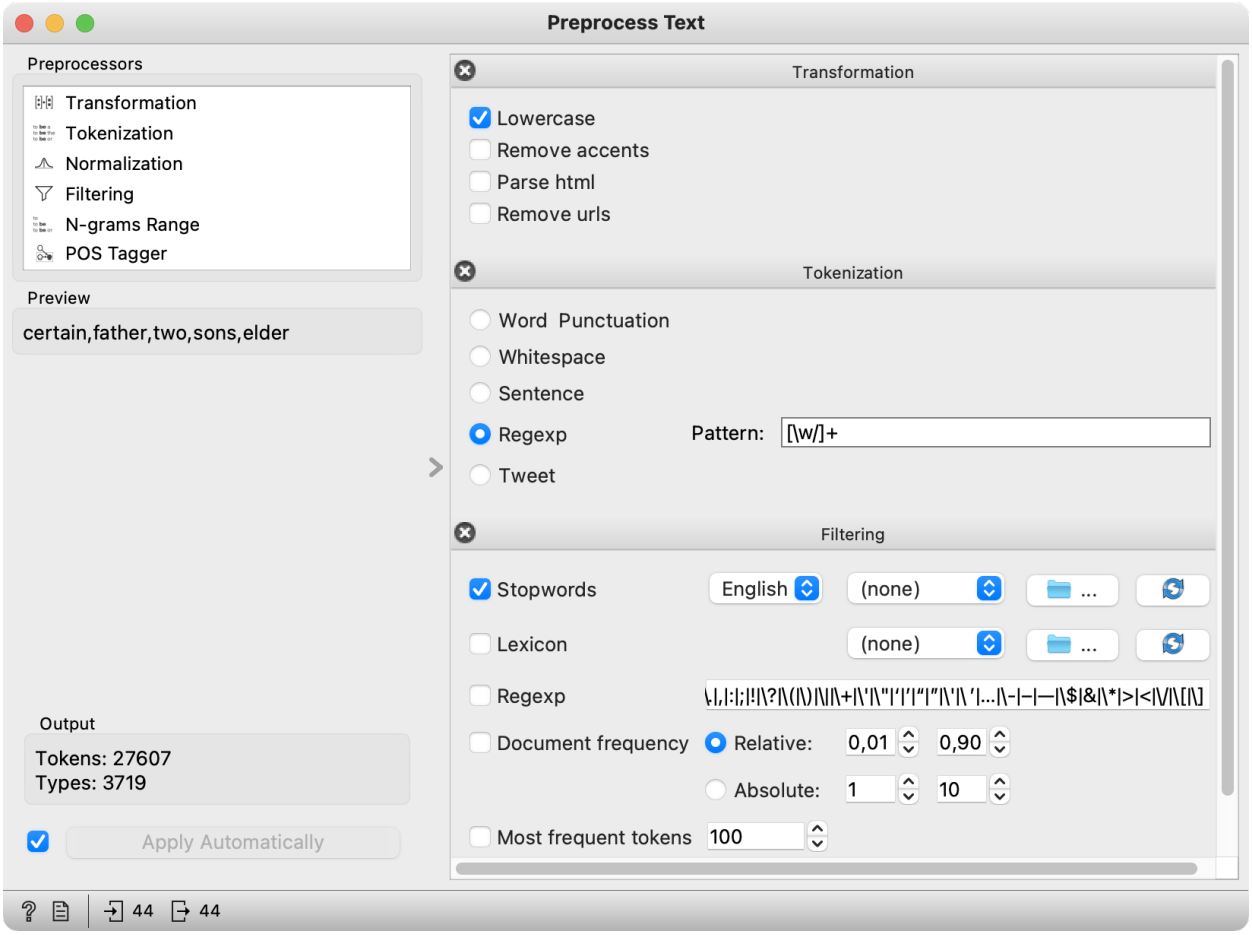


Počakajmo malo! Ta oblak besed je živa groza! Vidimo lahko celo kopico semantičnih smeti. Ali obstaja kakšen način, da to nekako uredimo?

Seveda! Odstraniti moramo vse delčke, ki ne vsebujejo nikakršne informacije, točneje ločila in odvečne besede (členke, pomožne glagole, veznike).

Priprava besedil

Word Cloud je preprosto prikazal vse besede in simbole, ki obstajajo v besedilu. Ampak običajno to ni to, kar hočemo. Ponavadi želimo prikazati zgolj pomenske enote, torej semantično bogate besede. Zato potrebujemo predprocesiranje.



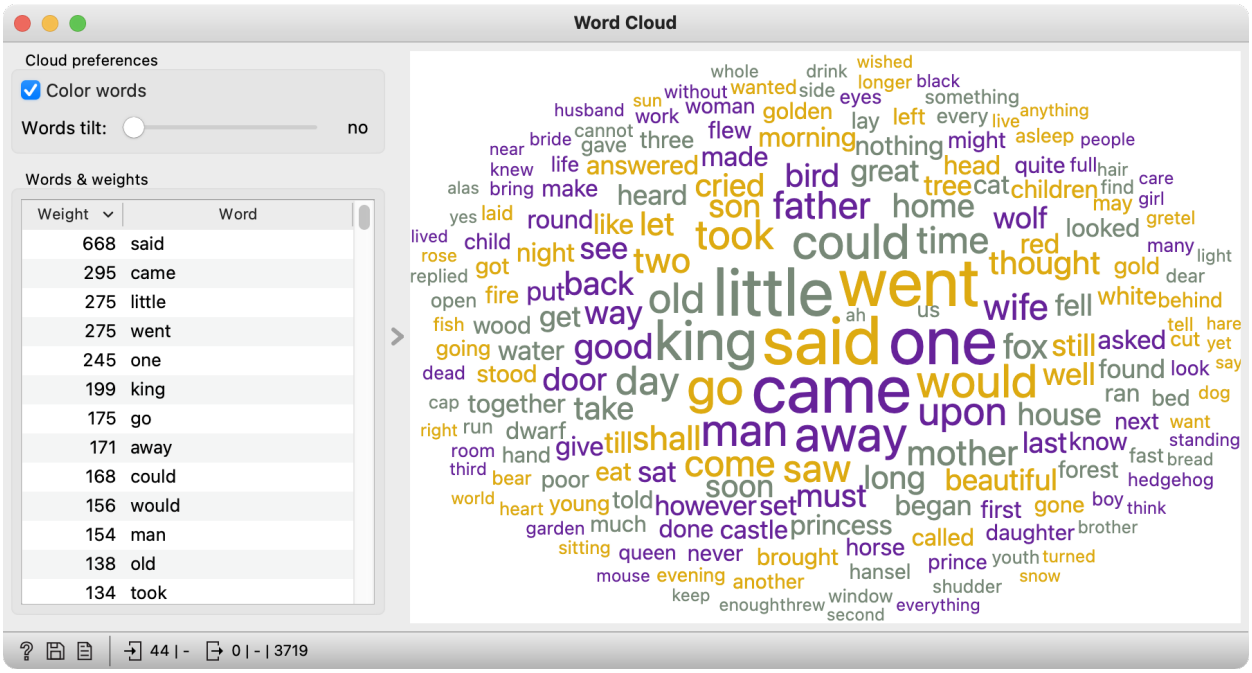
Preprocesiranje definira kaj je pomembno v podatikih. Je "Zdravnik" enako kot "zdravnik"? Naj upoštevamo besede kot so "in", "ali", "ko" ali naj jih izpustimo? Ali želimo upoštevati besedi "živel" in "živi" kot isto besedo? Preprocesiranje definira osnovne enote analize.

Token je osnovna enota naše analize. Lahko je beseda, besedna zveza, stavek... S predprocesiranjem definiramo osnovne enote za analizo.

V gradniku *Preprocess Text* smo vse besede pretvorili v male črke, vsako besedo smo obravnavali kot svojo *menoto (token)* in odstranili ločila, na koncu pa smo odstranili tudi nepomenske besede (npr. ‘in’, ‘da’, ‘čeprav’). Takšno predprocesiranje ustvari sledeče enote:

“To je vzorčni stavek.” → “vzorčni”, “stavek”

Rezultate predprocesiranja si lahko pogledamo v gradniku *Word Cloud*, kjer vidimo najpogostejše enote. S to vizualizacijo lahko identificiramo odvečne besede in nepravilnosti.



Rezultate predprocesiranja vidimo v gradniku Word Cloud. Dve najpogostejši besedi sta "would" in "could". Če se odločimo, da ti dve besedi nista primerni za našo analizo, ju moramo odstraniti. To lahko storimo s filtriranjem po meri.

Ker ne želimo upoštevati besed brez pomena, smo že odstranili nekaj nepomenskih besed. Ampak morda generično filtriranje ni dovolj za našo analizo.

V tem primeru vedno lahko naložimo seznam besed po meri. Odprite urejevalnik besedil in ustvarite seznam nepomenskih besed oz. besed, ki jih želite odstraniti. Vsako besedo zapišite v svojo vrstico in shranite dokument v obliki *.txt*.

Dober urejevalnik besedil je Sublime, lahko pa uporabite tudi WordPad ali Word.

Seznam besed naložite s klikom na ikono z mapo poleg opcije *Stopwords* v razdelku *Filtering*.

Filtriramo lahko tudi besede, ki so preredke ali prepogoste. Redke besede se pojavijo običajno le v nekaj dokumentih, medtem ko so prepogoste besede presplošne ali pa nimajo pomena (stopwords). Da bi ohranili le besede, ki zares predstavljajo naš korpus dokumentov, uporabimo filtriranje Document frequency (Pogostost v besedilu). Če nastavimo vrednosti na 0,1 and 0,9, bomo obdržali le tiste besede, ki se pojavijo v več kot 10 % in manj kot 90 % dokumentov.

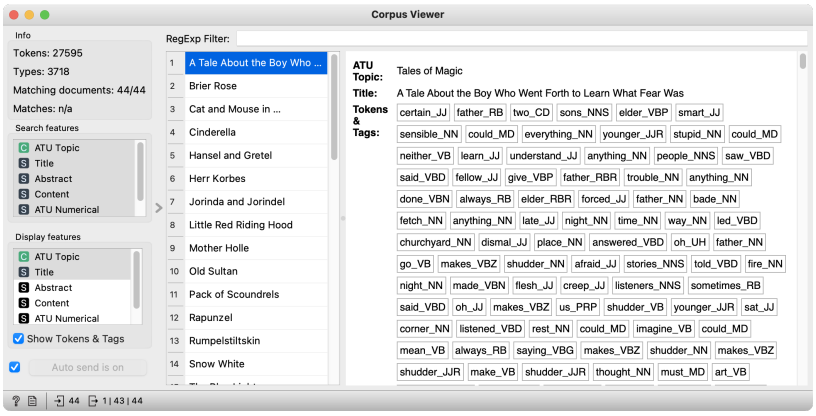
Predprocesiranje je ključ do uspešne analize besedil. Omenili smo le nekaj tehnik, sami pa lahko preizkusite še druge, na primer:

- *normalizacija (Normalization)* pretvori vse besede v korene oz. osnovne oblike (na primer sinovi v sin)
- *n-grami* so večje enote, na primer bigrami (par zaporednih besed) in trigrami (trojke besed)
- *oblikoskladenjsko označevanje* (POS tagging) označi vsako enoto s njeno oblikoskladenjsko vlogo (sinovi → samostalnik, množina, oznaka = NNS)

Pred kratkim smo za slovenščino dodali korenjenje z orodjem UDPipe

Za razlago POS oznak glejte: http://nl.ijs.si/imp/msd/html-sl/

Na sliki vidite gradnik Corpus Viewer, s katerim si lahko pogledamo naslove, besedila dokuemntov in enote na katere je preprocesiranje razbilo besedilo. V našem primeru imamo poleg enot prikazane tudi POS oznake.



Bibliography

Index

license, 2