Topic : Text mining using the Orange tool

Supervisor: Nikos Karacapilidis

Panagiotis Prasinos AM:247030

Department of Mechanical and Aeronautical Engineering

Purpose

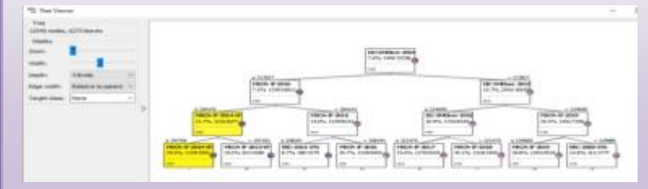
This thesis deals with the widest field of machine learning. It investigates some methods of extracting text data, which are analyzed on a theoretical and practical level, and how they lead in some cases to the categorization of texts. All this, in addition to a theoretical background, also acquires the practice of Orange tool [1].

Conclusions

The experiments that were carried out helped us to understand some of the possibilities of the program. In the end what we can mention as a safe conclusion is that text classification in Orange can bring excellent results implemented in a variety of ways and with quite a large amount of data. Also the preprocessing of our data seems necessary in most of the experiments that were performed. The export of keywords[2] can be applied to large volumes of data and give us the words we want at the level of importance that we define with p-value[3] . As for the other techniques, most of them help a lot in extracting useful information by improving the visualization and analysis of our data.

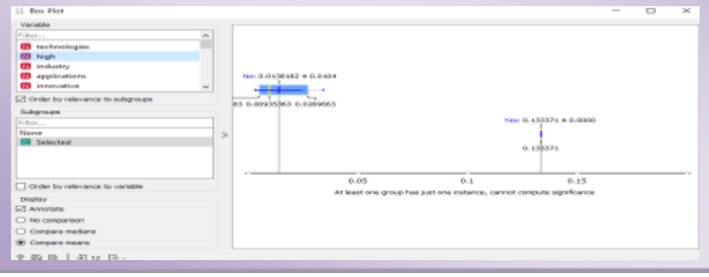
Experiments in Orange

1. Text classification

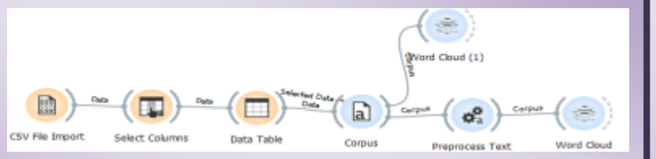


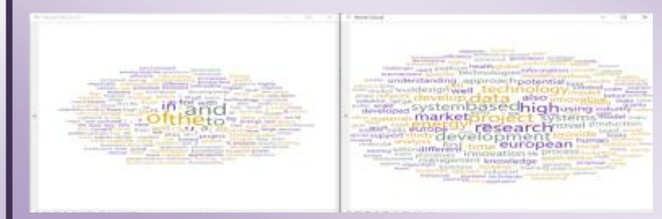
2. Keyword extraction using the word enrichment plugin

3.Topic modeling

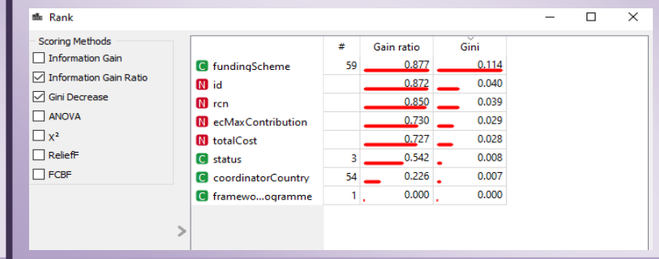


4.Text preprocessing

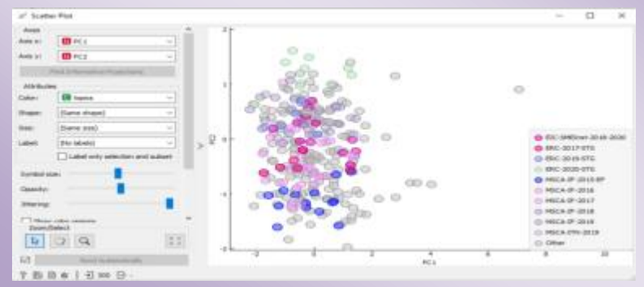




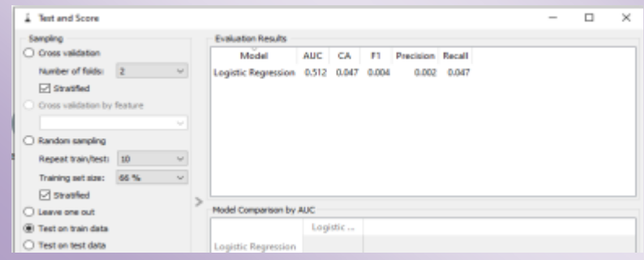
5.Feature Ranking



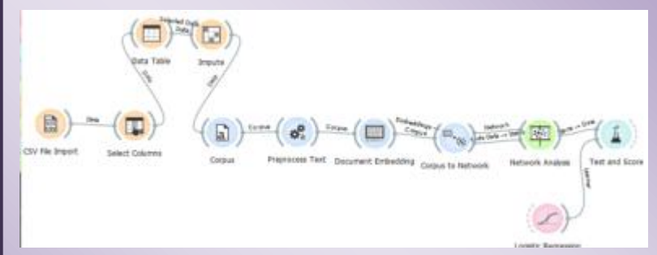
6.Principal Component Analysis (PCA)

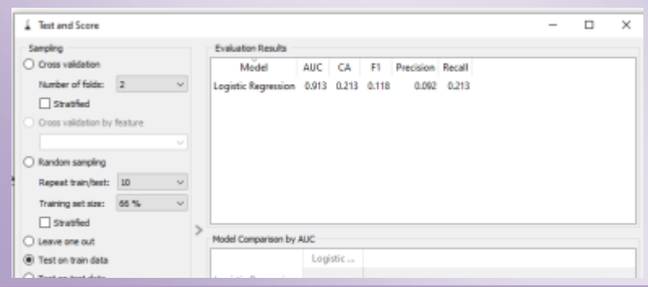


7. Text classification using Document Embedding

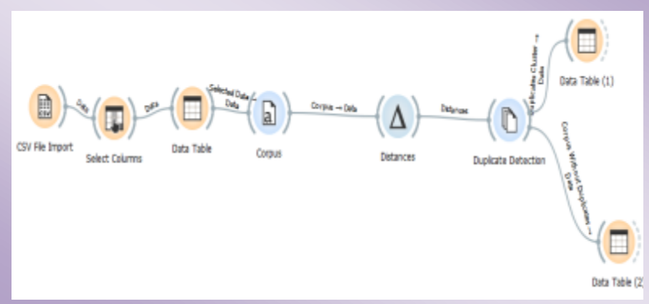


8. Text classification by combining Corpus to Network and Document Embedding





9. Duplicate detection



Bibliography : [1] https://orangedatamining.com/

[2] https://monkeylearn.com/keyword-extraction/

[3] https://en.wikipedia.org/wiki/P-value