

BIOLAB AND COLLABORATORS

USING ORANGE

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, November 2021

Contents

Network from Text 5

Bibliography 9

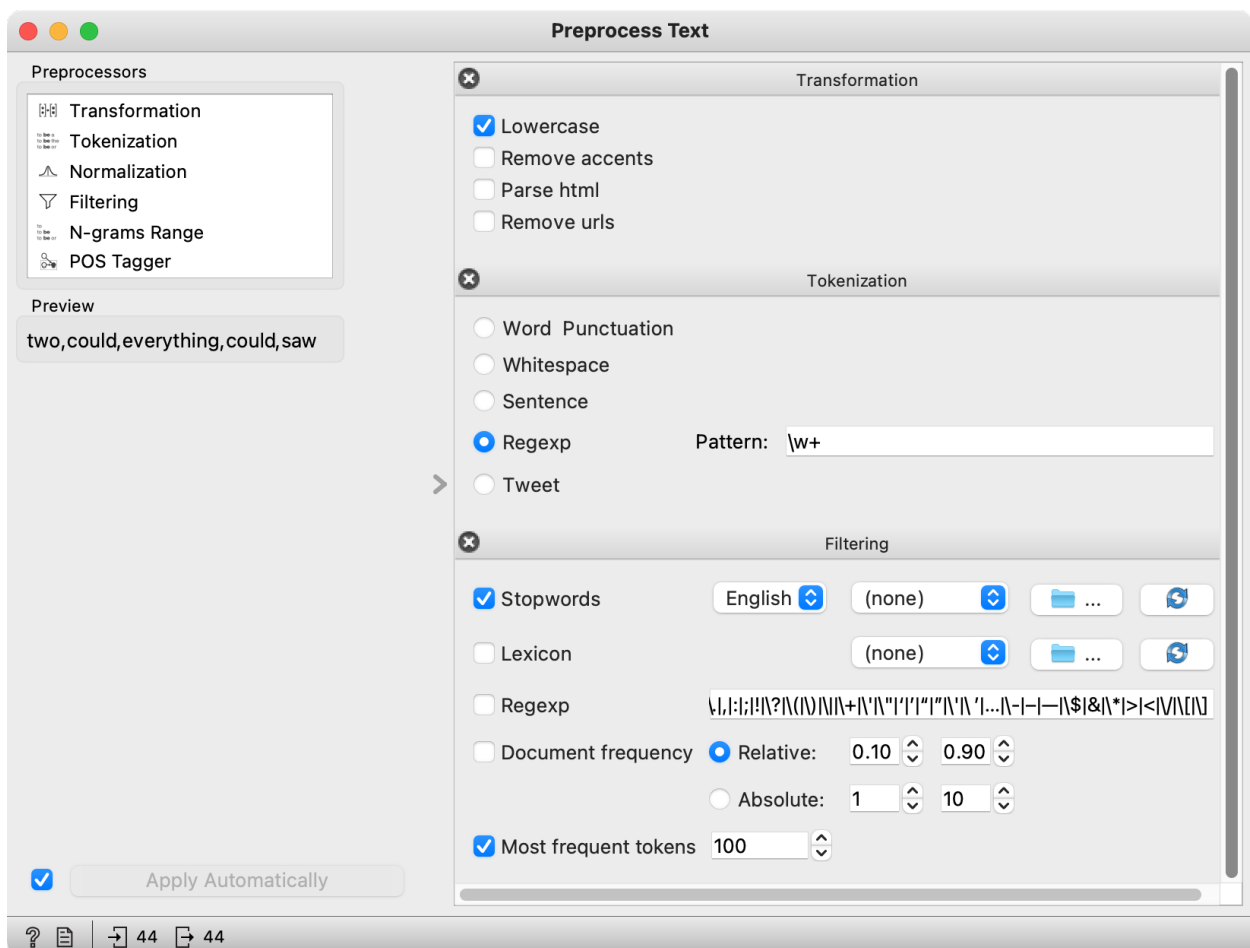
Index 10

Network from Text

Now, let us try to generate networks from text. What do we mean by that? How can a text be transformed to a network? Well, documents are nodes, while edges can be the number of shared words. Again, we are back to similarity, but applied to text.

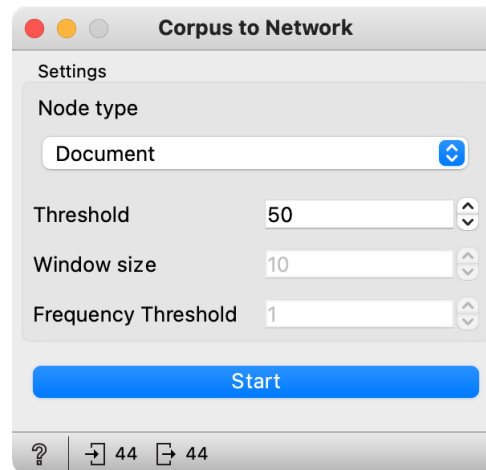
For this task, we will need the *Text add-on*. Load *grimm-tales-selected* with the *Corpus* widget. This data set contains 44 Grimm's tales, some of which are tales of magic and some are animal tales.

Every text needs to be preprocessed, that is we have to split the text to words and remove those words that have no meaning (such as stopwords). To speed up the analysis, we will keep only 100 most frequent tokens.

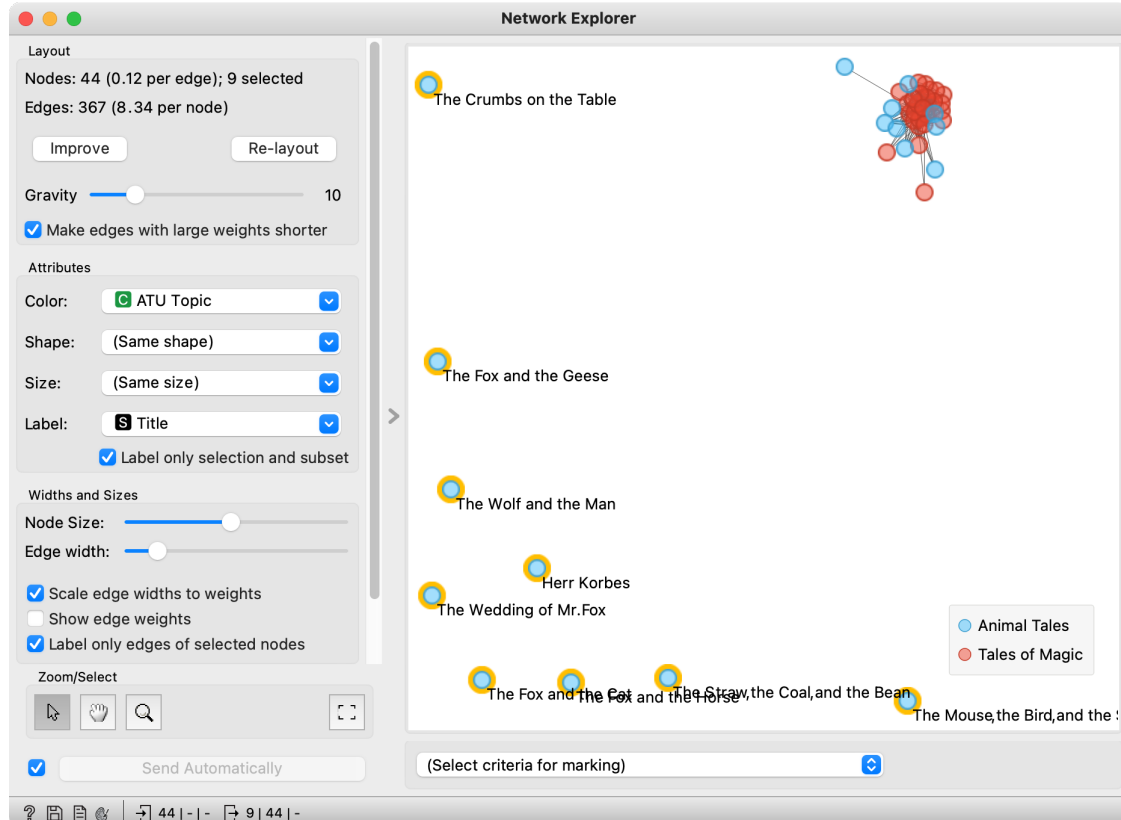


Now pass the data to *Corpus to Network*. With this widget, we will generate the graph. If we are generating network where nodes are documents, then we need to set a single parameter, namely the

threshold. This is similar to the similarity threshold in *Network from Distances*. *Threshold* will define how many words the documents have to share for them to have a connecting edge. In our case, we will set the threshold quite high - two documents have to share at least 50 words to be connected with an edge.

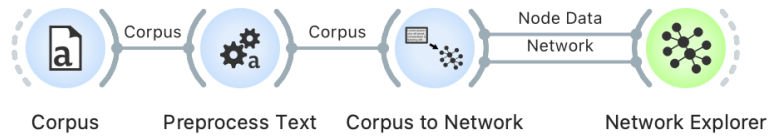


Let us observe the end result in *Network Explorer*. Seem like Tales of Magic are well-connected, even with some Animal Tale, while certain Animal Tales are quite distinct and don't share as many words with the other tales.



A task for the reader: play with the threshold and observe how

the graph changes. Does a lower threshold results in a more or less connected graph? What happens if words are used as Node types in *Corpus to Network*? What does such a graph show?



Bibliography

<https://github.com/biolab/orange3>, a.

<https://github.com/quasars>, b.

<https://quasar.codes>.

<https://quasar.codes>.

Robert Bringhurst. *The Elements of Typography*. Hartley & Marks, 3.1 edition, 2005. ISBN 0-88179-205-5.

Frank Mittelbach and Michel Goossens. *The L^AT_EX Companion*. Addison–Wesley, second edition, 2004. ISBN 0-201-36299-6.

Edward R. Tufte. *Envisioning Information*. Graphics Press, Cheshire, Connecticut, 1990. ISBN 0-9613921-1-8.

Edward R. Tufte. *Visual Explanations*. Graphics Press, Cheshire, Connecticut, 1997. ISBN 0-9613921-2-6.

Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 2001. ISBN 0-9613921-4-2.

Edward R. Tufte. *Beautiful Evidence*. Graphics Press, LLC, first edition, May 2006. ISBN 0-9613921-7-7.

Hideo Umeki. The geometry package. <http://ctan.org/pkg/geometry>, December 2008.

Index

license, [2](#)