

BIOLAB AND COLLABORATORS

UVOD V RUDARJENJE BES

BIOLAB

Copyright © 2021 Biolab and Collaborators

PUBLISHED BY BIOLAB

TUFTE-LATEX.GOOGLECODE.COM

Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

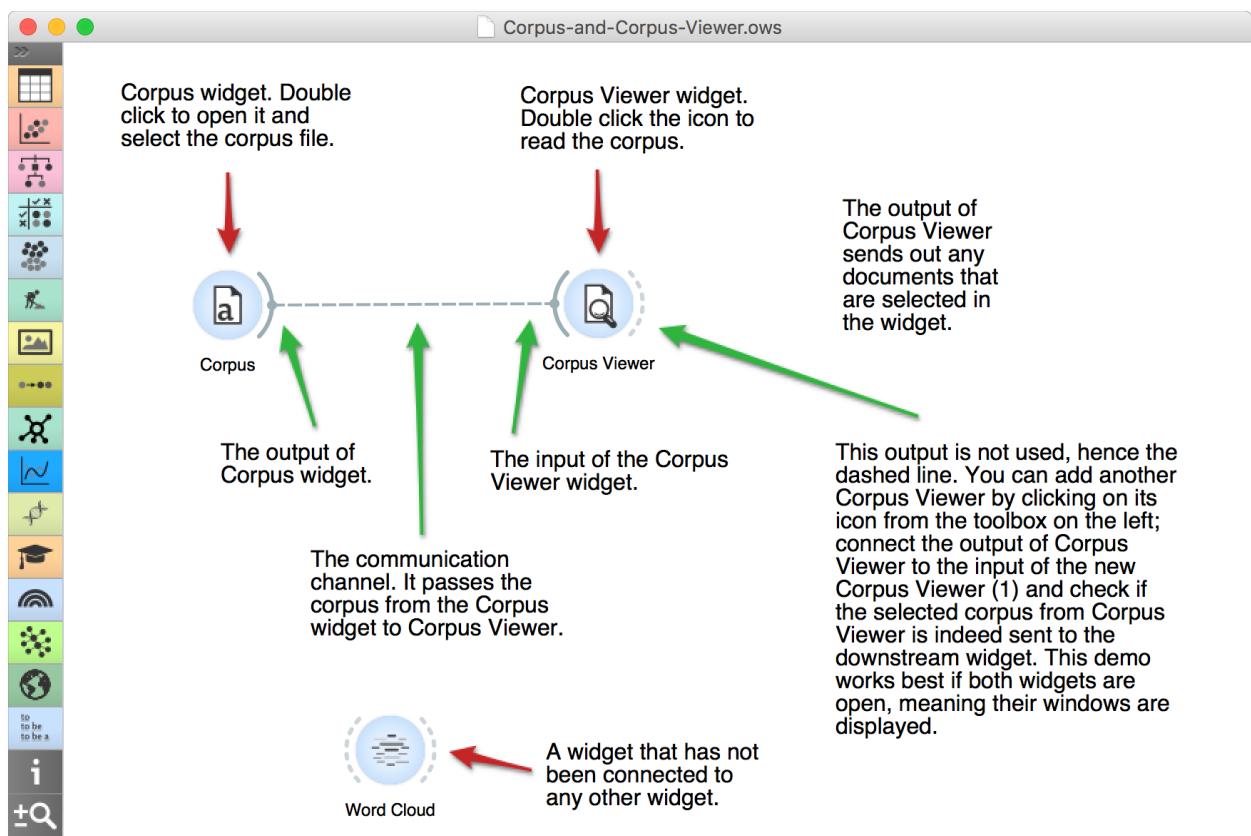
First printing, October 2021

Contents

<i>Delotoki v Orangeu</i>	5
<i>Priprava besedil</i>	8
<i>Kontekst</i>	11
<i>Vreča besed</i>	12
<i>Hierarhično razvrščanje v skupine</i>	13
<i>Hierarhično razvrščanje besedil</i>	15
<i>Bibliography</i>	19
<i>Index</i>	20

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.

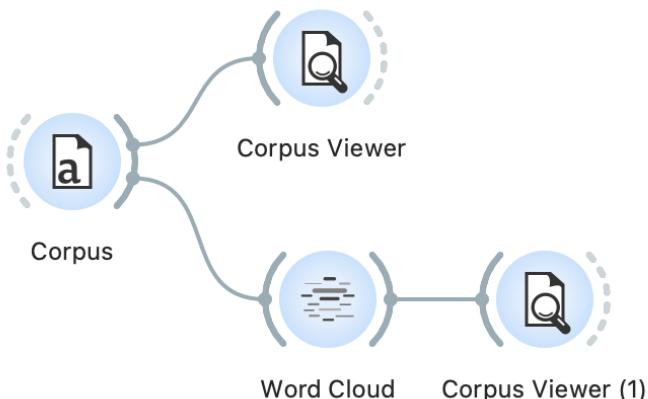


Delotoke sestavljamo tako, da polagamo gradnike na platno in jih povezujemo. Povezavo ustvarimo tako, da potegnemo črto od izhodnega v vhodni gradnik. Izhodi gradnika so na desni, vhodi pa na levi strani. V zgornjem delotoku gradnik *Corpus* pošilja podatke v gradnik *Corpus Viewer*.

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

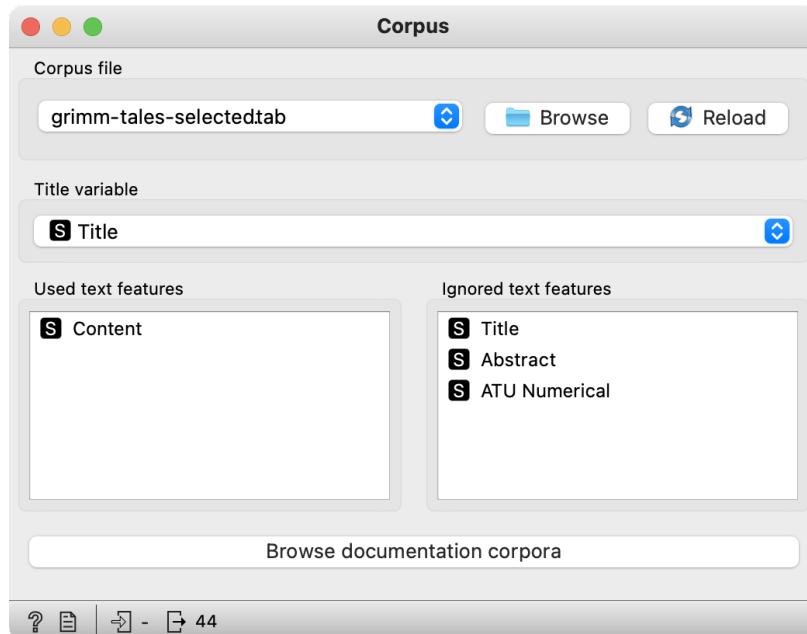
Pričnite z gradnjo delotoka, ki vsebuje gradnik *Corpus*, dva gradnika *Corpus Viewer* in gradnika *Word Cloud*:

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 najpogosteje 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 *Grimm-tales-selected.tab*, korpus z izbranimi Grimmovimi pravljicami.

Oranjevi 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.).



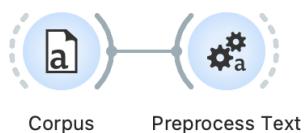
Odprite *Word Cloud*. *Word Cloud* (oblak besed) prikaže pogostost besed v dokumentih, kjer so pogostejše besede prikazane sorazmerno večje. Izberite besedo v oblaku in jo pošiljte v *Corpus Viewer (1)*. Sedaj lahko pregledate samo dokumente, ki vsebujejo izbrano besedo.



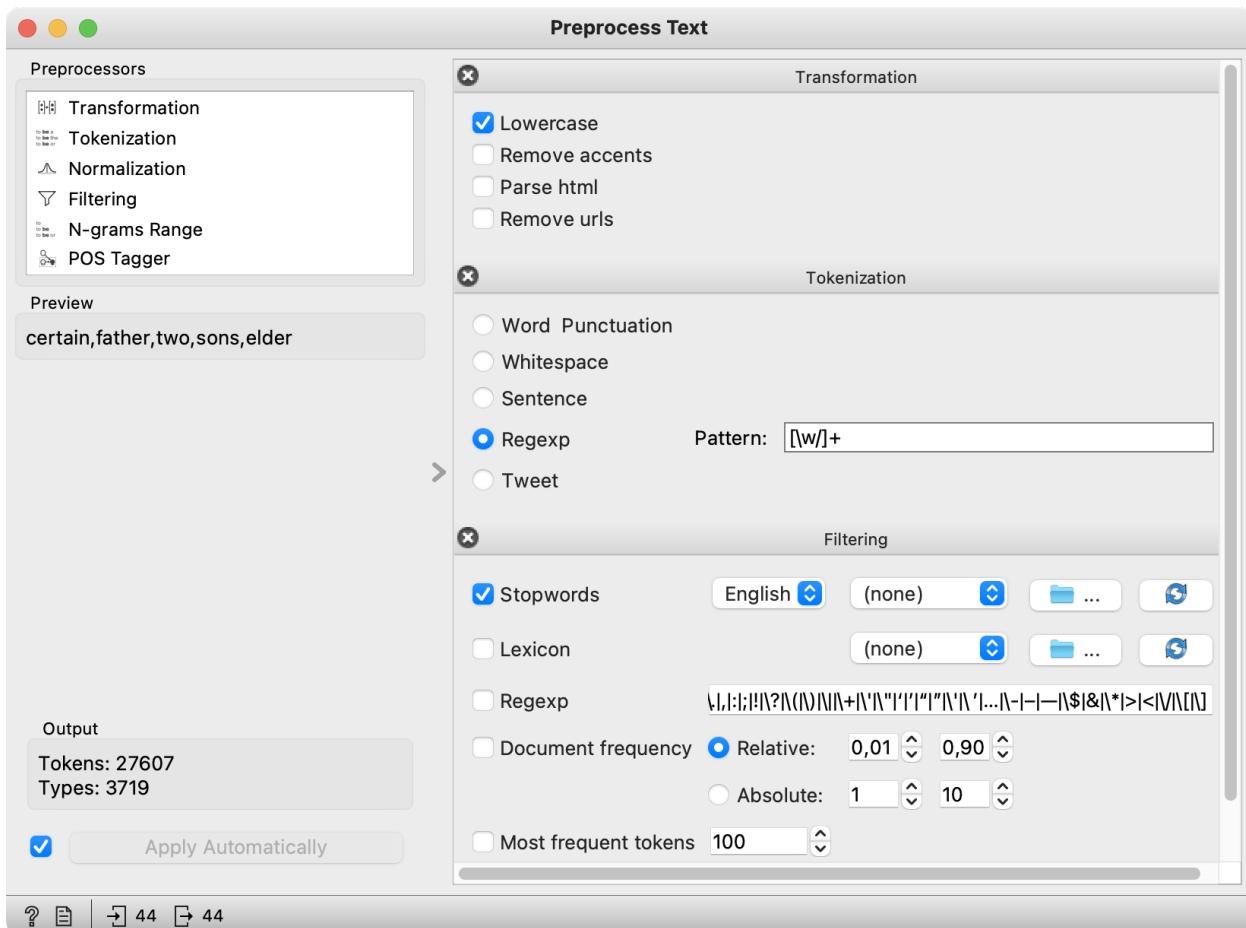
Počakajmo malo! Ta oblak besed je živa groza! Vidimo lahko celo kopico semantičnih smeti. Ali obstaja kakšen način, da to nekako uredimo?

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

Príprava 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.



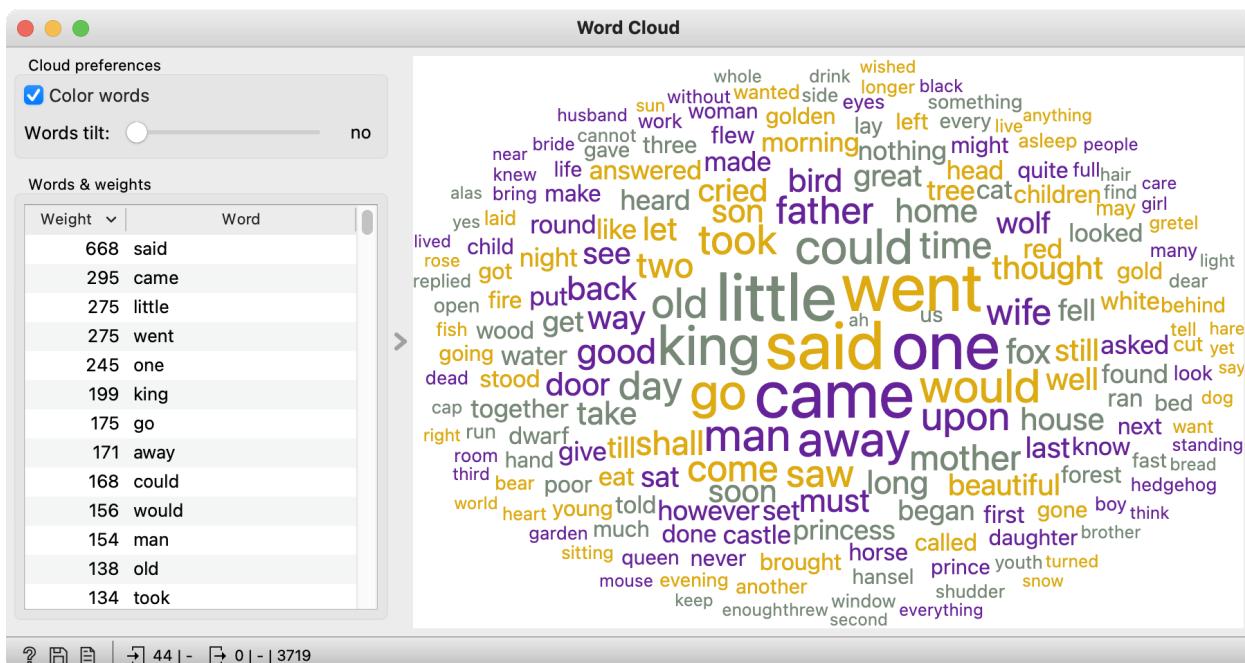
Preprocesiranje definira kaj je pomembno v podatkih. Je "Zdravnik" enako kot "zdravnik"? Naj upoštevamo besede kot so "in", "ali", "ko" ali naj jih izpustimo? Ali želimo upoštevati besedi "živel" in "živi" kot isto besedo? Preprocesiranje definira osnovne enote analize.

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"

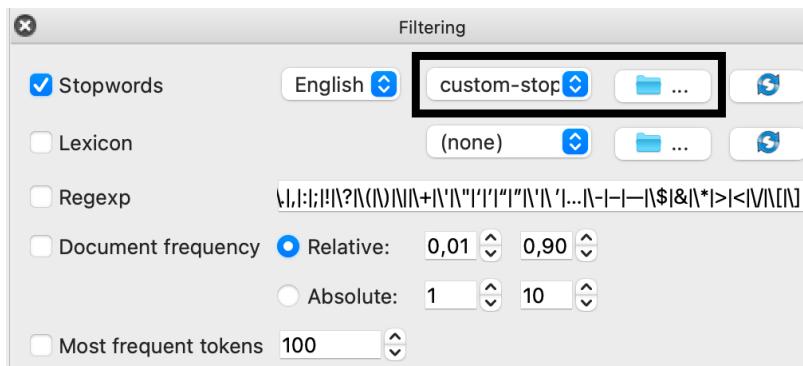
Rezultate predprocesiranja si lahko pogledamo v gradniku *Word Cloud*, kjer vidimo najpogosteje enote. S to vizualizacijo lahko identificiramo odvečne besede in nepravilnosti.



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*.

Rezultate predprocesiranja vidimo v gradniku Word Cloud. Dve najpogosteši 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.



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.

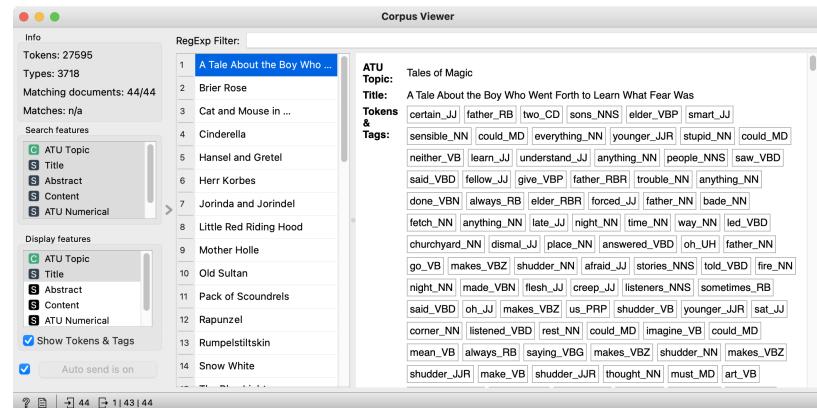
Predprocesiranje je ključ do uspešne analize besedil. Omenili smo le nekaj tehnik, sami pa lahko preizkusite še druge, na primer:

- *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 njenim oblikoskladenjskim 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 preprocesirane razbilo besedilo. V našem primeru imamo poleg enot prikazane tudi POS oznake.



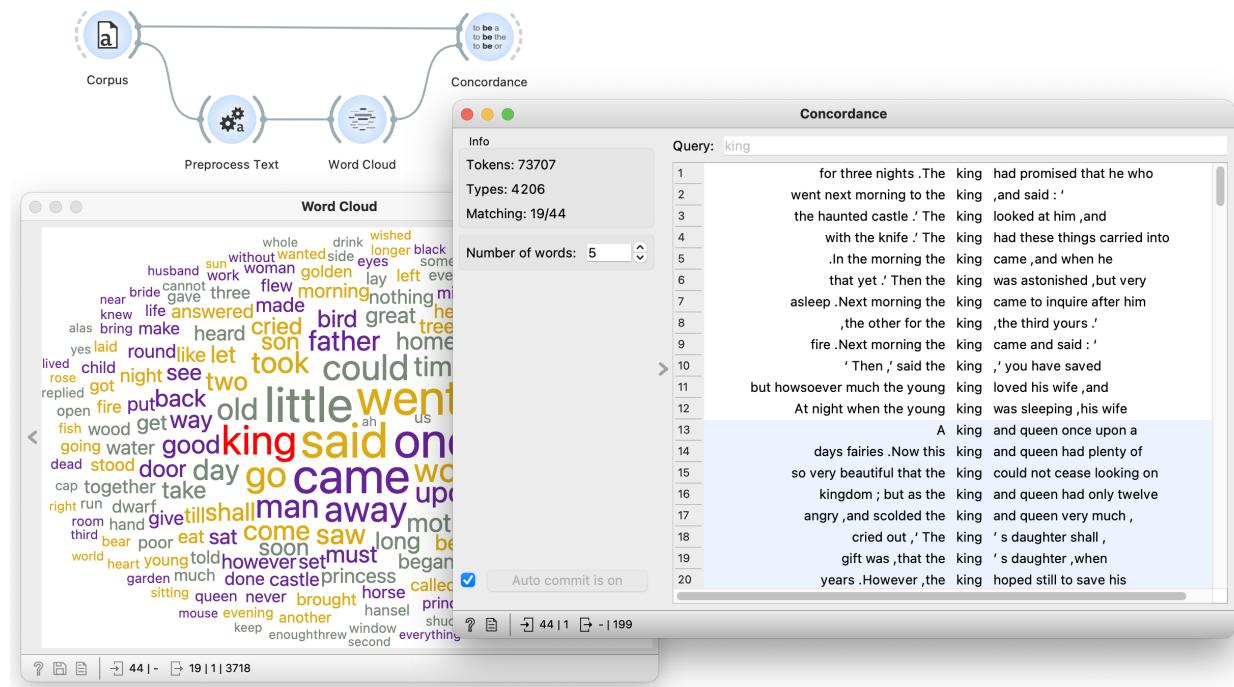
Kontekst

Sedaj smo pripravili korpus in čas je, da ga prikažemo. En način prikaza je oblak besed, ki ga že poznamo. Word Cloud nam prikaže pogostost besed. Pogostejsa kot je beseda, z večjimi črkami bo zapisana.

Še vedno pa ne vemo, kako se besede uporabljajo v besedilu. Na primer 'oh' je lahko maločrkovna verzija besede OH (kemijska spojina hidroksid), preprost vzklik 'Oh!' ali pa kratica za ameriško zvezno državo Ohio.

Da bi preverili kontekst posamezne besede, lahko uporabimo gradnik Concordance. Concordance na pokaže besedilo okrog izbrane besede. Pozvežite Concordance z gradnikom Corpus. Tako Concordance dobi vhodno besedilo. Besedo lahko poiščemo z iskalnikom na vrhu gradnika ali pa jo izberemo v gradniku Word Cloud.

Vizualizacije v Orangeu so narejene tako, da podpirajo izbor podmnožic. Odkrivanje zanimivih podmnožic in raziskovanje njihovih podobnosti je ključni del odkrivanja znanj iz podatkov.



V tem primeru smo izbrali besedo 'king' v oblaku besed in preverili njen kontekst v gradniku Concordance.

Dokumente, ki vsebujejo izbrano besedo, si pogledamo tako, da izberemo dokumente v gradniku Concordance in jih pošljemo v Corpus Viewer za podrobnejšo analizo.

Vreča besed

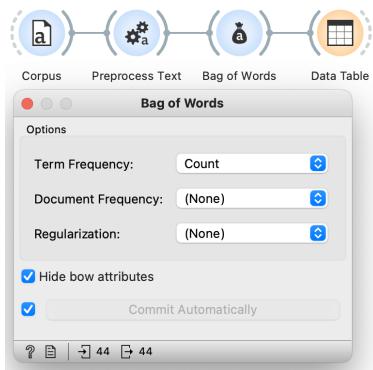
Sedaj imamo predprocesirano besedilo, s pravimi enotami, ampak še vedno pa ne moremo odkriti nikakršnih vzorcev v besedilu. Za to potrebujemo številke in preprost način, kako spremenimo besedila v številske vektorje je... da preštejemo besede v vsakem dokumentu!

	this	is	an	example	another	apple
"This is an example"	1	1	1	1	0	0
"Another example"	0	0	0	1	1	0
"This is another apple"	1	1	0	0	1	1

Gradnik Bag of Words ustvari tabelo z besedami v stolpcih in dokumenti v vrsticah. Vrednosti so pojavitve besed v vsakem dokumentu.

Besede lahko preprosto preštejemo (TF ali term frequency) ali pa besede utežimo glede na to, kako pogosto se pojavijo v dokumentih (IDF ali inverse document frequency). S TF-IDF bodo pogoste besede imele nizko vrednost, saj se pojavi velikokrat pri velikem deležu dokumentov, medtem ko bodo imele pomembne besede visoko vrednost, saj se pojavi pogosto v majhnem deležu dokumentov.

Podatke pošilje v gradnik Bag of Words in od tam naprej v Data Table. Vidimo nov stolpec, ki vsebuje pojavitve besed za vsak dokument. Sedaj imamo številke in končno lahko pričnemo z analizo!

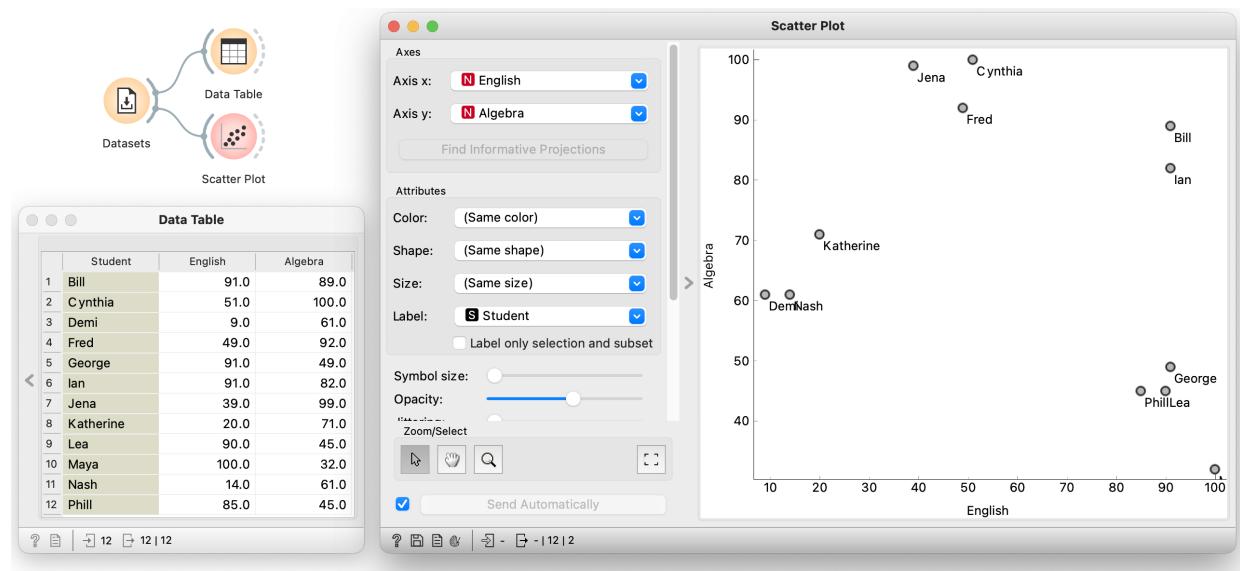


Data Table		
bow-feature hidden include skip-normalization title	ATU Topic	Title
1	Tales of Magic	A Tale About...
2	Tales of Magic	Brier Rose
3	Animal Tales	Cat and Mou...
4	Tales of Magic	Cinderella
5	Tales of Magic	Hansel and ...
6	Animal Tales	Herr Korbes
7	Tales of Magic	Jorinda and ...
8	Tales of Magic	Little Red Ri...
9	Tales of Magic	Mother Holle
10	Animal Tales	Old Sultan
11	Animal Tales	Pack of Sco...
12	Tales of Magic	Rapunzel
13	Tales of Magic	Rumpelstilts...
14	Tales of Magic	Snow White
15	Tales of Magic	The Blue Light
16	Animal Tales	The Bremen ...
17	Animal Tales	The Crumbs ...
18	Animal Tales	The Dog and...
19	Tales of Magic	The Elves an...
20	Tales of Magic	The Fisher...

Hierarhično razvrščanje v skupine

Ena od nalog rudarjenja besedil je iskanje zanimivih skupin dokumentov. Torej radi bi odkrili dokumente, ki so si podobni mes sabo.

Poglejmo si preproste podatke z dvema stolpcema (glejte opombo) in jih prikažimo v gradniku Scatter Plot. Koliko skupin imamo? Kaj predstavlja različne skupine? Kateri primeri sodijo v posamezno skupino?



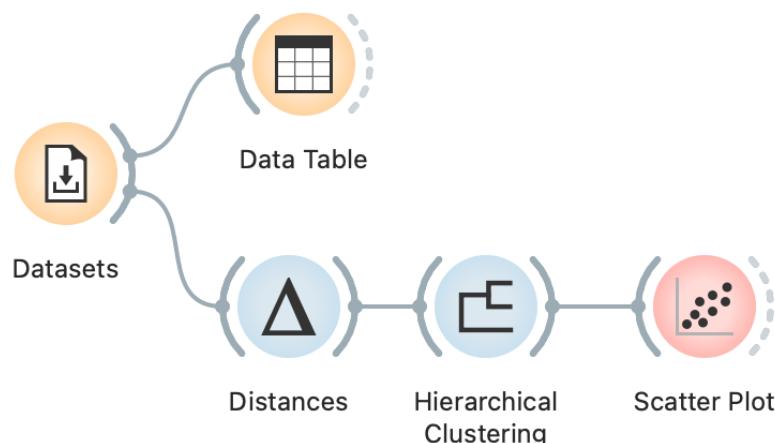
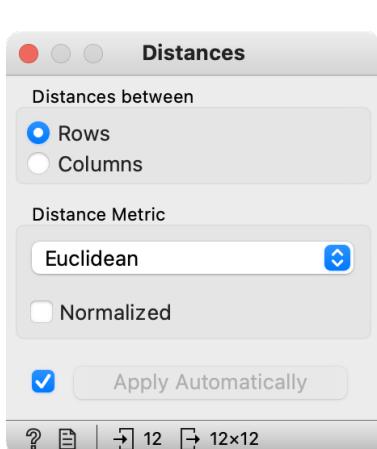
Kaj sploh pomeni "podobno"? Študenti so opisani s številskimi spremenljivkami, torej s ocenami pri predmetu. Ena od mer podobnosti je kosinusna razdalja. Pogostost besed iz vreče besed je predstavljena z vektorji, ki kažejo vsak v svojo smer glede na vsebino posameznega dokumenta. Kosinusna razdalja je kot med temi vektorji.

Sedaj definirajmo še postopek za razvrščanje v skupine. Recimo, da začnemo z vsakim dokumentom v svoji skupini, nato pa v vsakem koraku združimo skupini, ki sta si najbolj podobni. Razdaljo med skupinami izračunamo kot povprečje razdalj med posameznimi elementi skupine. Tak postopek imenujemo hierarhično razvrščanje v skupine.

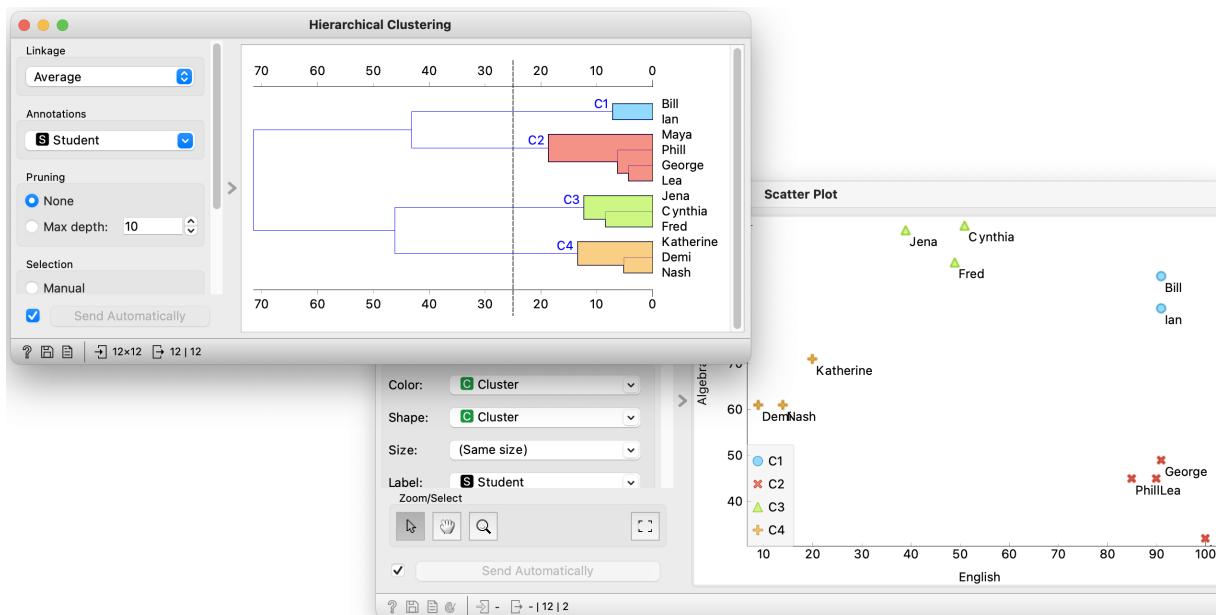
Razvrščanje v skupine bomo predstavili s preprostimi podatki o študentih in njihovih ocenah pri angleščini in matematiki. Podatki so dostopni v Datasets widgetu.

Načinov merjenja razdalj med skupinami je več. Način, ki smo ga opisali, se imenuje povprečna razdalja (average linkage). Lahko bi računali tudi razdaljo med najbližjima točkama v skupini (single linkage) ali pa med točkama, ki sta si najbolj oddaljeni (complete linkage).

Rezultate razvrščanja v skupine na primeru naših študentov si lahko pogledamo v sledečem delotoku:



Naložite podatke z gradnikom File, izračunajte razdalje z gradnikom Distances, uporabite Hierarchical Clustering in si poglejte rezultate v gradniku Scatter Plot. Gradnik Hierarchical Clustering omogoča, da hierarhijo skupin odrežemo pri določeni meri podobnosti in tako definiramo skupine.



Hierarhično razvrščanje besedil

Vrnimo se h Grimmovim pravljicam in ustvarimo naslednji delotok:

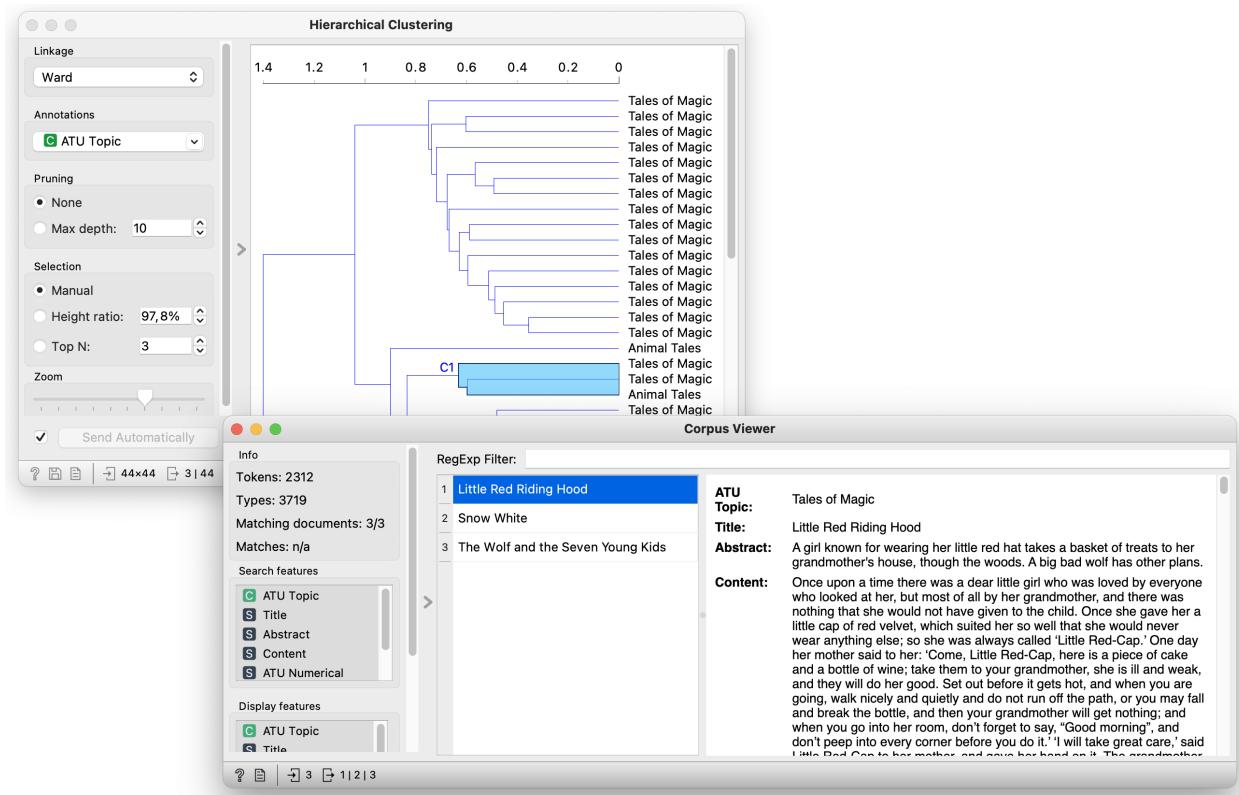


Gradnik Hierarchical Clustering prikaže gruče v obliki dendrograma. Hierarchical Clustering povežite z gradnikom Corpus Viewer in odprite oba gradnika. Izberite gručo v dendrogramu in v gradniku Corpus Viewer poglejte, kateri dokumenti pripadajo izbrani skupini.

Raziščite različne gruče. Zakaj so nekatere magične pravljice (Tales of Magic) pomešane z živalskimi pravljicami (Animal Tales)? Kaj imajo skupnega?

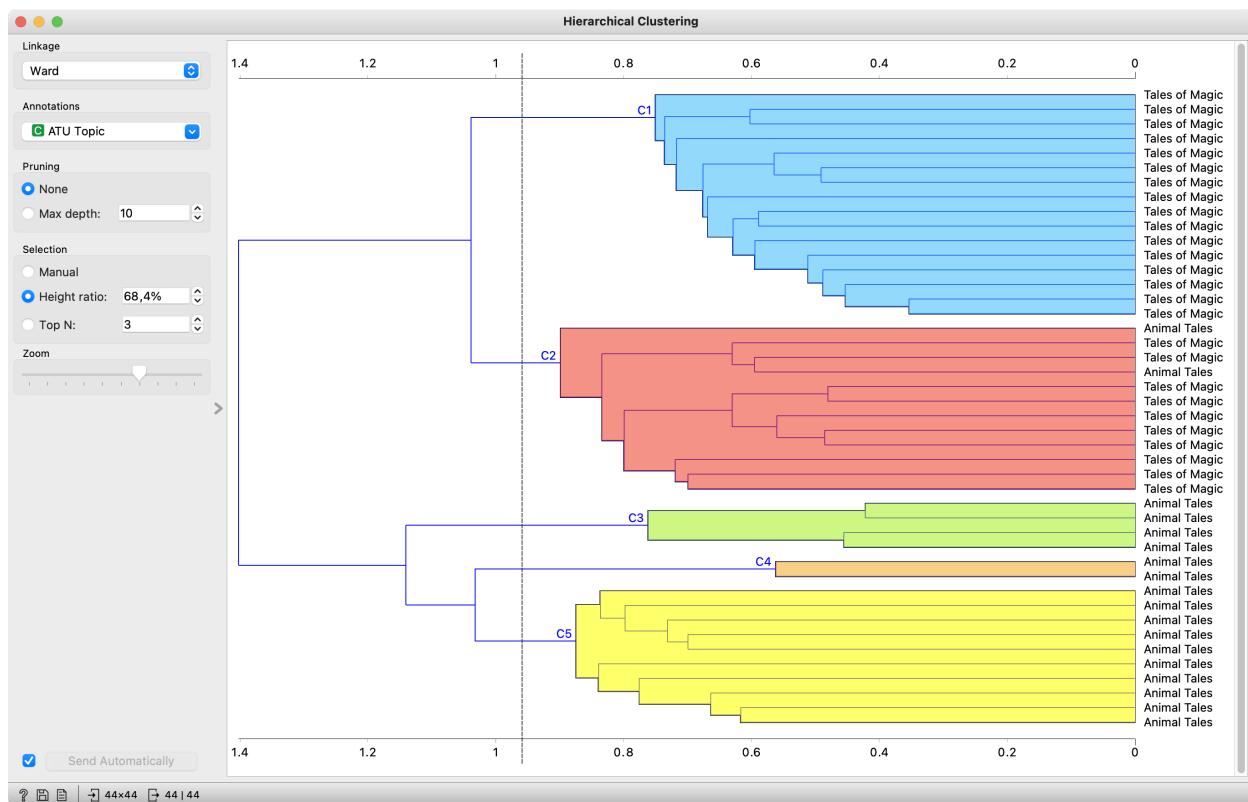
Enak delotok lahko preizkusite tudi na drugih podatkih, na primer na bookexcerpt.tab, ki vsebuje izvlečke knjig za odrasle in otroke.

Beseda dendrogram je sestavljena iz grških besede dendro "drevo" in gramma "risba" in pomeni hierarhično vizualizacijo v obliki drevesa.

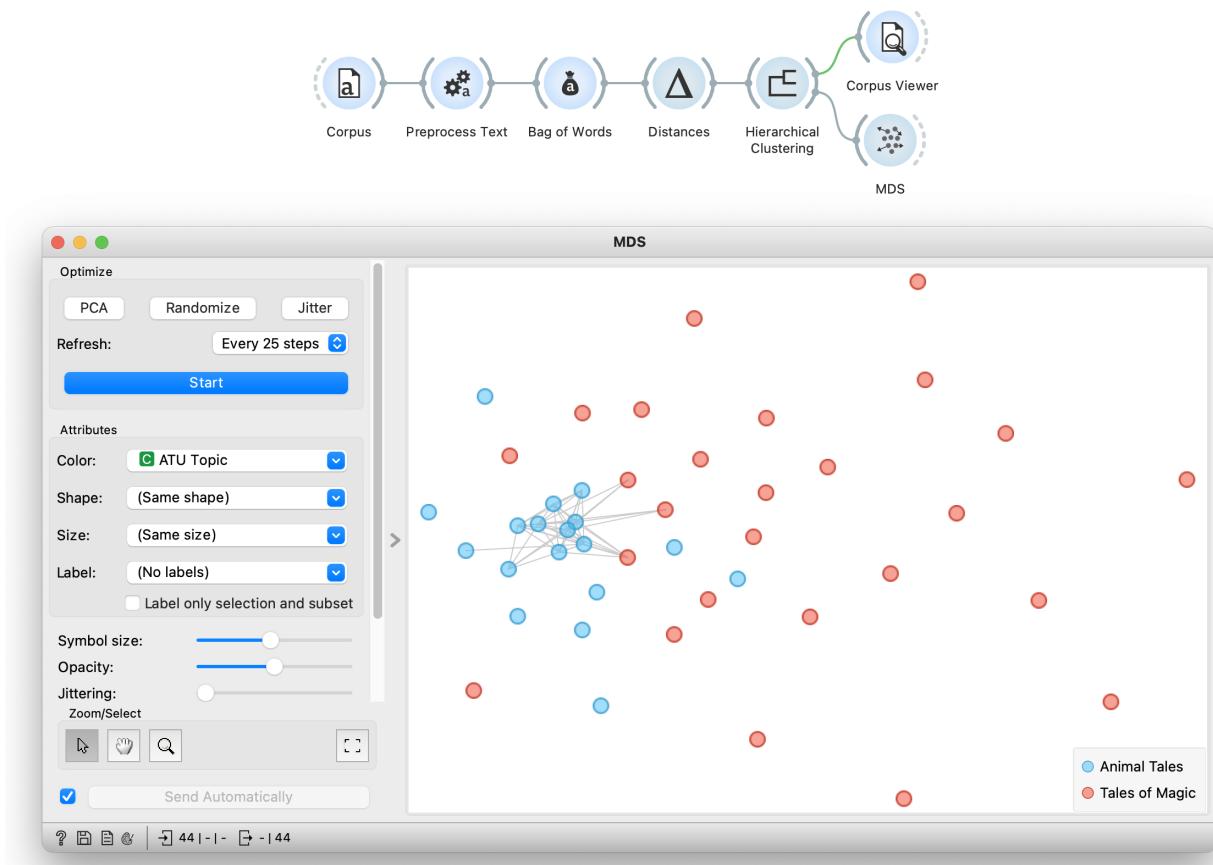


Hierarhično razvrščanje zgradi hierarhijo dokumentov, mi pa se moramo odločiti, kje zamejimo podobnost znotraj skupine. Mejo podobnosti nastavimo tako, da na ravnili zgoraj povlečemo črto desno ali levo in s tem zamejimo skupine.

Odločili smo se za pet skupin, saj po tem razdalja med skupinami kar precej naraste. Primerjajte pet skupin s štirimi, šestimi ali sedmimi. Gruče, ki jih odkrijemo, hkrati sovpadajo z oznako tipa Aarne-Thompson (ATU Topic).



Kako blizu pa so si v resnici živalske pravljice iz tretje in te iz četrte skupine? Ali ne bi bilo bolj zanimivo pogledati dokumente v ravnini, kjer bi se podobni dokumenti nahajali skupaj, različni pa narazen? Taka vizualizacija se imenuje večrazsežnostno lestvičenje oziroma Multidimensional Scaling (MDS).



Magične pravljice tvorijo eno skupino, živalske pa drugo - tako kot smo pričakovali. Zanimivo, magične pravljice so si med sabo bolj podobne kot živalske (bolj so povezne). Raziščite podobne pravljice tako, da jih izberete v vizualizaciji in jih pogledate v gradniku Corpus Viewer.

Bibliography

Index

license, [2](#)