HỆ HỖ TRỢ QUYẾT ĐỊNH

Bài 10(b): Text Mining

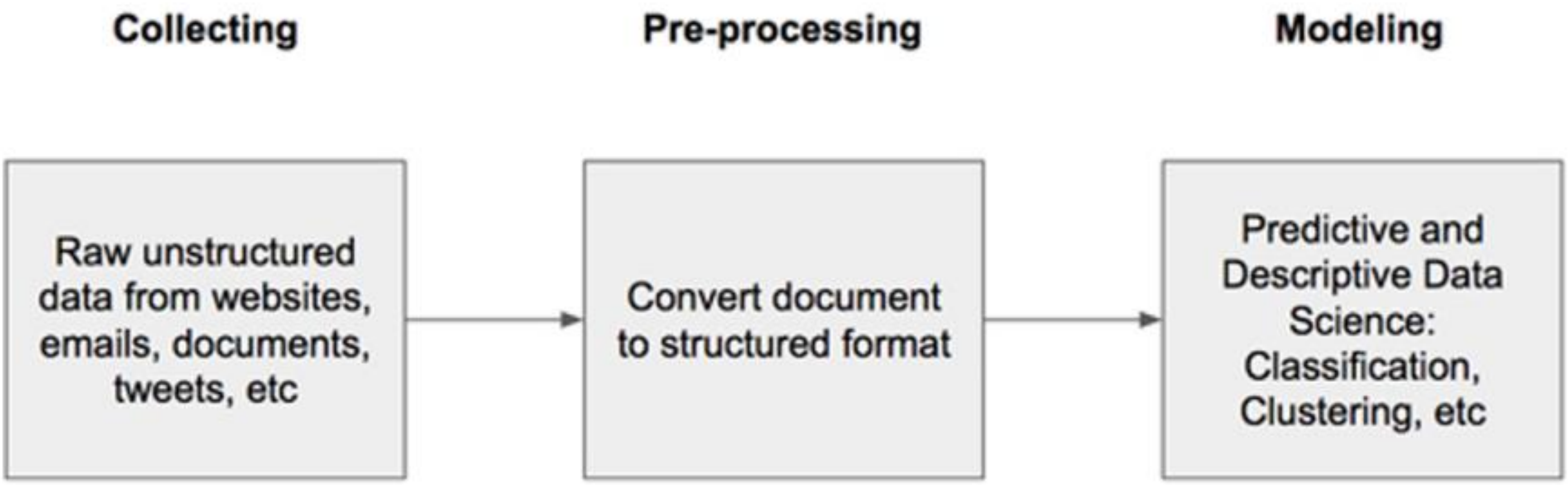
Contents

https://ucilnica.fri.uni-lj.si/pluginfile.php/164808/mod_resource/content/2/Text%20Mining.pdf

Text data

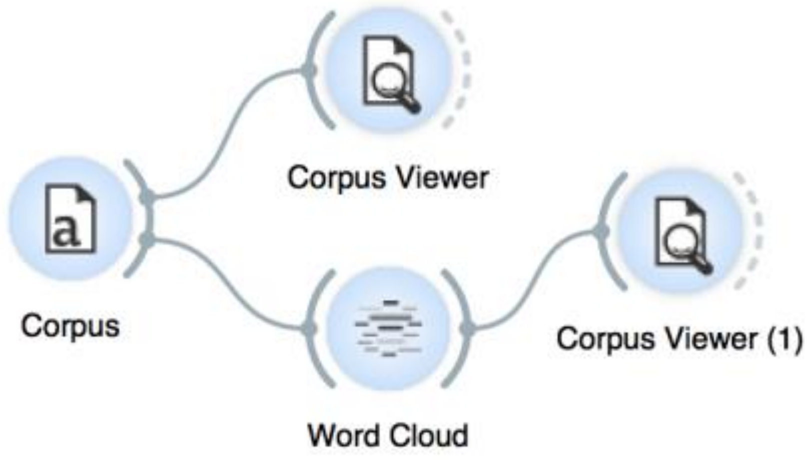
- Unstructured data (including text, audio, images, videos, etc.) is the new frontier of data science
- If all the data in the world was equivalent to the water on earth, then textual data is like the ocean, making up a majority of the volume
- Text analytics is driven by the need to process natural human language, but unlike numeric or categorical data, natural language does not exist in a structured format consisting of rows (of examples) and columns (of attributes)
- Text mining is, therefore, the domain of unstructured data science

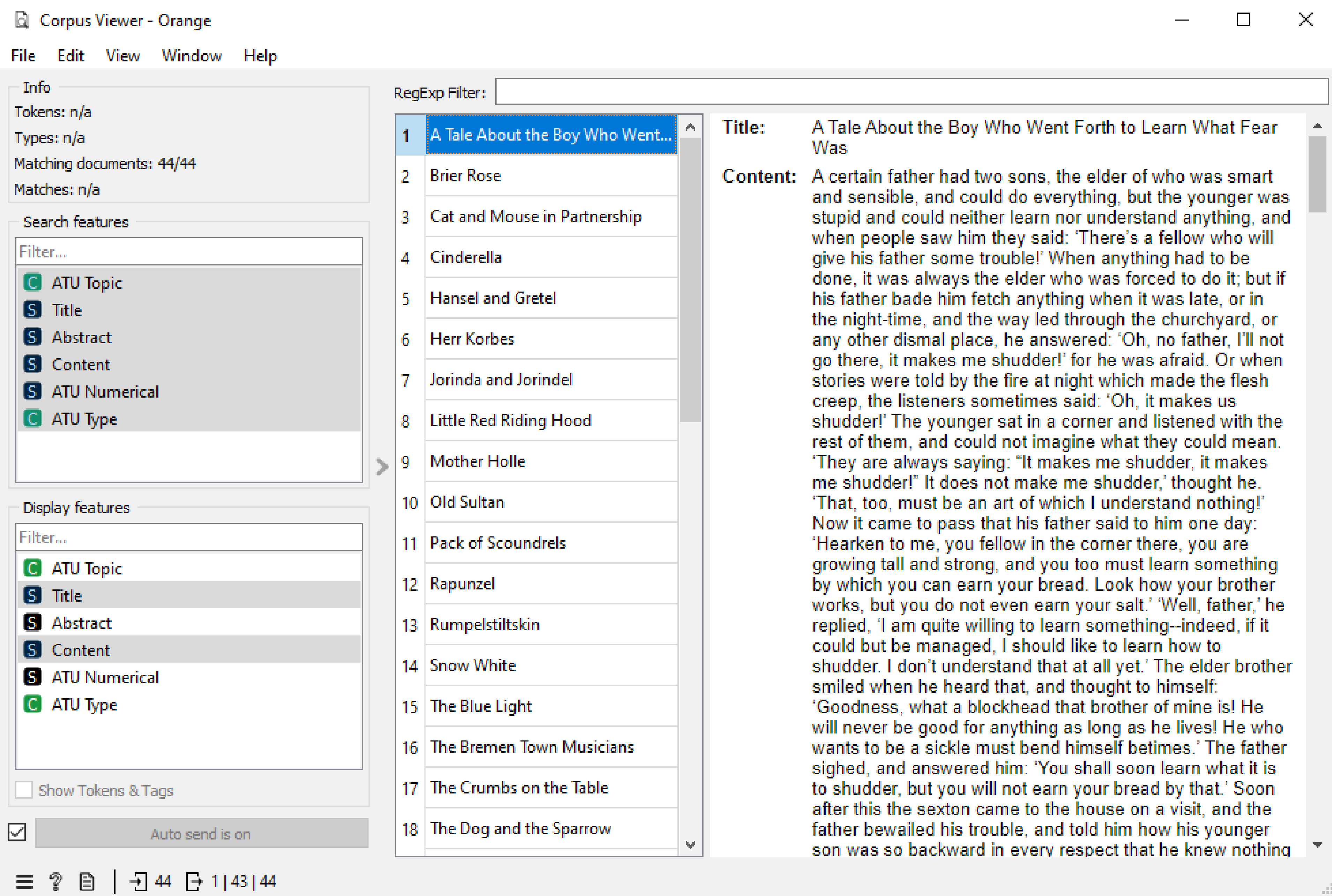
High-level process for text mining

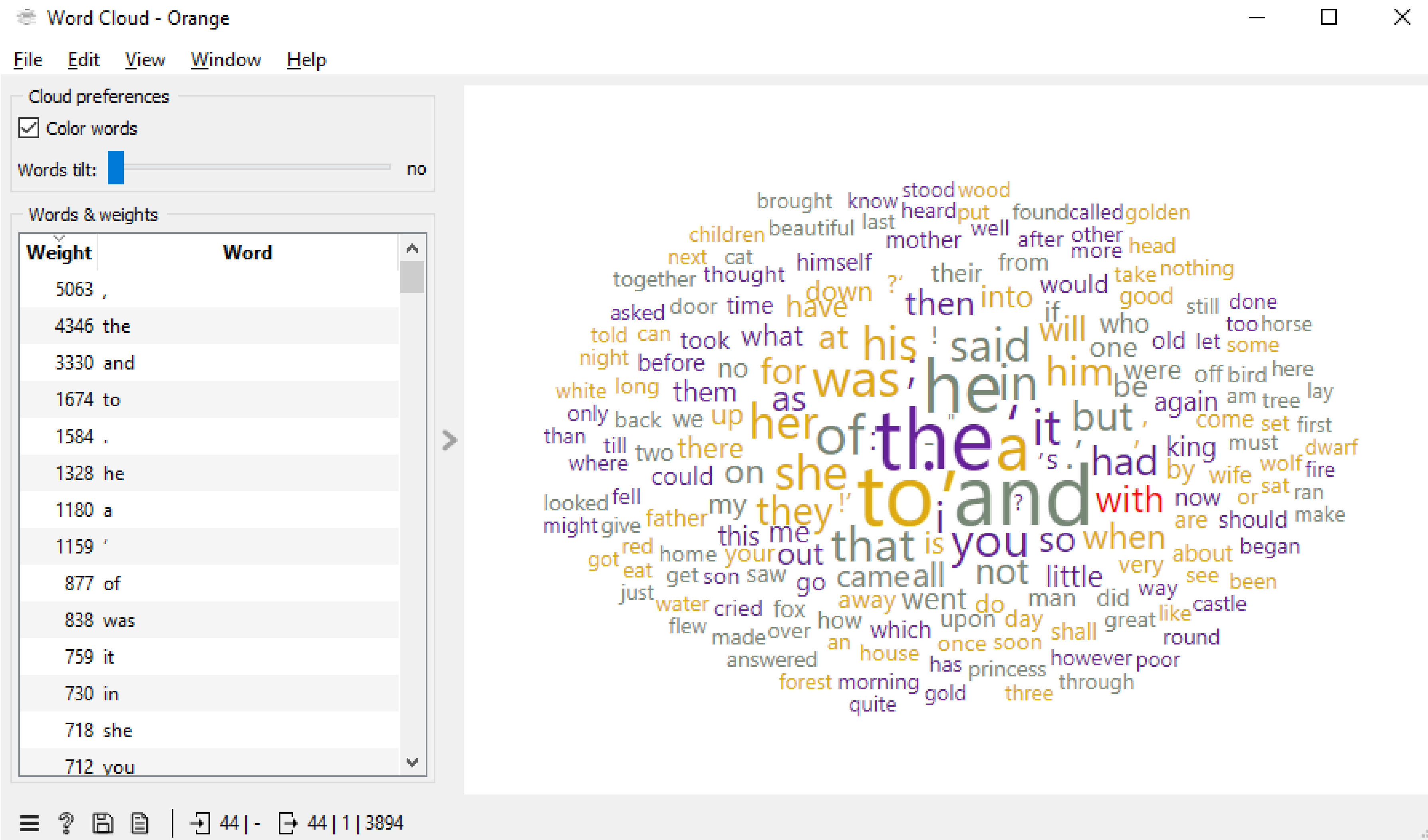


Corpus

- A collection of documents
- A document: a collection of sentences/words/characters
- Example: *Grimm-talesselected.tab*



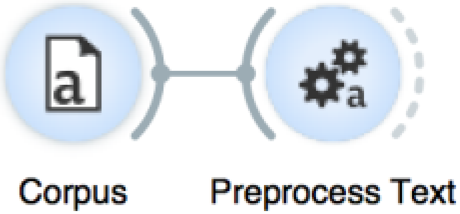

Corpus Viewer - Orange
File Edit View Window Help
Info
Tokens: n/a
Types: n/a
Matching documents: 44/44
Matches: n/a
Search features
Filter...
ATU Topic
Title
Abstract
Content
ATU Numerical
ATU Type
Display features
Filter...
ATU Topic
Title
Abstract
Content
ATU Numerical
ATU Type
Show Tokens & Tags
Auto send is on
RegExp Filter:
1 A Tale About the Boy Who Went...
2 Brier Rose
3 Cat and Mouse in Partnership
4 Cinderella
5 Hansel and Gretel
6 Herr Korbes
7 Jorinda and Jorindel
8 Little Red Riding Hood
9 Mother Holle
10 Old Sultan
11 Pack of Scoundrels
12 Rapunzel
13 Rumpelstiltskin
14 Snow White
15 The Blue Light
16 The Bremen Town Musicians
17 The Crumbs on the Table
18 The Dog and the Sparrow
Title: A Tale About the Boy Who Went Forth to Learn What Fear Was
Content: A certain father had two sons, the elder of who was smart and sensible, and could do everything, but the younger was stupid and could neither learn nor understand anything, and when people saw him they said: 'There's a fellow who will give his father some trouble!' When anything had to be done, it was always the elder who was forced to do it; but if his father bade him fetch anything when it was late, or in the night-time, and the way led through the churchyard, or any other dismal place, he answered: 'Oh, no father, I'll not go there, it makes me shudder!' for he was afraid. Or when stories were told by the fire at night which made the flesh creep, the listeners sometimes said: 'Oh, it makes us shudder!' The younger sat in a corner and listened with the rest of them, and could not imagine what they could mean. 'They are always saying: "It makes me shudder, it makes me shudder!" It does not make me shudder,' thought he. 'That, too, must be an art of which I understand nothing!' Now it came to pass that his father said to him one day: 'Hearken to me, you fellow in the corner there, you are growing tall and strong, and you too must learn something by which you can earn your bread. Look how your brother works, but you do not even earn your salt.' 'Well, father,' he replied, 'I am quite willing to learn something--indeed, if it could but be managed, I should like to learn how to shudder. I don't understand that at all yet.' The elder brother smiled when he heard that, and thought to himself: 'Goodness, what a blockhead that brother of mine is! He will never be good for anything as long as he lives! He who wants to be a sickle must bend himself betimes.' The father sighed, and answered him: 'You shall soon learn what it is to shudder, but you will not earn your bread by that.' Soon after this the sexton came to the house on a visit, and the father bewailed his trouble, and told him how his younger son was so backward in every respect that he knew nothing
44
1 | 43 | 44

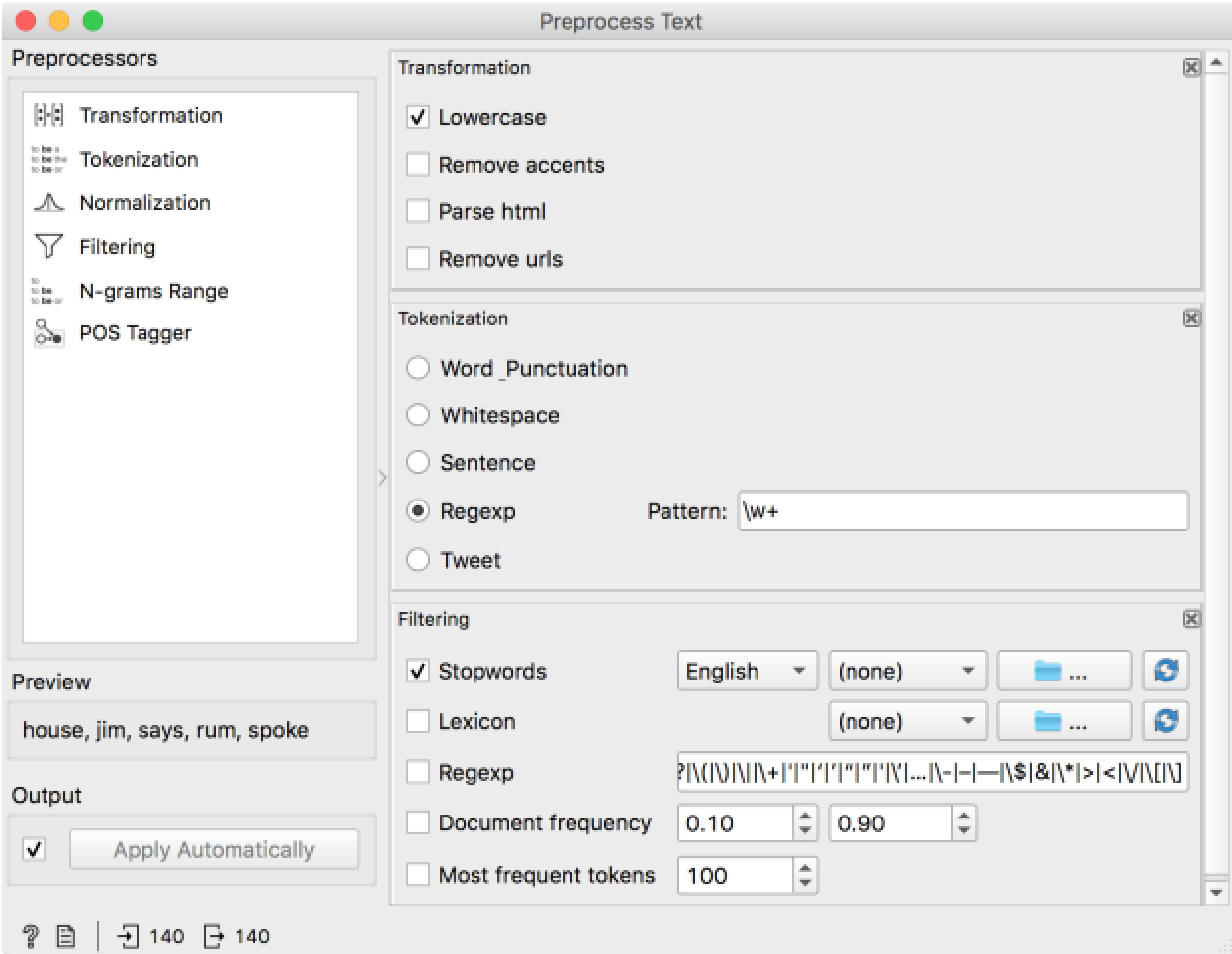


We need to remove all the bits that carry no information, namely **punctuation** and **stopwords**.

Preprocessing Text

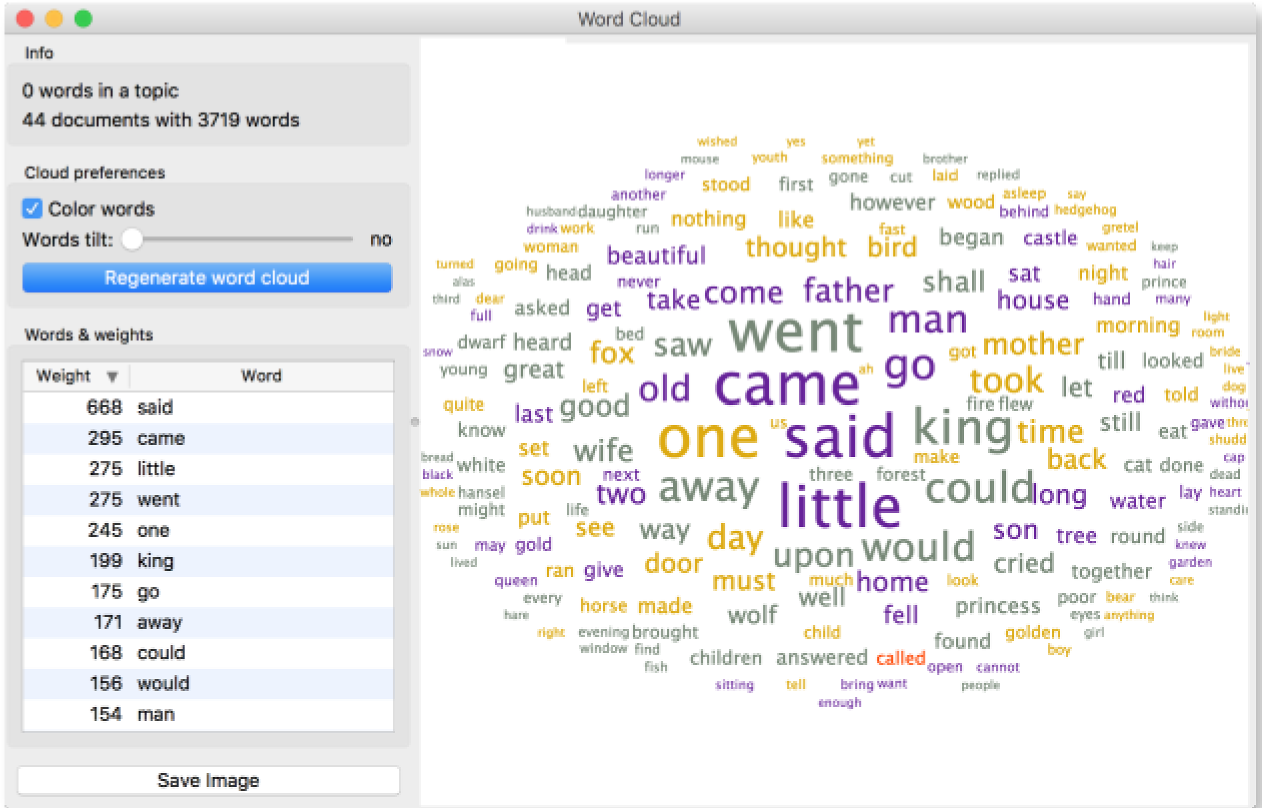
- Preprocessing is key to defining what is important in the data. Is “Doctor” the same as “doctor”?
- Should we consider words such as “and”, “the”, “when” or omit them?
- Do we wish to treat “said” and “say” as the same word?
- Preprocessing defines the core units of the analysis.
- **Token** is a basic unit of the analysis. It can be a word, a bigram, a sentence… With preprocessing we define our tokens for the analysis.



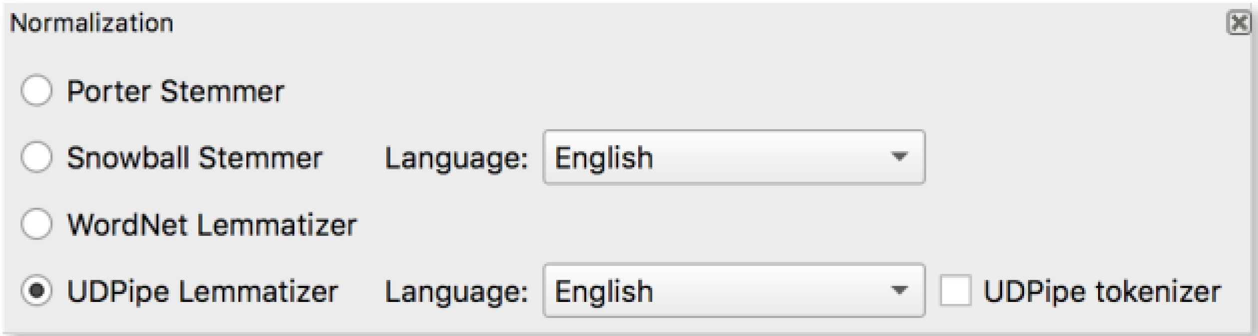

Preprocess Text
Preprocessors
Transformation
Tokenization
Normalization
Filtering
N-grams Range
POS Tagger
Preview
house, jim, says, rum, spoke
Output
Apply Automatically
Transformation
Lowercase
Remove accents
Parse html
Remove urls
Tokenization
Word_Punctuation
Whitespace
Sentence
Regexp
Pattern: \w+
Tweet
Filtering
Stopwords
English
(none)
Lexicon
(none)
Regexp
Document frequency
0.10
0.90
Most frequent tokens
100
140
140

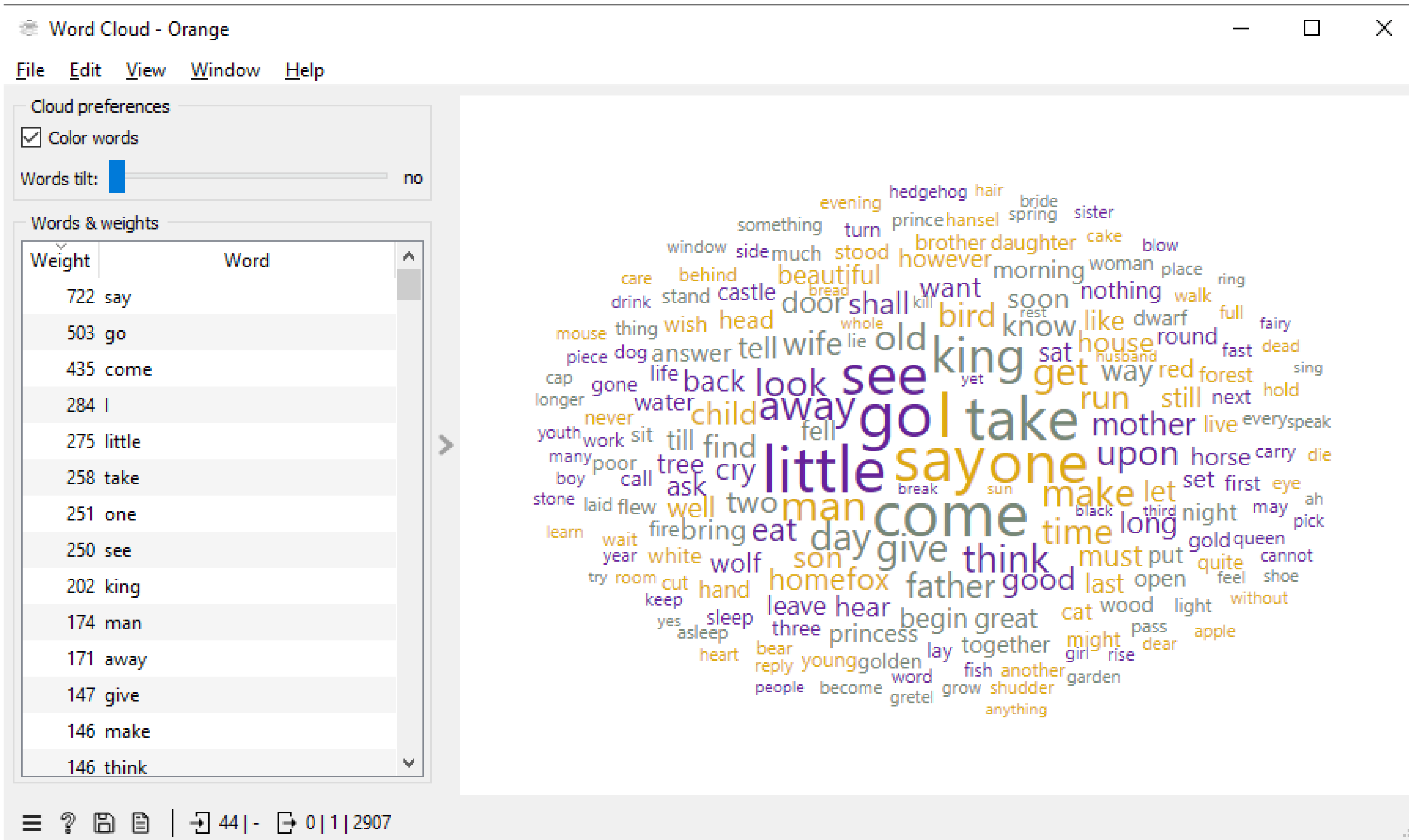
Preprocessing terminology

- **Stopwords**: articles, conjunctions, pronouns, prepositions, and other similar terms that need to be filtered before additional analysis. The process of removing these words is called Stop word filtering
- **Term filtering**: process to remove some normal terms in specific domains
- **Stemming**: process to convert words into their stem.
- **n-gram**: group n words into a term
- **POS tagger**: tagging tags each token with a corresponding part-of-speech tag (sons → noun, plural, tag = NNS)



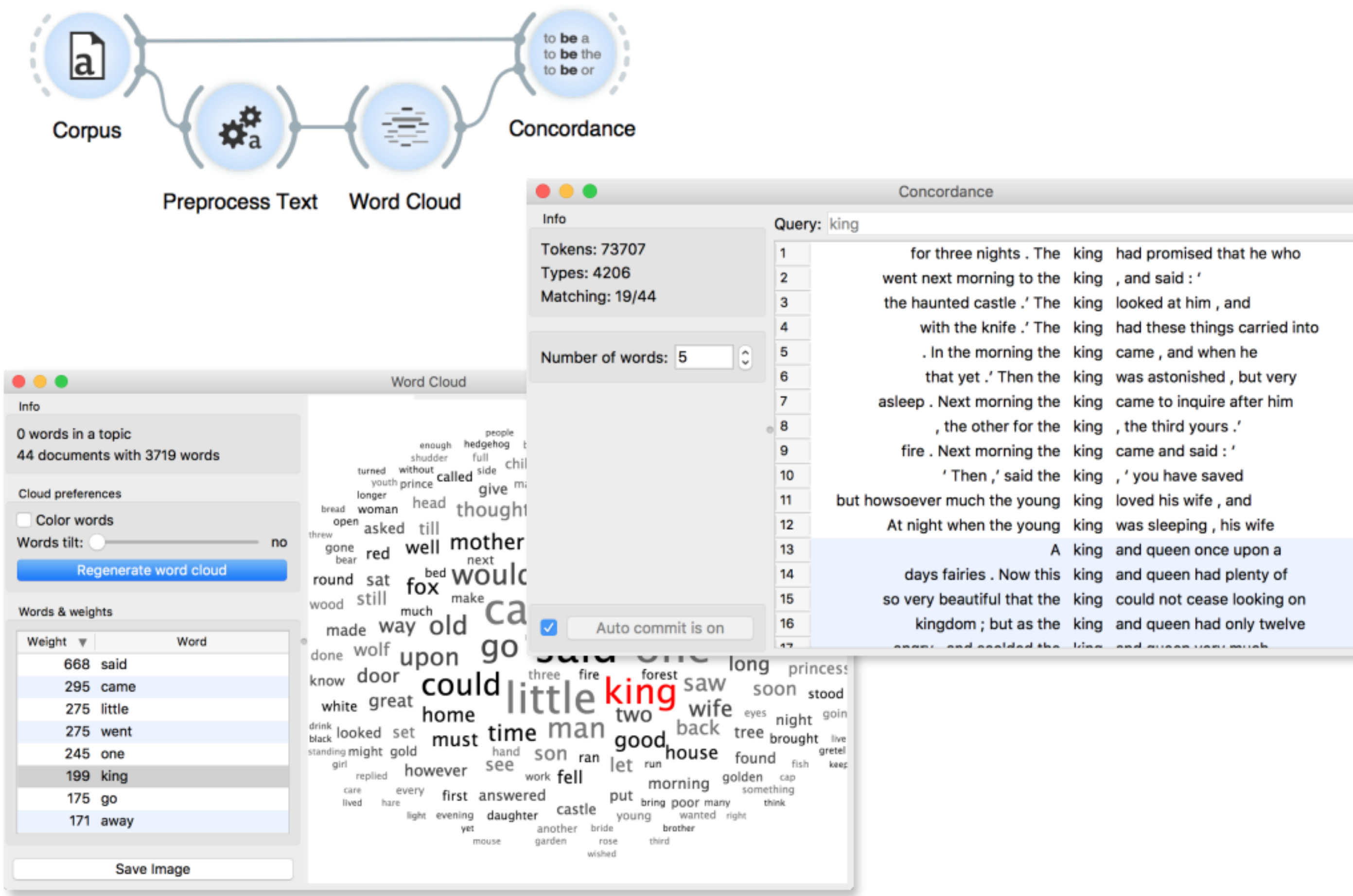
Two of the most frequent words are “would” and “could”.
If we decide these two words are not important for our analysis, it would be good to omit them.
We can do this with custom filtering.


Normalization
Porter Stemmer
Snowball Stemmer
Language: English
WordNet Lemmatizer
UDPipe Lemmatizer
Language: English
UDPipe tokenizer


Word Cloud - Orange
File Edit View Window Help
Cloud preferences
Color words
Words tilt: no
Words & weights
Weight Word
722 say
503 go
435 come
284 I
275 little
258 take
251 one
250 see
202 king
174 man
171 away
147 give
146 make
146 think
44 | -
0 | 1 | 2907

Context

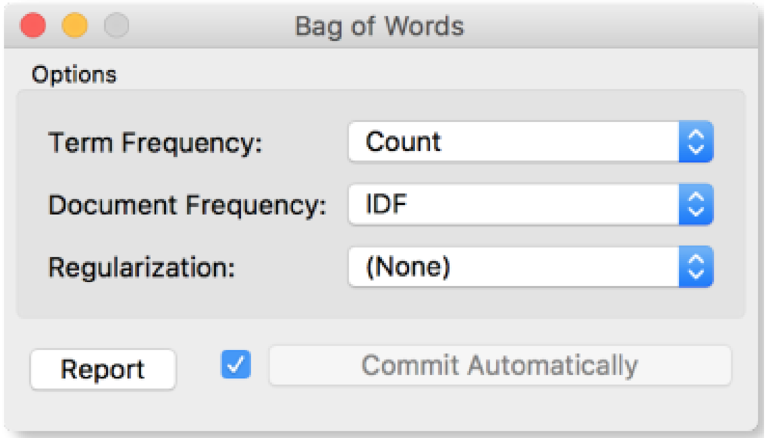
- **Concordance** shows the text around the given word.



Bag of Words

- Bag of Words creates a table with words in columns and documents in rows. Values are word occurrences in each document. They can be binary, but normally they are counts.

	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



Term Frequency-Inverse Document Frequency

- Example: search web pages with keywords "*RapidMiner books that describe text mining.*"
  1. Give a high weightage to those keywords that are relatively rare.
  2. Give a high weightage to those web pages that contain a large number of instances of the rare keywords.
- The highest-weighted web pages are the ones for which the product of these two weights is the highest
- The technique of calculating this weighting is called term **TF-IDF**, which stands for term frequency-inverse document frequency.

TF-IDF

- **Term frequency (TF)**: the ratio of the number of times a keyword appears in a given document, n_k (where k is the keyword), to the total number of terms in the document, n:

$$TF = \frac{n_k}{n}$$

- E.g. "that" has a fairly high TF score, and "RapidMiner" will have a much lower TF score
- **Inverse document frequency (IDF)**:

$$IDF = \log_2\left(\frac{N}{N_k}\right)$$

- N is the number of documents, and N_k is the number of documents that contain the keyword, k
- "that" would arguably appear in every document and, thus, the ratio (N/N_k) would be close to 1, and the IDF score would be close to zero. "RapidMiner" would possibly appear in a relatively fewer number of documents and so the ratio (N/N_k) would be much greater than 1

TF-IDF

$$TF - IDF = \frac{n_k}{n} \times \log_2\left(\frac{N}{N_k}\right)$$

- In the example, when the high TF for “that” is multiplied by its corresponding low IDF, a low (or zero) TF-IDF will be reached, whereas when the low TF for “RapidMiner” is multiplied by its corresponding fairly high IDF, a relatively higher TF-IDF would be obtained
- Typically, TF-IDF scores for every word in the set of documents is calculated in the preprocessing step of the three-step process described earlier.

Example

- Corpus

Document 1	This is a book on data mining
Document 2	This book describes data mining and text mining using RapidMiner

- **Document vector or term document matrix (TDM)**: the matrix with columns consist of all the tokens found in the documents and the cells of the matrix are the counts of the number of times a token appears

Table 9.1 Building a Matrix of Terms From Unstructured Raw Text

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

Example

- **TDM using TF**

Table 9.2 Using Term Frequencies Instead of Term Counts in a TDM

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1/7 = 0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0	0	0	0	0
Document 2	1/10 = 0.1	0	0	0.1	0	0.1	0.2	0.1	0.1	0.1	0.1	0.1

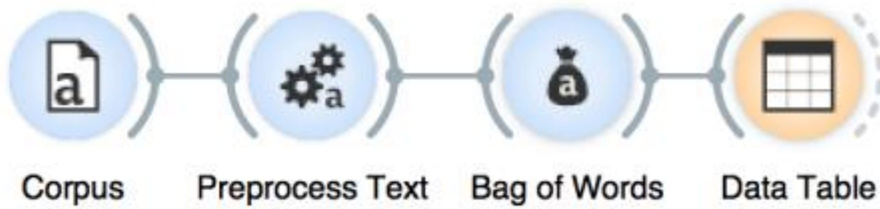
TDM, *Term document matrix.*

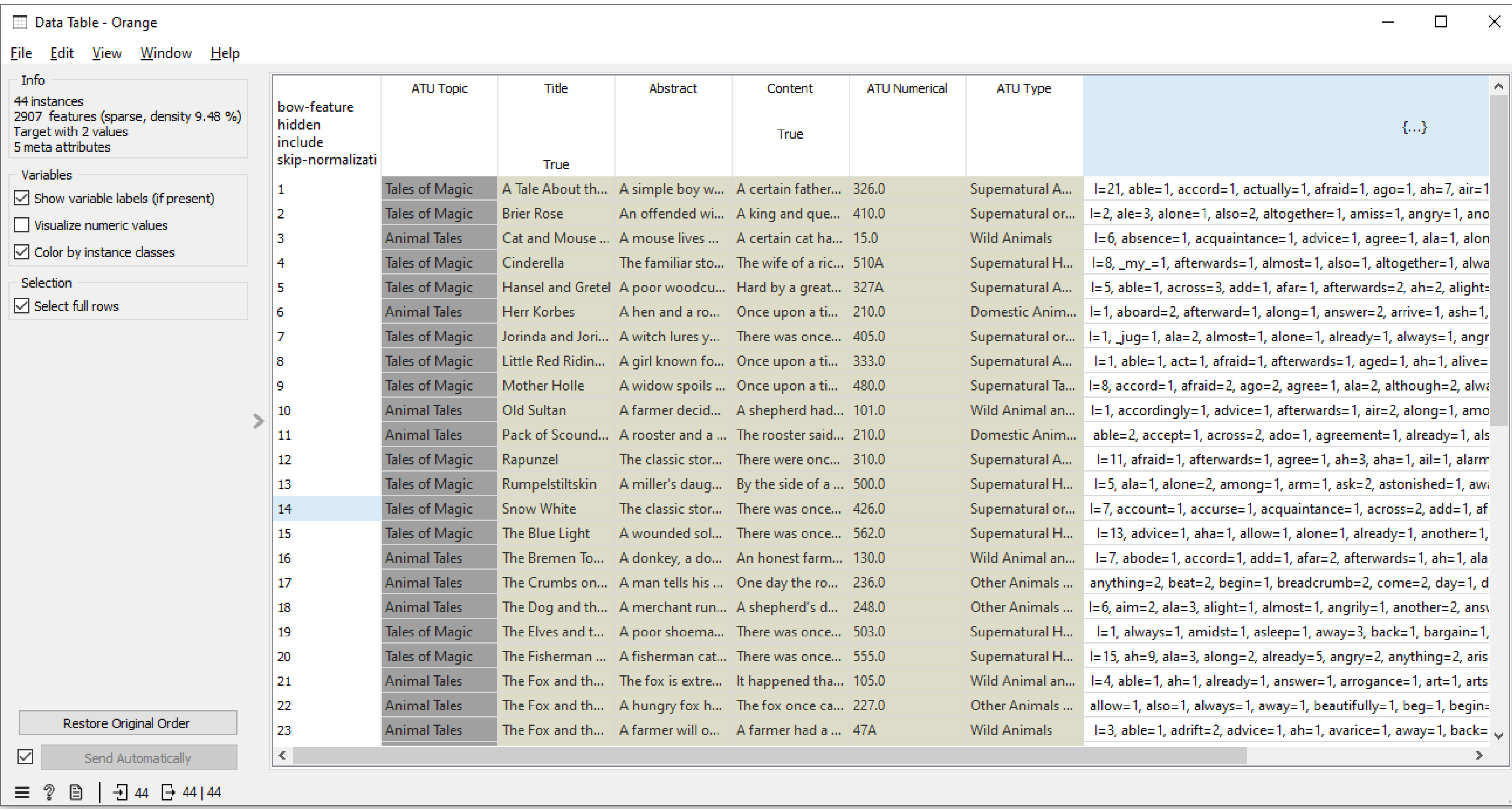
- **TDM using TF-IDF**

ExampleSet (2 examples, 0 special attributes, 12 regular attributes)

Row No.	RapidMiner	This	a	and	book	data	describes	is	mining	on	text	using
1	0	0	0.577	0	0	0	0	0.577	0	0.577	0	0
2	0.447	0	0	0.447	0	0	0.447	0	0	0	0.447	0.447

FIGURE 9.2
Calculating TF—IDF scores for the sample TDM. *TF—IDF*, Term Frequency—Inverse Document Frequency; *TDM*, term document matrix.



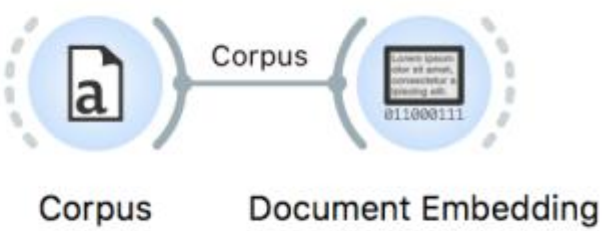


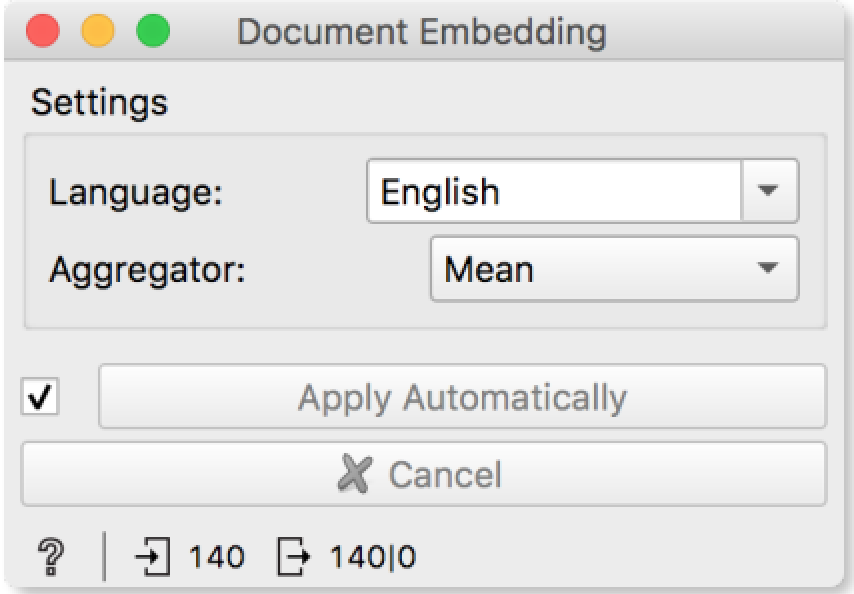
bow-feature hidden include skip-normalizati	ATU Topic	Title True	Abstract	Content True	ATU Numerical	ATU Type	{...}
1	Tales of Magic	A Tale About th...	A simple boy w...	A certain father...	326.0	Supernatural A...	I=21, able=1, accord=1, actually=1, afraid=1, ago=1, ah=7, air=1
2	Tales of Magic	Brier Rose	An offended wi...	A king and que...	410.0	Supernatural or...	I=2, ale=3, alone=1, also=2, altogether=1, amiss=1, angry=1, ano
3	Animal Tales	Cat and Mouse ...	A mouse lives ...	A certain cat ha...	15.0	Wild Animals	I=6, absence=1, acquaintance=1, advice=1, agree=1, ala=1, alon
4	Tales of Magic	Cinderella	The familiar sto...	The wife of a ric...	510A	Supernatural H...	I=8, _my_=1, afterwards=1, almost=1, also=1, altogether=1, alwa
5	Tales of Magic	Hansel and Gretel	A poor woodcu...	Hard by a great...	327A	Supernatural A...	I=5, able=1, across=3, add=1, afar=1, afterwards=2, ah=2, alight=
6	Animal Tales	Herr Korbes	A hen and a ro...	Once upon a ti...	210.0	Domestic Anim...	I=1, aboard=2, afterward=1, along=1, answer=2, arrive=1, ash=1,
7	Tales of Magic	Jorinda and Jori...	A witch lures y...	There was once...	405.0	Supernatural or...	I=1, _jug=1, ala=2, almost=1, alone=1, already=1, always=1, angr
8	Tales of Magic	Little Red Ridin...	A girl known fo...	Once upon a ti...	333.0	Supernatural A...	I=1, able=1, act=1, afraid=1, afterwards=1, aged=1, ah=1, alive=
9	Tales of Magic	Mother Holle	A widow spoils ...	Once upon a ti...	480.0	Supernatural Ta...	I=8, accord=1, afraid=2, ago=2, agree=1, ala=2, although=2, alwa
10	Animal Tales	Old Sultan	A farmer decid...	A shepherd had...	101.0	Wild Animal an...	I=1, accordingly=1, advice=1, afterwards=1, air=2, along=1, amo
11	Animal Tales	Pack of Scound...	A rooster and a ...	The rooster said...	210.0	Domestic Anim...	able=2, accept=1, across=2, ado=1, agreement=1, already=1, als
12	Tales of Magic	Rapunzel	The classic stor...	There were onc...	310.0	Supernatural A...	I=11, afraid=1, afterwards=1, agree=1, ah=3, aha=1, ail=1, alarm
13	Tales of Magic	Rumpelstiltskin	A miller's daug...	By the side of a ...	500.0	Supernatural H...	I=5, ala=1, alone=2, among=1, arm=1, ask=2, astonished=1, awa
14	Tales of Magic	Snow White	The classic stor...	There was once...	426.0	Supernatural or...	I=7, account=1, accurse=1, acquaintance=1, across=2, add=1, af
15	Tales of Magic	The Blue Light	A wounded sol...	There was once...	562.0	Supernatural H...	I=13, advice=1, aha=1, allow=1, alone=1, already=1, another=1,
16	Animal Tales	The Bremen To...	A donkey, a do...	An honest farm...	130.0	Wild Animal an...	I=7, abode=1, accord=1, add=1, afar=2, afterwards=1, ah=1, ala
17	Animal Tales	The Crumbs on...	A man tells his ...	One day the ro...	236.0	Other Animals ...	anything=2, beat=2, begin=1, breadcrumb=2, come=2, day=1, d
18	Animal Tales	The Dog and th...	A merchant run...	A shepherd's d...	248.0	Other Animals ...	I=6, aim=2, ala=3, alight=1, almost=1, angrily=1, another=2, ansv
19	Tales of Magic	The Elves and t...	A poor shoema...	There was once...	503.0	Supernatural H...	I=1, always=1, amidst=1, asleep=1, away=3, back=1, bargain=1,
20	Tales of Magic	The Fisherman ...	A fisherman cat...	There was once...	555.0	Supernatural H...	I=15, ah=9, ala=3, along=2, already=5, angry=2, anything=2, aris
21	Animal Tales	The Fox and th...	The fox is extre...	It happened tha...	105.0	Wild Animal an...	I=4, able=1, ah=1, already=1, answer=1, arrogance=1, art=1, arts
22	Animal Tales	The Fox and th...	A hungry fox h...	The fox once ca...	227.0	Other Animals ...	allow=1, also=1, always=1, away=1, beautifully=1, beg=1, begin=
23	Animal Tales	The Fox and th...	A farmer will o...	A farmer had a ...	47A	Wild Animals	I=3, able=1, adrift=2, advice=1, ah=1, avarice=1, away=1, back=

Document Embedding

- Word embedders are based on pre-trained deep models that map words in the language space. In such a model, words with similar meaning and words from the same family (car, Toyota, vehicle) would be placed close together. Computing a vector for an individual word based on the model is called embedding.

Orange uses **fastText** pre-trained models to embed words. Then is averages word vectors to produce a single document vector (one can also use sum, min or max aggregation)





Data Table - Orange

File Edit View Window Help

Info

44 instances (no missing data)
384 features
Target with 2 values
5 meta attributes

Variables

- [x] Show variable labels (if present)
- [ ] Visualize numeric values
- [x] Color by instance classes

Selection

- [x] Select full rows

embedding-featu hidden include	5	Dim376 True True	Dim377 True True	Dim378 True True	Dim379 True True	Dim380 True True	Dim381 True True	Dim382 True True	Dim383 True True	Dim384 True True
1	249426	-0.229823	-0.0634372	0.265132	-0.146508	-0.0758463	0.0870508	0.341508	0.338517	-0.00145558
2	792258	-0.169549	-0.00454659	0.242163	-0.144142	-0.0581509	0.139689	-0.0959346	0.119618	-0.222717
3	197353	0.0643316	0.0430067	0.18408	0.00729054	0.0209662	0.193949	0.0181829	0.0519499	0.16875
4	179557	-0.103695	0.191948	0.0519287	-0.106108	0.171174	-0.113088	-0.204811	0.245933	0.0449549
5	131421	-0.0155922	-0.05335	0.167877	-0.0816081	-0.118979	0.076868	0.039222	0.0622554	0.0884285
6	821059	0.0163652	0.0518748	-0.0167726	0.0884752	-0.162041	0.0368435	0.485523	-0.0433478	-0.126416
7	115041	-0.163554	0.0252003	0.24984	-0.191082	-0.11889	0.0120342	-0.0989318	0.138529	0.0829685
8	461769	-0.115161	0.0990756	0.0475949	-0.11352	0.135115	0.181822	0.0837991	0.402852	0.0173896
9	742271	-0.317256	0.0998543	0.0577679	-0.246113	0.126758	0.490267	0.172305	0.33863	0.138665
10	202021	-0.00229644	0.245644	-0.0710565	0.0610021	0.318116	0.0973136	-0.15862	0.151904	0.100394
11	640289	-0.152391	0.101379	-0.0709826	0.0984134	-0.256391	0.0732751	0.180499	-0.0212355	0.0520328
12	279669	-0.137338	-0.213381	0.0141127	-0.129649	-0.139572	-0.0162687	0.177856	-0.0851691	0.254717
13	228903	-0.0940455	0.0229882	-0.193839	-0.326457	-0.245273	0.225871	-0.0546229	0.130358	-0.015537
14	269349	-0.14028	0.0581219	0.118484	-0.0354942	-0.049622	0.113807	0.238216	0.267491	0.198762
15	639879	-0.0722747	0.167852	0.268028	-0.000667217	-0.0653031	0.10719	-0.084943	0.104572	-0.295958
16	505448	-0.178639	-0.0278463	0.0136296	-0.151518	-0.0130818	-0.112767	0.059074	0.190738	0.0565506
17	).15043	-0.107508	0.308656	-0.122127	0.328695	-0.135214	-0.141353	0.269168	0.245852	-0.0656533
18	065451	-0.179756	0.0773088	0.139588	-0.0303722	0.0545047	0.298319	0.203727	-0.0807735	0.0524545
19	407366	-0.217026	-0.109342	0.147077	-0.350409	-0.208917	0.0333798	-0.322399	0.150257	0.0615454
20	298968	0.155971	-0.089627	-0.0609872	-0.246158	0.0284492	0.204959	0.0237995	0.0166593	0.194241
21	113902	0.0626725	0.199996	-0.0438538	-0.0496769	0.0117078	0.119782	0.249785	-0.0533808	0.17462
22	167204	0.0156366	0.201158	-0.201029	-0.00384904	-0.187391	0.413832	0.171721	0.236455	0.186764
23	209974	-0.0760384	0.00606273	0.0578814	0.0460216	-0.0777282	-0.0386007	0.231094	0.270226	0.0152654
24	117644	0.0675185	-0.100894	0.087747	-0.191176	0.118478	0.0103163	0.0372159	0.0304638	0.405901

Restore Original Order

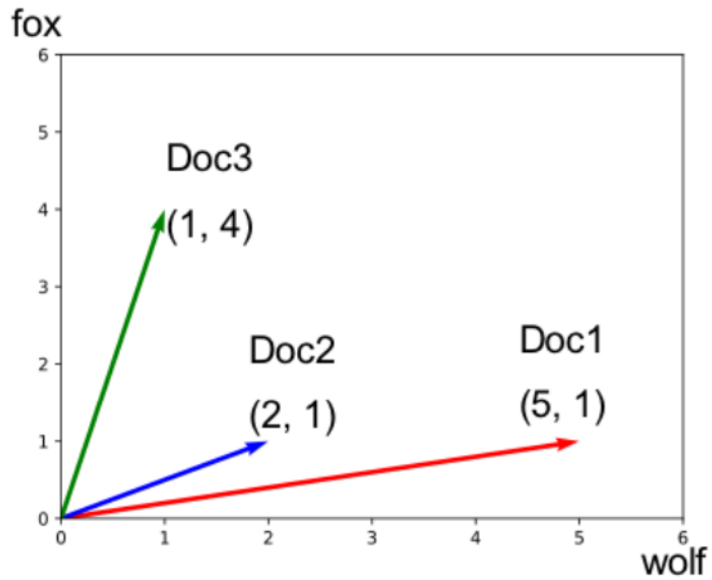
- [x] Send Automatically

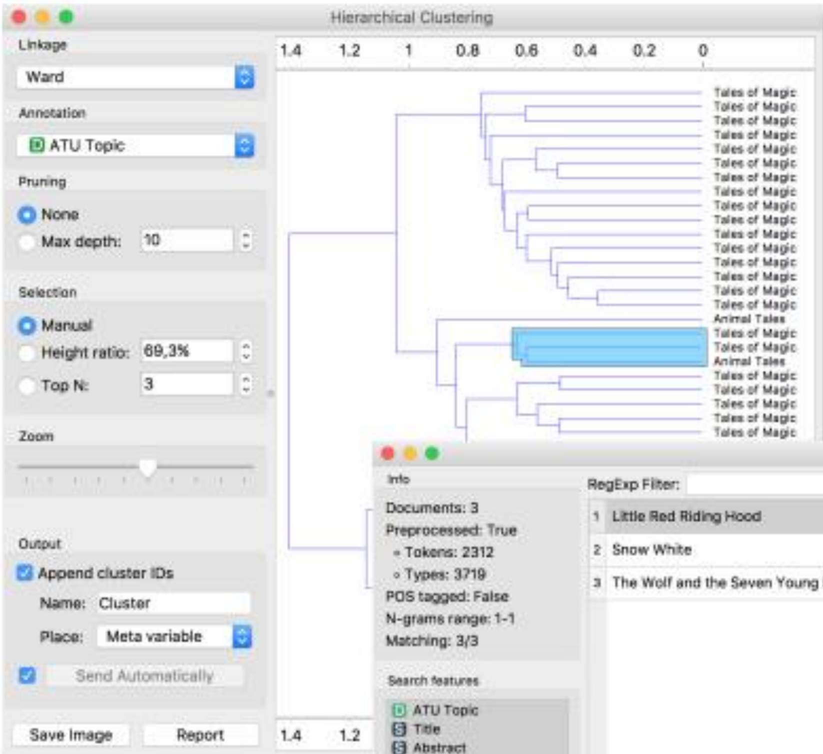
44 | 44 | 44

Clustering & Distances

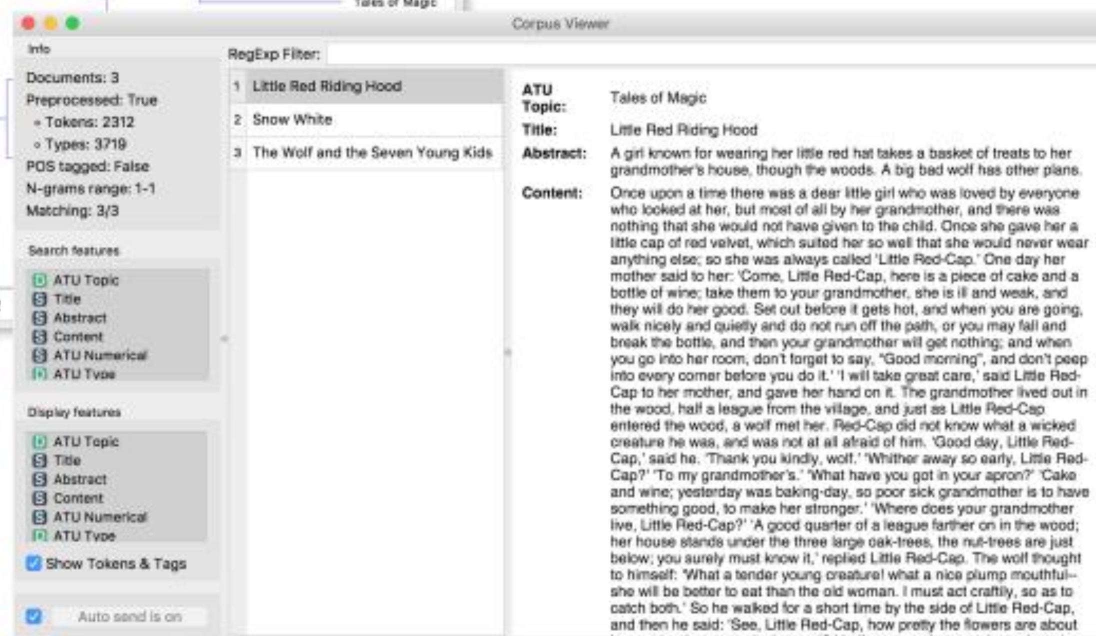
- One common task in text mining is finding interesting groups of similar documents. That is, we would like to identify documents that are similar to each other.
- We normally use Euclidean distance to measure the similarity, but the Euclidean distance is not the only option.
- There are many distance measures and Euclidean doesn't work very well for text.

An example of the similarity





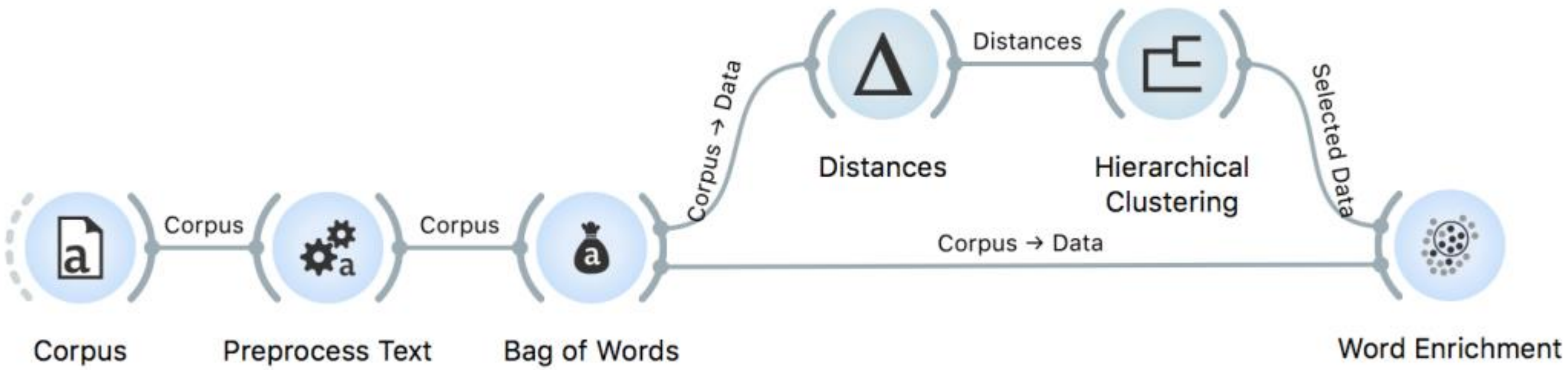
Connect Corpus Viewer to Hierarchical Clustering and open both widgets. Now click on a cluster in the dendrogram and observe the documents from the selected cluster in Corpus Viewer. Explore different clusters. Why are some Tales of Magic mixed with Animal Tales? What do they have in common?

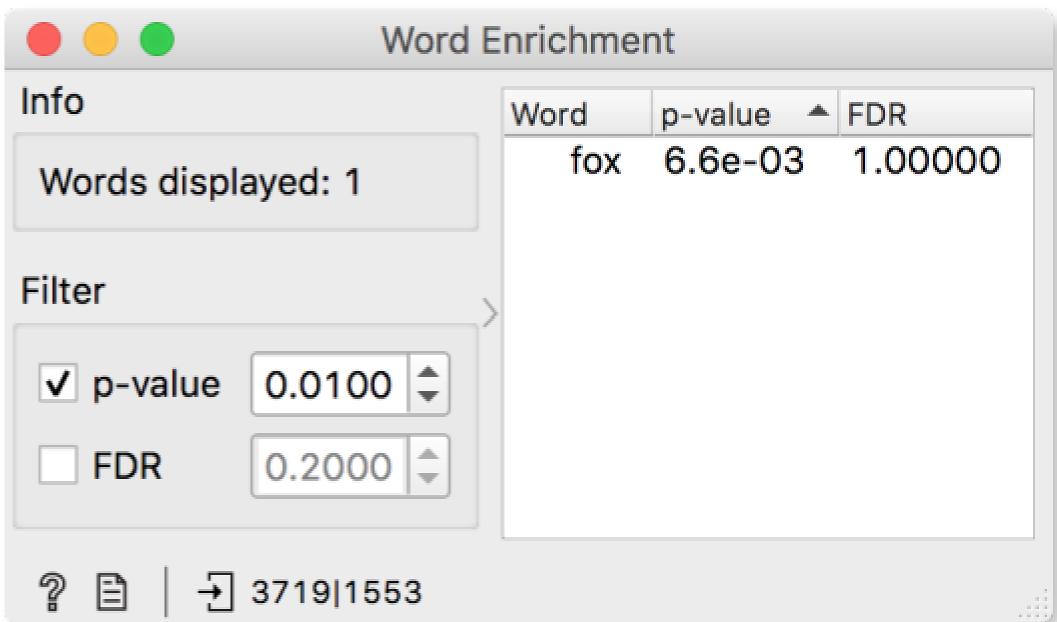


Word Enrichment

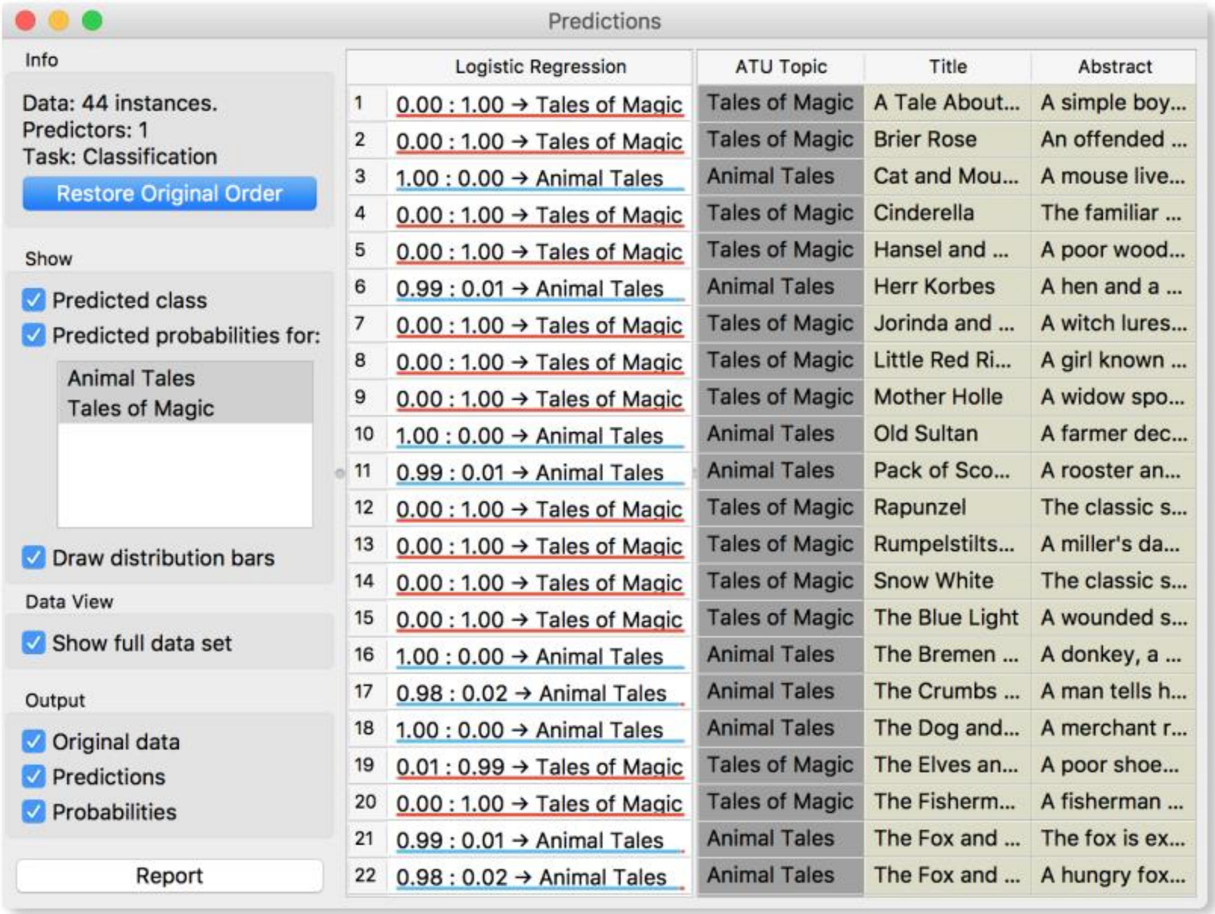
- Word Enrichment compares a subset of documents against the entire corpus and finds statistically significant words for the selected subset. It uses hypergeometric p-value to find words, that are overrepresented in the subset.

$$p = \frac{\binom{\text{term in corpus}}{\text{term in subset}} \times \binom{\text{other terms}}{\text{other terms in subset}}}{\binom{\text{all terms}}{\text{terms in subset}}}$$


Corpus
Corpus
Corpus
Preprocess Text
Corpus
Bag of Words
Corpus → Data
Distances
Distances
Distances
Hierarchical Clustering
Selected Data
Corpus → Data
Word Enrichment

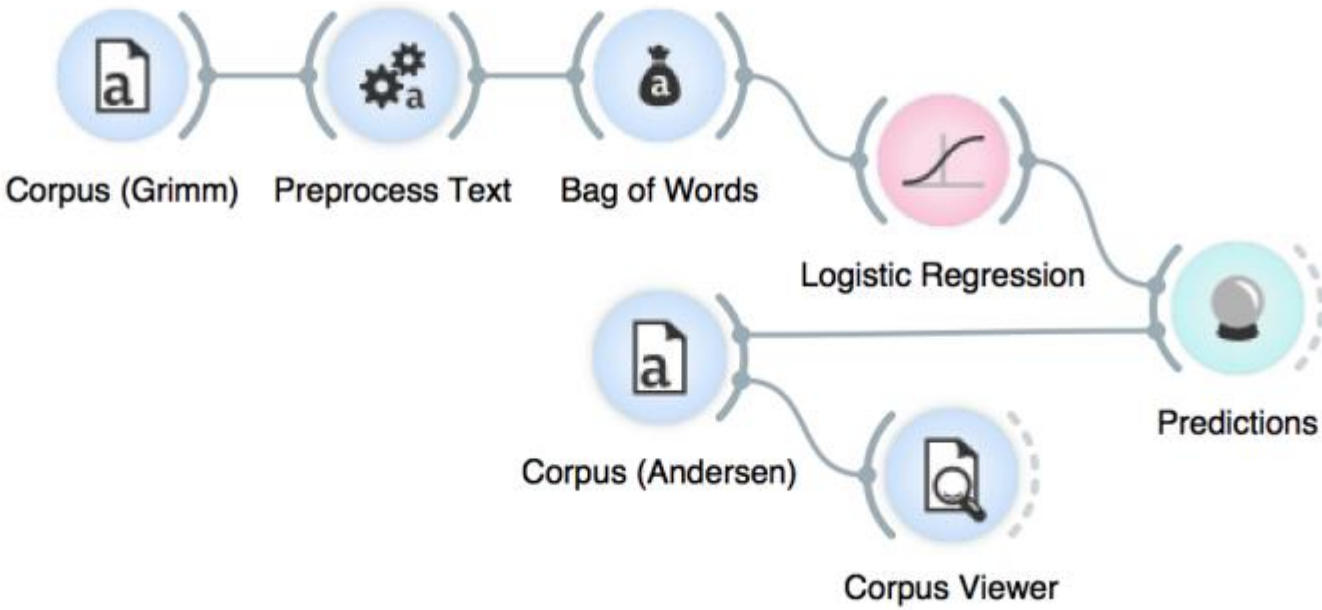

Word Enrichment
Info
Words displayed: 1
Filter
p-value 0.0100
FDR 0.2000
Word p-value FDR
fox 6.6e-03 1.00000
3719|1553

Classification



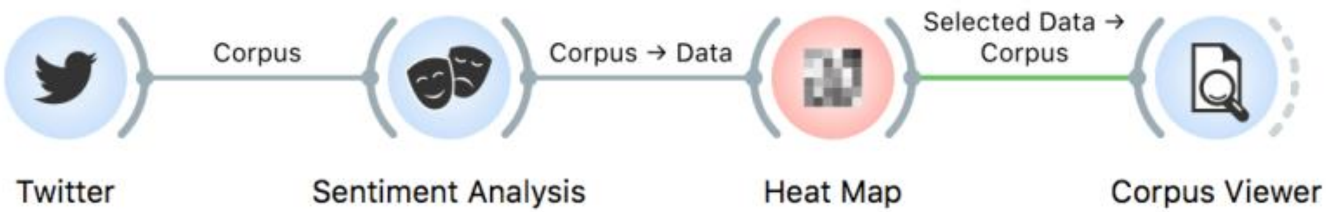
	Logistic Regression	ATU Topic	Title	Abstract
1	0.00 : 1.00 → Tales of Magic	Tales of Magic	A Tale About...	A simple boy...
2	0.00 : 1.00 → Tales of Magic	Tales of Magic	Brier Rose	An offended ...
3	1.00 : 0.00 → Animal Tales	Animal Tales	Cat and Mou...	A mouse live...
4	0.00 : 1.00 → Tales of Magic	Tales of Magic	Cinderella	The familiar ...
5	0.00 : 1.00 → Tales of Magic	Tales of Magic	Hansel and ...	A poor wood...
6	0.99 : 0.01 → Animal Tales	Animal Tales	Herr Korbes	A hen and a ...
7	0.00 : 1.00 → Tales of Magic	Tales of Magic	Jorinda and ...	A witch lures...
8	0.00 : 1.00 → Tales of Magic	Tales of Magic	Little Red Ri...	A girl known ...
9	0.00 : 1.00 → Tales of Magic	Tales of Magic	Mother Holle	A widow spo...
10	1.00 : 0.00 → Animal Tales	Animal Tales	Old Sultan	A farmer dec...
11	0.99 : 0.01 → Animal Tales	Animal Tales	Pack of Sco...	A rooster an...
12	0.00 : 1.00 → Tales of Magic	Tales of Magic	Rapunzel	The classic s...
13	0.00 : 1.00 → Tales of Magic	Tales of Magic	Rumpelstilts...	A miller's da...
14	0.00 : 1.00 → Tales of Magic	Tales of Magic	Snow White	The classic s...
15	0.00 : 1.00 → Tales of Magic	Tales of Magic	The Blue Light	A wounded s...
16	1.00 : 0.00 → Animal Tales	Animal Tales	The Bremen ...	A donkey, a ...
17	0.98 : 0.02 → Animal Tales	Animal Tales	The Crumbs ...	A man tells h...
18	1.00 : 0.00 → Animal Tales	Animal Tales	The Dog and...	A merchant r...
19	0.01 : 0.99 → Tales of Magic	Tales of Magic	The Elves an...	A poor shoe...
20	0.00 : 1.00 → Tales of Magic	Tales of Magic	The Fisherm...	A fisherman ...
21	0.99 : 0.01 → Animal Tales	Animal Tales	The Fox and ...	The fox is ex...
22	0.98 : 0.02 → Animal Tales	Animal Tales	The Fox and ...	A hungry fox...

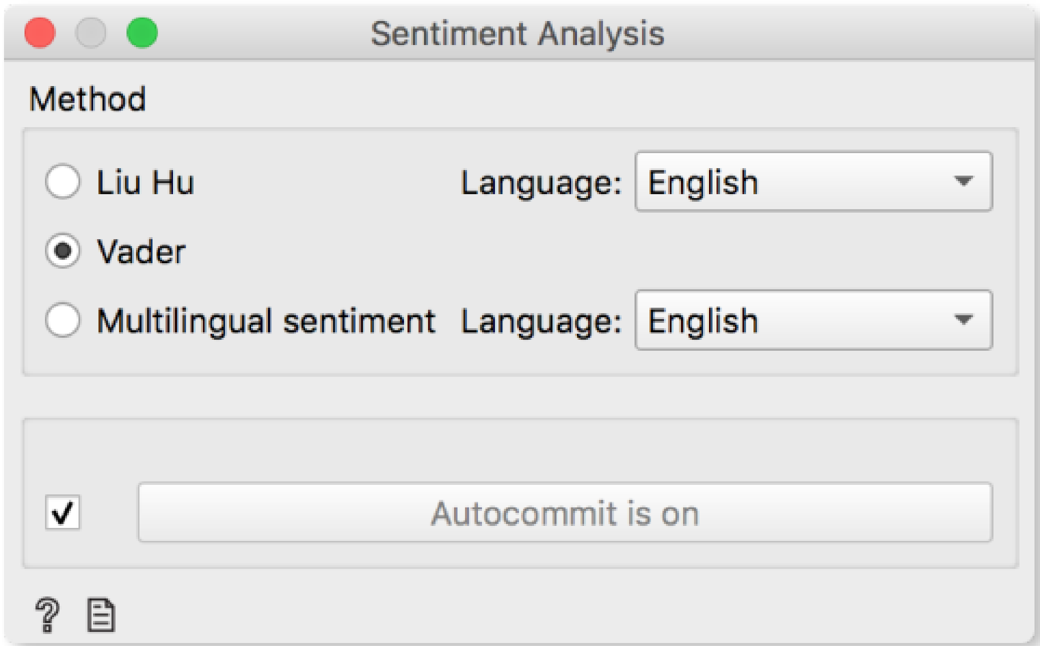
Predictions



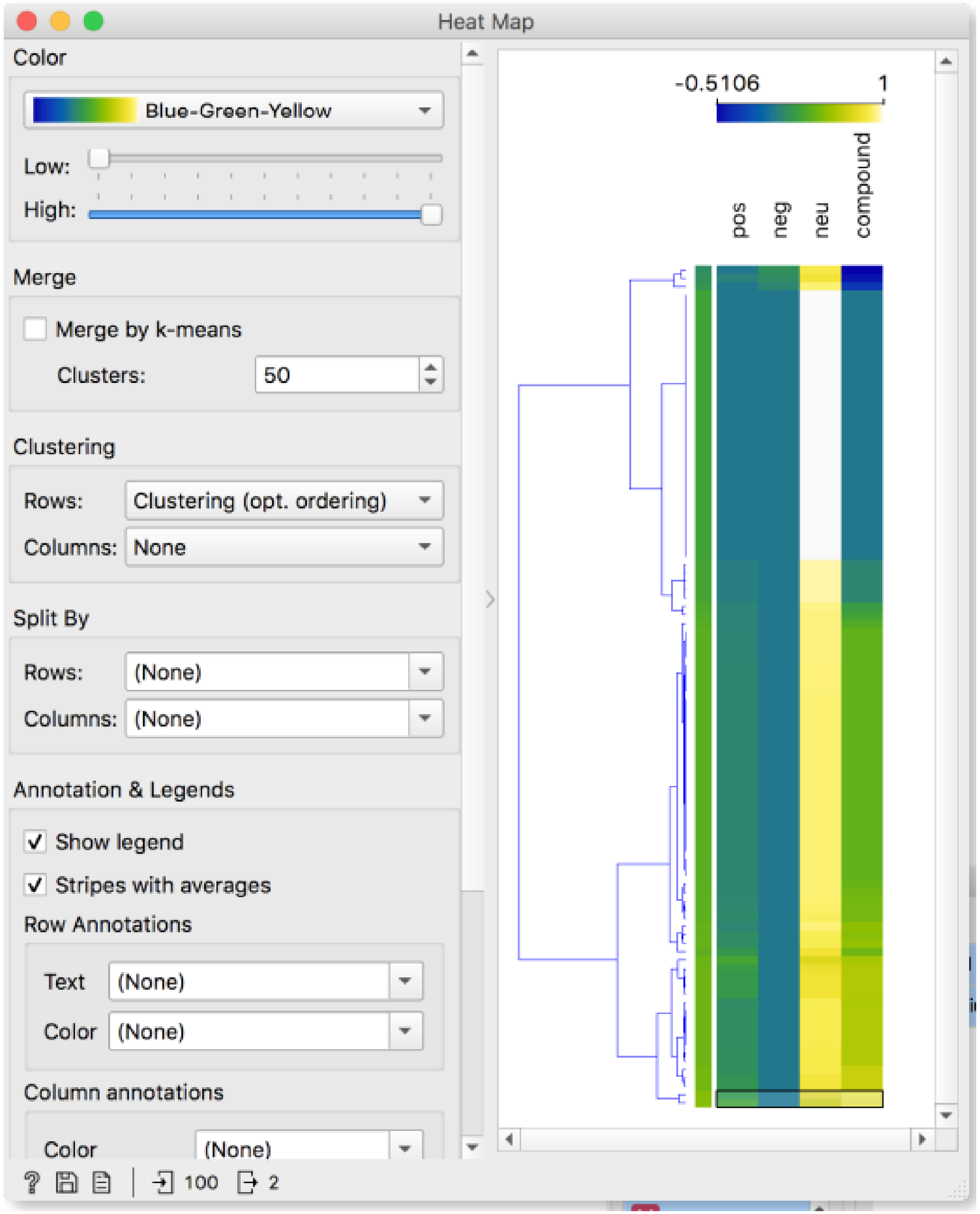
	Logistic Regression	Title	Content
1	0.01 : 0.99 → Tales of Magic	The Little Match-Seller	It was terribly cold and nearly dark on...
2	0.00 : 1.00 → Tales of Magic	The Philosopher's Stone	Far away towards the east, in India, w...
3	0.90 : 0.10 → Animal Tales	The Ugly Duckling	It was lovely summer weather in the c...

Sentiment Analysis





More advanced techniques for sentiment analysis are based on models, usually with deep neural networks that learn from a large amount of labelled texts.


Heat Map
Color
Blue-Green-Yellow
Low:
High:
Merge
Merge by k-means
Clusters: 50
Clustering
Rows: Clustering (opt. ordering)
Columns: None
Split By
Rows: (None)
Columns: (None)
Annotation & Legends
Show legend
Stripes with averages
Row Annotations
Text (None)
Color (None)
Column annotations
Color (None)
100
2
-0.5106
1
pos
neg
neu
compound