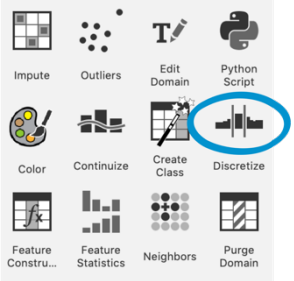
Data Science Infrastructures – Exercise 03 (DSI E03 ST 24)

02/05/2024 / Philipp Wieder / philipp.wieder@gwdg.de

General Prerequisites

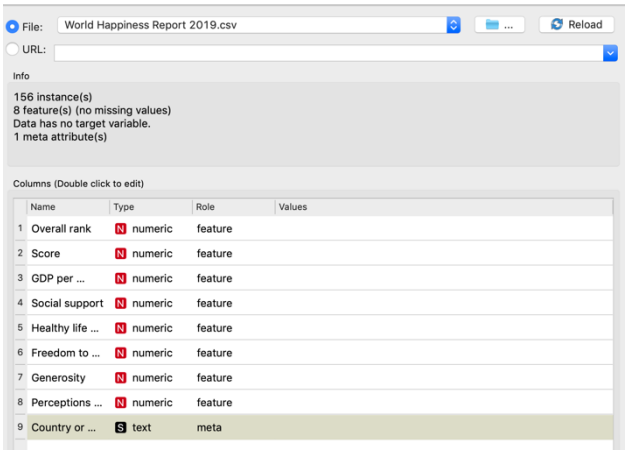
To fulfil this assignment, you need the Software Orange, which you can download from https://orangedatamining.com/. Please follow the installation instructions for your respective operation system (https://orangedatamining.com/download/). Orange is freely available and comes with a number of so-called widgets that offer user interfaces for particular data science tasks. It has a comprehensive documentation including tutorials (https://orangedatamining.com/docs/) and comes with a number of example workflows (https://orangedatamining.com/examples/).

Assignment I – Data Type Portability: Discretization



As part of the Pre-Processing lecture we came across a step called data type portability, which is used to convert between different data types to make it easier to process certain data sets or apply standard algorithms. The conversion of numeric data types into categorical ones is called discretization. Orange has a particular widget for this task.

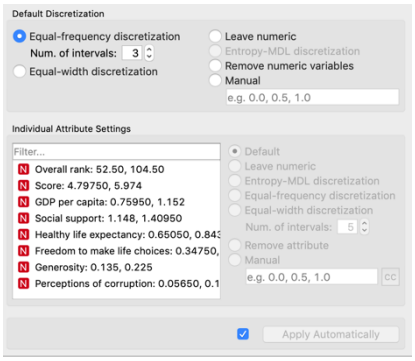
Select a Data Set



The first thing to be done is to select a data set. In the exercise folder, you can find the "*World Happiness Report 2019.csv*[1]". Select the "File" widget, choose the aforementioned file from wherever you have stored it on your machine, and it will be loaded into your widget. Double-clicking on the icon in Orange opens a dialogue that provides basic information about the data (like e.g. the number of instances and features) and lists all features and their respective data types. In our example, all data is numeric except the name of the country.

[1] Downloaded from https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2019.csv.

Apply Discretization

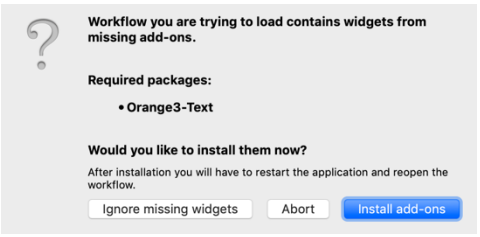


Now choose the "Discretization" widget and link it to the "File" widget. You can either do this by clicking on the dotted curve at the right of the "File" widget and then select the "Discretization" widget form the drop-down list or you place the "Discretization" widget directly on the main panel of Orange and then link both widgets. Just make sure that they are in the right order, i.e. the output from the "File" widget (which is the data set) is the input to the "Discretization" widget. Per default, equal-frequency discretization with three intervals is chosen. This splits each feature/attribute into three intervals, so that they each contain approximately the same number of instances. You can change the number of intervals or chose another discretization method, either for all attributes or individually[2]. It does not necessarily make sense to discretize all attributes, so you can also leave it numeric.

Assignment

Discretize the attribute "Overall rank" into the Top 10, Top 11-50, and Top 51-100 of the happiest countries in the world in 2019. Leave all other attributes as is (you can apply settings to a selection of attributes at once). Look at the results and compare it to the original data set (you can use the "Data Table" widget for that). Then use the "Select Rows" widget to select the Top 10 and look at the results. Make screenshots of the complete workflow, the discretization settings, and the resulting data.

Assignment II – Pre-Processing of Text Data

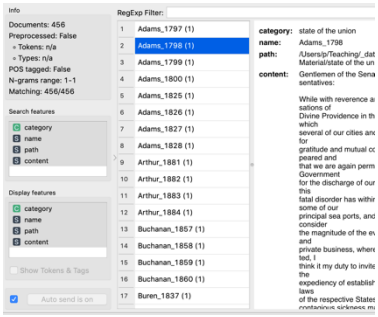


For this assignment, you need a widget collection called "Text Mining". One way is to download a text mining workflow example (e.g. "*620-text-clustering.ows*" from the exercise folder on Stud.IP[3]), because Orange then recognizes the missing widgets and asks whether you want to install them. Install the add-ons and you have everything you need for the assignment. But you can also install it using the "Add-ons" manager from the "Options" menu.

[2] More info can be found here: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/data/discretize.html.
[3] More Workflows can be found here: https://orangedatamining.com/examples/.

Select a Data Set

A data set comprising text data is often called a corpus. This is also the term used in Orange. Use the "Import Documents" widget and select "state of the union" folder, which contains the annual addresses by the presidents of the United States between 1719 and 2018. It takes a while to import the texts. To have a closer look at the corpus, select the "Corpus Viewer" widget and connect it to the "Import Documents" widget. You can use it to have a closer look at your corpus, the features, and other information.



Pre-Processing of Text Data

Chose the "Preprocess Text" widget and connect it to the "Import Documents" widget. Double-clicking on it opens the pre-processing dialogue. You can remove or add pre-processing steps as you like. Some of the steps have already been discussed during the lecture. Please have a look at the documentation https://orangedatamining.com/widget-catalog/text-mining/preprocesstext/ for more information on the different steps/methods.

Assignment

a) First, remove all preprocessing steps except "Transformation" and transform the corpus into lowercase. This avoids to have multiple versions ("Large" and "large") of the same word. If you checked the box "Apply Automatically" on the lower left, the changes are directly applied. You can use the "Word Cloud" widget to observe the effect of the transformation. Please note that also punctuation symbols are also counted. Save the word cloud clicking on the disk symbol. 

   NOTE: For some reason, the "Word Cloud" widget does sometimes only show about 10 words. By ticking the "Color words" box once or twice, the problem is resolved.

b) Now apply "Tokenization". This is used to break the text into smaller pieces like sentences, words, … Please have a look at the aforementioned documentation to understand the different tokenization methods. Use Regular Expressions to split the text by words without keeping punctuation. This is quite a common way to break down text. What are the top 3 words? Save the word cloud clicking on the disk symbol.
c) Now filter the stopwords. What are the top 3 words now? Save the word cloud clicking on the disk symbol.
d) Apply the standard normalization (Porter Stemmer). This does, to put it simple, convert words to their base form, like e.g. "the boy's cars are different colors" → "the boy car be differ color". What do you observe?

Please also provide screenshots of the complete workflow, the preprocessing settings, and the different word clouds.

If you are interested, try out the other pre-processing steps or use different settings.