UPH

UNIVERSITAS PELITA HARAPAN

Data Science

Introduction to Orange3

Training Series

Week 3

2 Februari 2024

Orange3

Prepared by I Made Murwantara, Universitas Pelita Harapan

Overview

1. Image Clustering
2. Text Analytics
3. Deployment in Streamlit

Add-on:

- Image Analytics
- Text Mining

Image Embedding

- a lower-dimensional representation of the image
- a **dense vector representation of the image** which can be used for many tasks such as classification
- Embeddings are different from images in their raw form. An image file contains RGB data that says exactly what colour each pixel is.
- Embeddings encode information that represents the contents of an image.
- These embeddings are unintelligible in their raw form, just as images are when read as a list of numbers.
- It is when you use embeddings that they start to make sense.

https://www.activeloop.ai/resources/generate-image-embeddings-using-a-pre-trained-cnn-and-store-them-in-hub/
https://blog.roboflow.com/what-is-an-image-embedding/

Image Embedding

- This image contains a bowl of fruit.
- An image embedding will encode this information
- We could then compare the image embedding to a text embedding like "fruit" to see how similar the concept of "fruit" is to the contents of the image.

- We could take two prompts, such as "fruit" and "vegetable", and see how similar each one is.
- The most similar prompt is considered the most representative of the image.

https://blog.roboflow.com/what-is-an-image-embedding/

Image Embedding

- Deep learning is used to develop models that transform complex objects to vectors of numbers.
- Deep learning requires a lot of data (thousands, possibly millions of data instances) and processing power to prepare the network.
- We will use one which is already prepared.

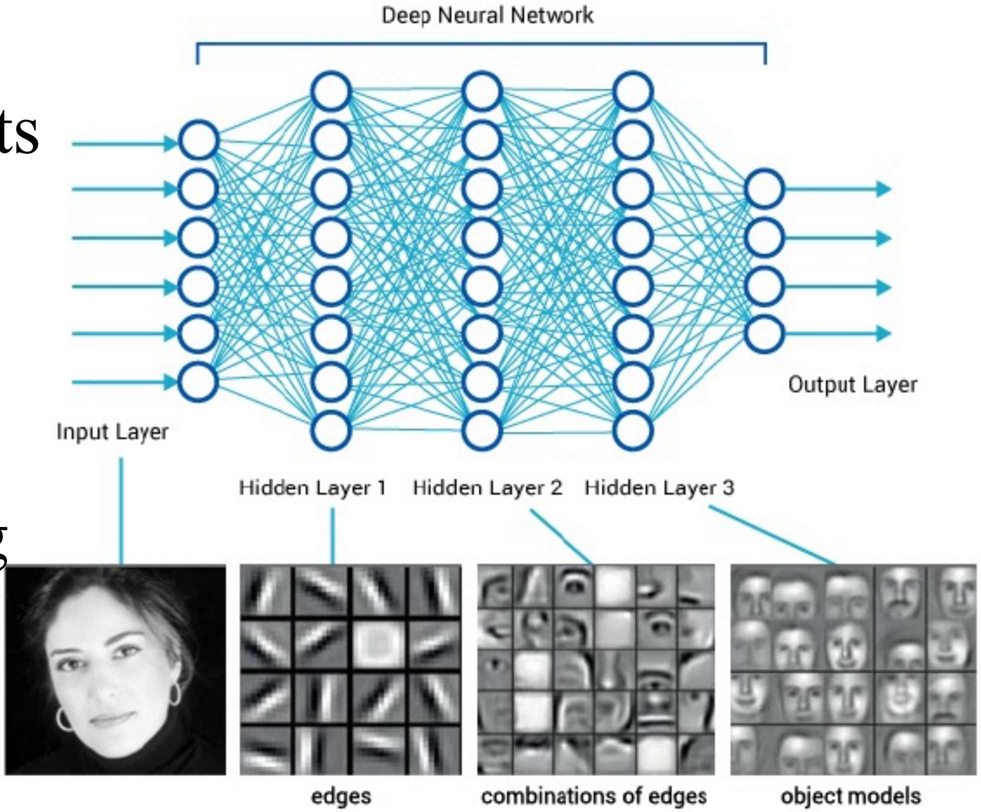


Image Embedding

- Dataset:
  - – http://file.biolab.si/images/domestic-animals.zip

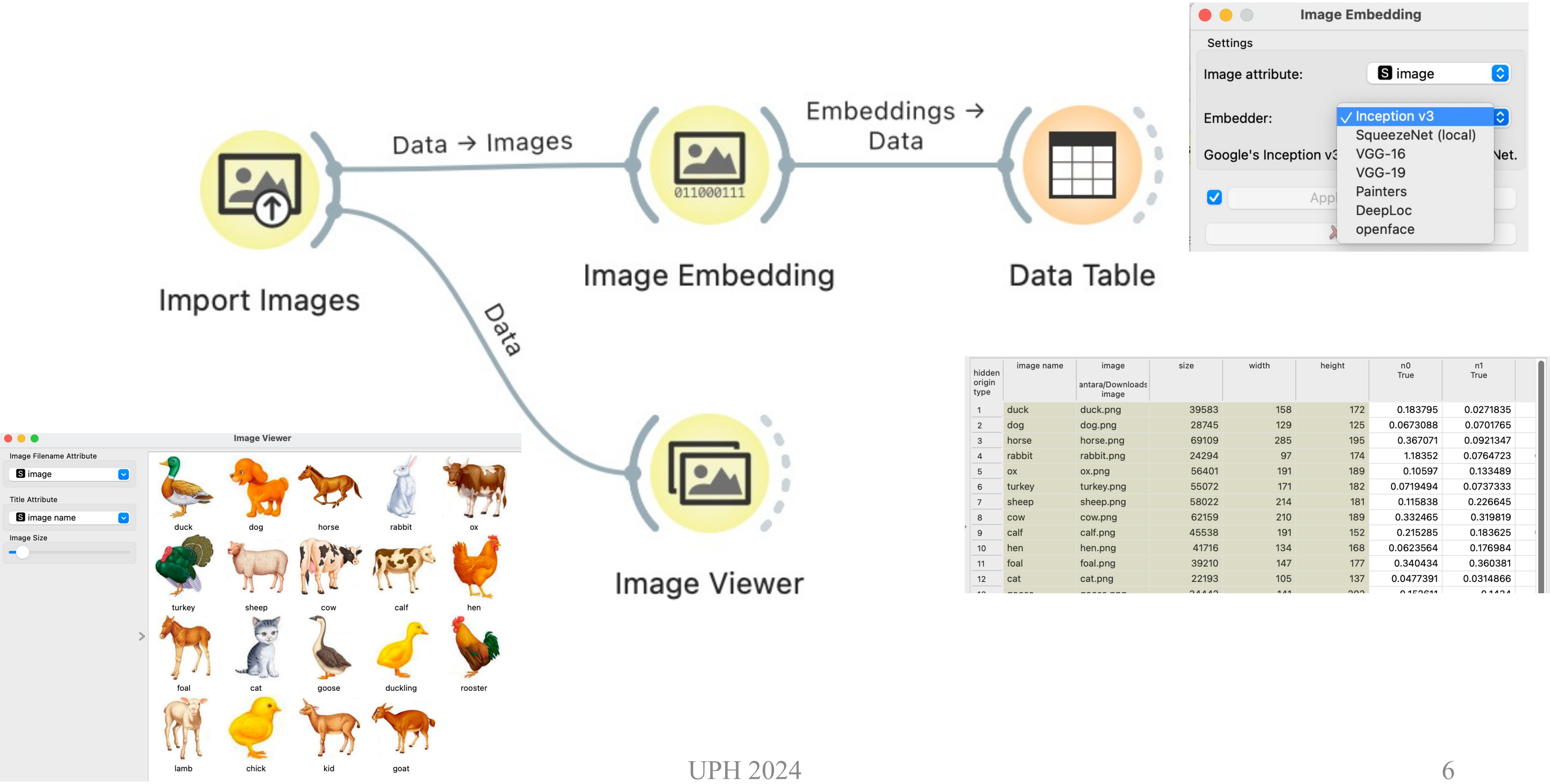


Image Embedding

Settings

Image attribute: image

Embedder: ✓ Inception v3

Google's Inception v3 … Net.

☑ Apply

Image Viewer

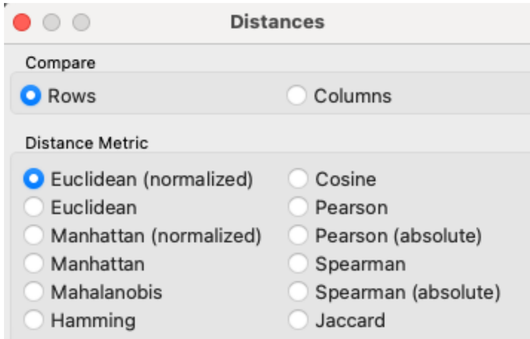
Image Filename Attribute: image

Title Attribute: image name

Image Size

hidden origin type	image name	image antara/Downloads image	size	width	height	n0 True	n1 True
1	duck	duck.png	39583	158	172	0.183795	0.0271835
2	dog	dog.png	28745	129	125	0.0673088	0.0701765
3	horse	horse.png	69109	285	195	0.367071	0.0921347
4	rabbit	rabbit.png	24294	97	174	1.18352	0.0764723
5	ox	ox.png	56401	191	189	0.10597	0.133489
6	turkey	turkey.png	55072	171	182	0.0719494	0.0737333
7	sheep	sheep.png	58022	214	181	0.115838	0.226645
8	cow	cow.png	62159	210	189	0.332465	0.319819
9	calf	calf.png	45538	191	152	0.215285	0.183625
10	hen	hen.png	41716	134	168	0.0623564	0.176984
11	foal	foal.png	39210	147	177	0.340434	0.360381
12	cat	cat.png	22193	105	137	0.0477391	0.0314866

Image Embedding



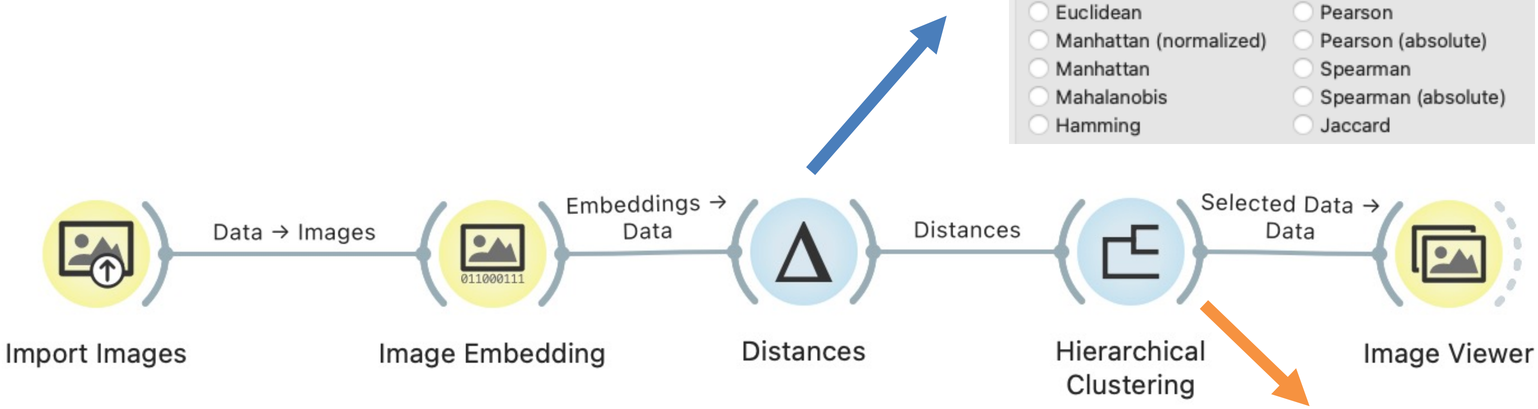
Distances

Compare

● Rows ○ Columns

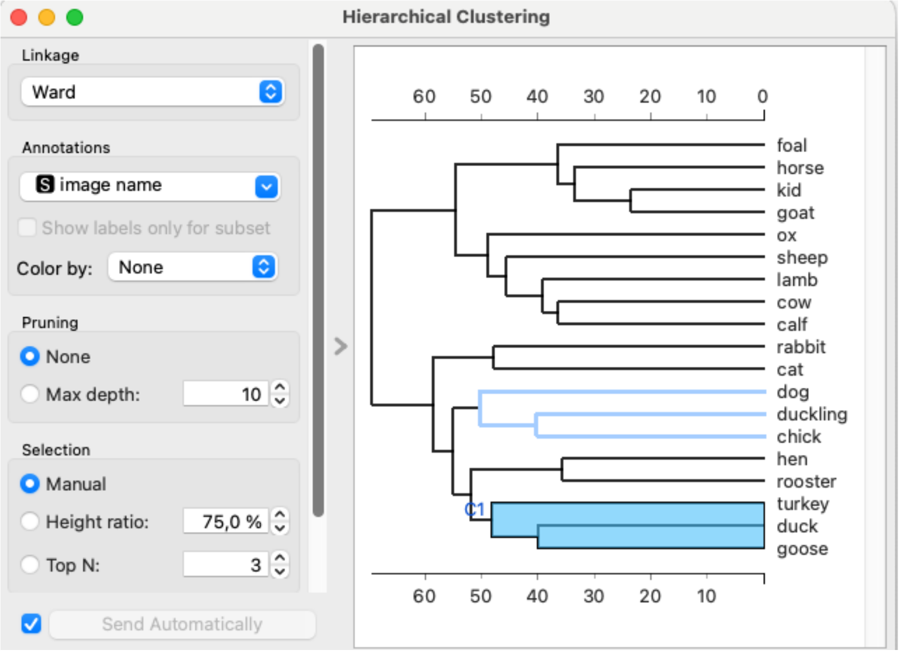
Distance Metric

● Euclidean (normalized) ○ Cosine
○ Euclidean ○ Pearson
○ Manhattan (normalized) ○ Pearson (absolute)
○ Manhattan ○ Spearman
○ Mahalanobis ○ Spearman (absolute)
○ Hamming ○ Jaccard



Import Images —Data → Images→ Image Embedding —Embeddings → Data→ Distances —Distances→ Hierarchical Clustering —Selected Data → Data→ Image Viewer

1. Load images.
2. Turned images into numbers.
3. Distances widget computes distances between rows or columns in a dataset.
4. Group items visualization
5. View group of images



Hierarchical Clustering

Linkage: Ward

Annotations: image name

□ Show labels only for subset

Color by: None

Pruning

● None

○ Max depth: 10

Selection

● Manual

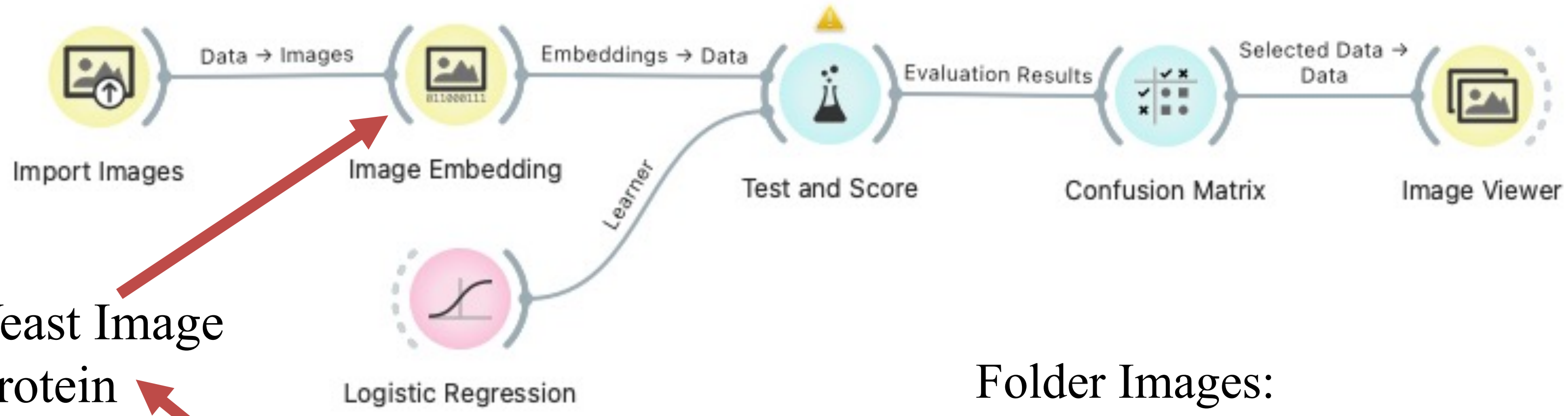
○ Height ratio: 75,0 %

○ Top N: 3

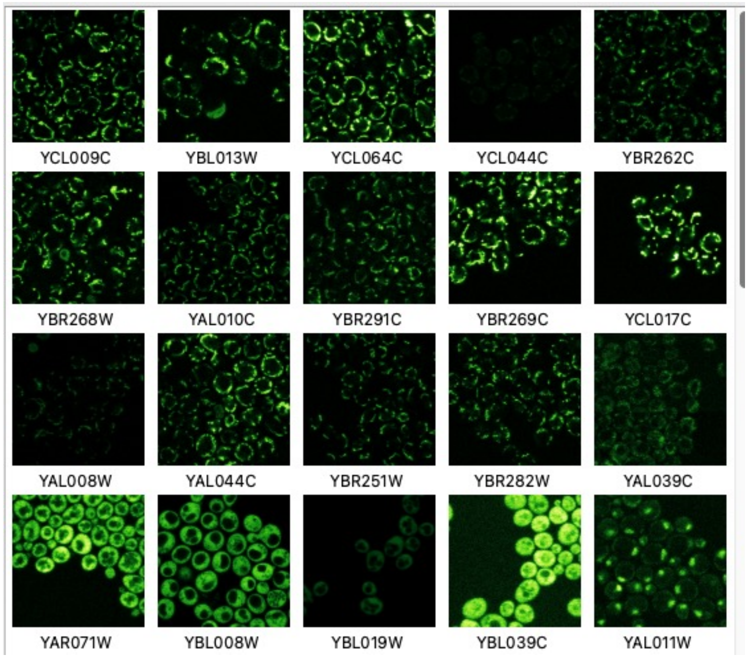
☑ Send Automatically

Images Classification

http://file.biolab.si/files/yeast-localization-small.zip



Yeast Image Protein



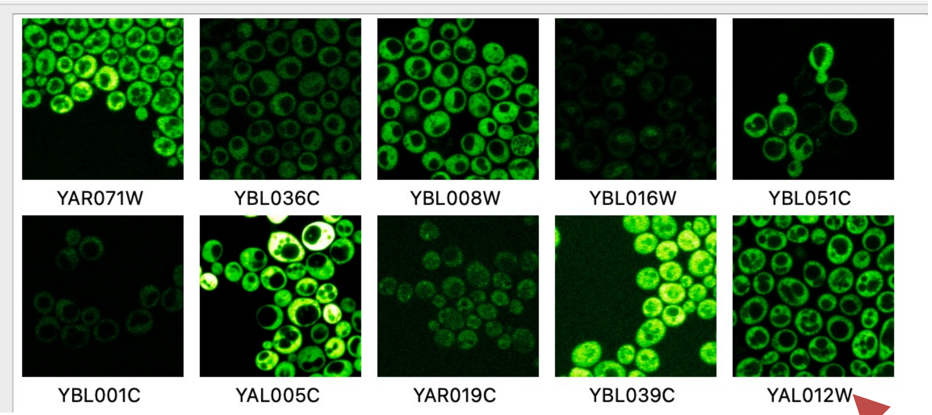
Folder Images:

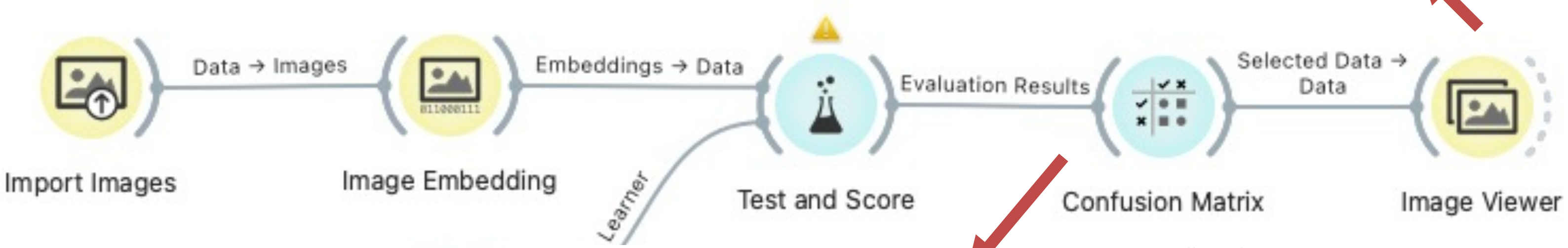
- Cytoplasm
- Endosome
- Mitochondria
- Nucleus

Steps

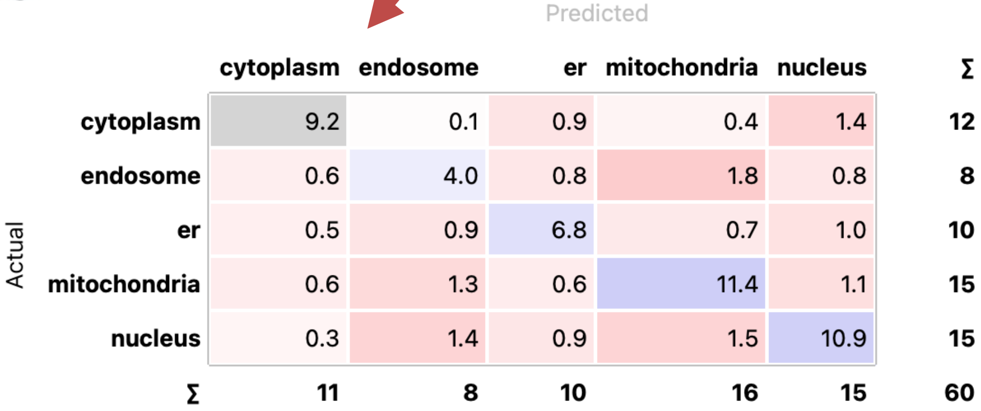
1. Import images
2. Construct matrix table from images
3. Evaluates using Machine Learning
4. Results of Testing Classification
5. View Images

Images Classification



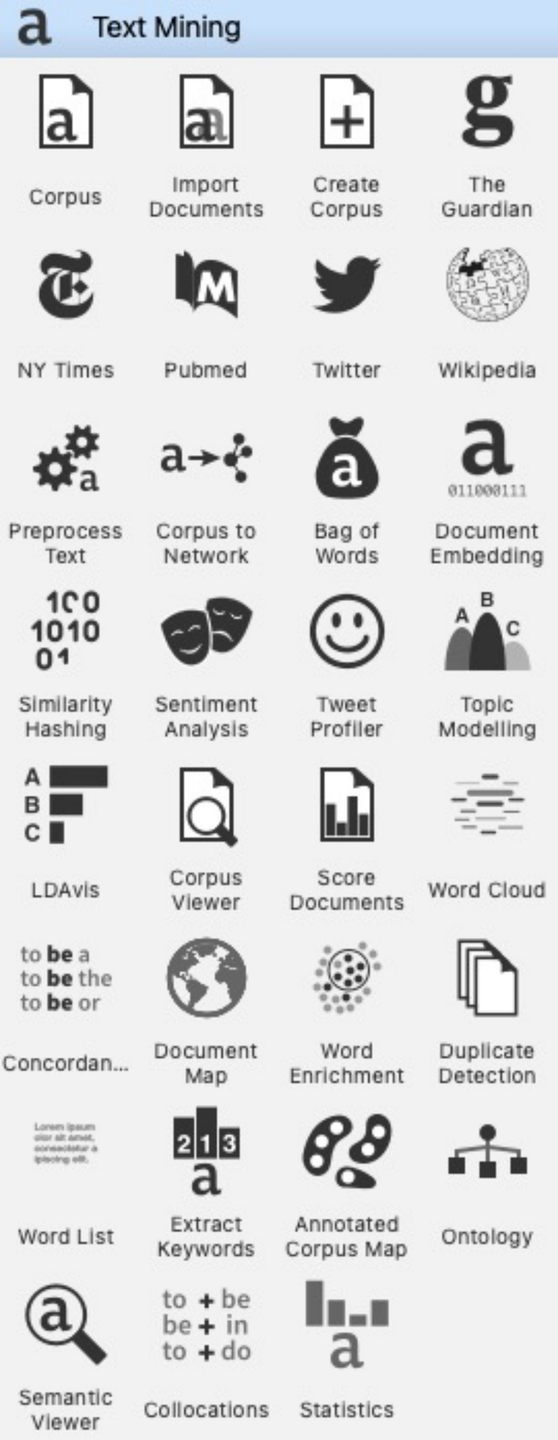


Logistic Regressior



Predicted

Actual	cytoplasm	endosome	er	mitochondria	nucleus	Σ
cytoplasm	9.2	0.1	0.9	0.4	1.4	12
endosome	0.6	4.0	0.8	1.8	0.8	8
er	0.5	0.9	6.8	0.7	1.0	10
mitochondria	0.6	1.3	0.6	11.4	1.1	15
nucleus	0.3	1.4	0.9	1.5	10.9	15
Σ	11	8	10	16	15	60



Text Mining

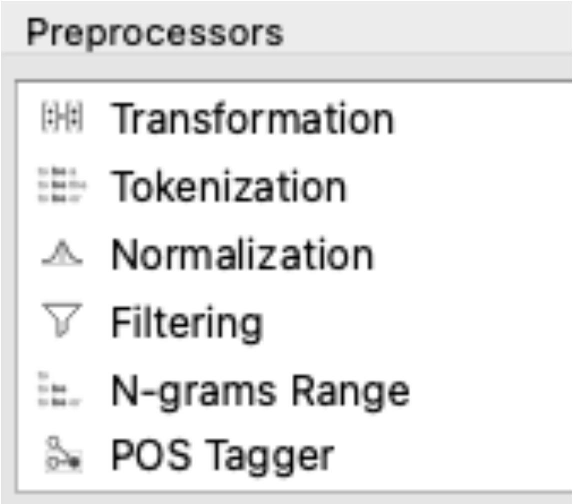
Text Mining & Text Analysis – Identifies textual patterns & trends within unstructured data through the use of machine leaning, statistics & linguistics [IBM]

Text Mining is the process of obtaining meaningful information from large collections of unstructured data using Natural Language Processing (NLP)

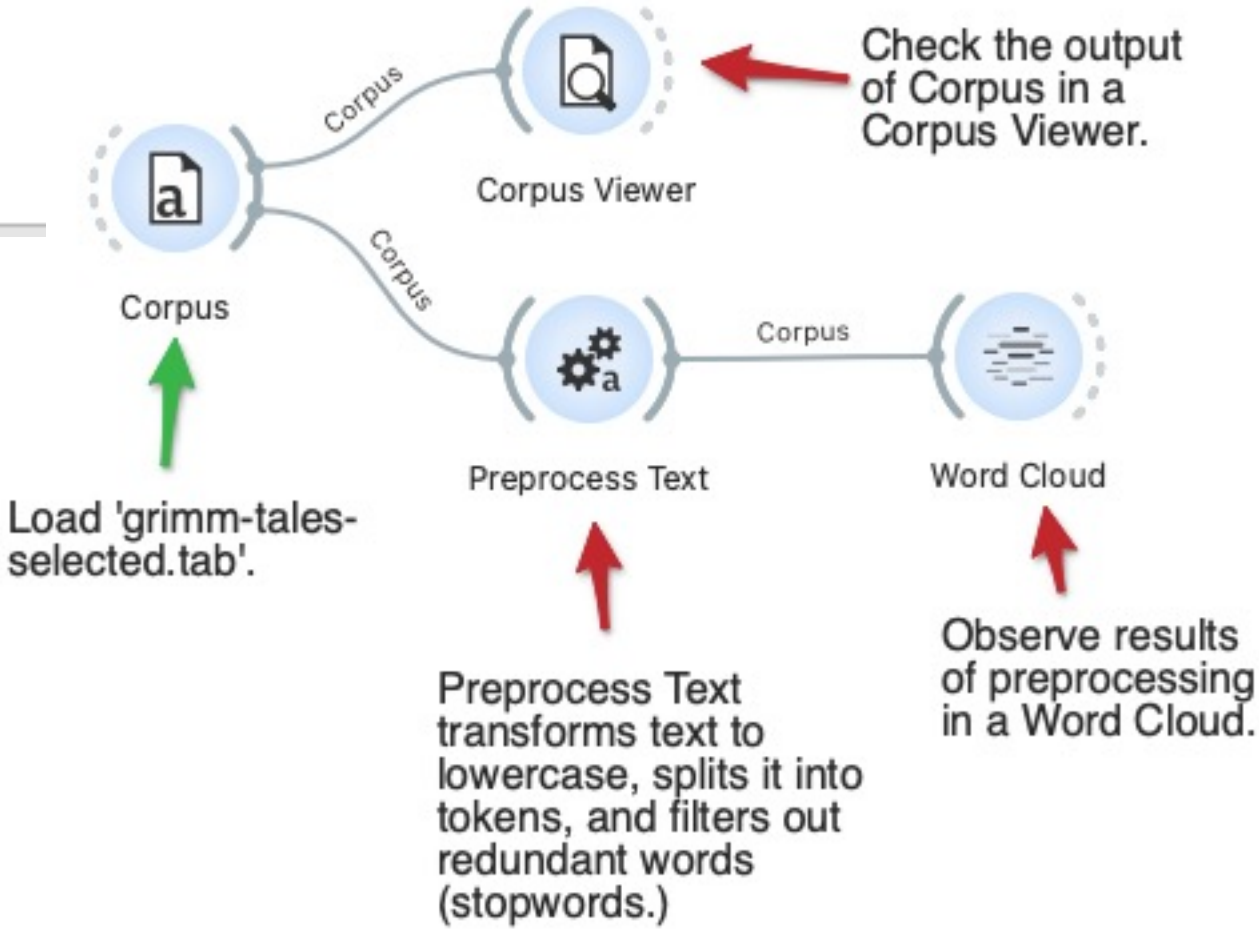
Text mining is the data mining technique or process which discovers earlier unfamiliar and valuable information from a huge quantity of unstructured text data

https://doi.org/10.1016/B978-0-323-90240-3.00026-6

Text Preprocessing


Preprocessors

- Transformation
- Tokenization
- Normalization
- Filtering
- N-grams Range
- POS Tagger



Check the output of Corpus in a Corpus Viewer.

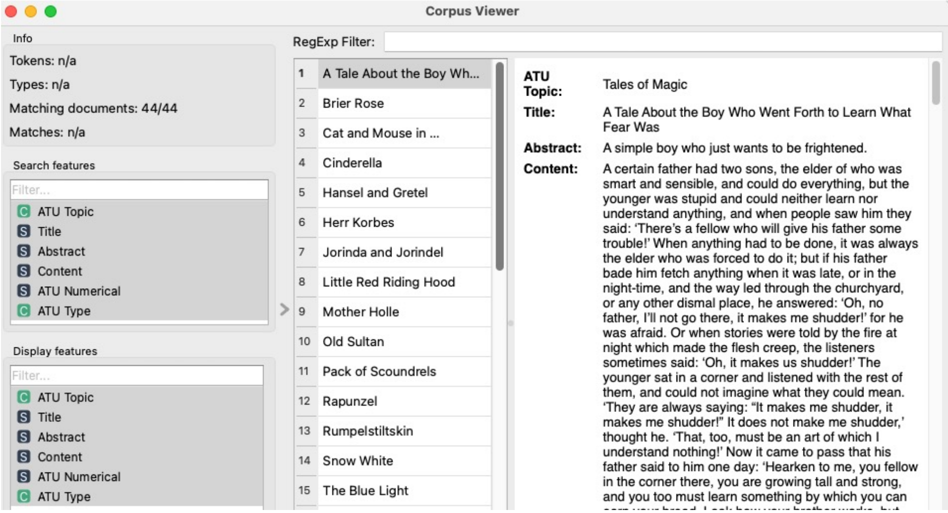
Load 'grimm-tales-selected.tab'.

Preprocess Text transforms text to lowercase, splits it into tokens, and filters out redundant words (stopwords.)

Observe results of preprocessing in a Word Cloud.

Example Workflows in Orange3

Text Preprocessing



Corpus Viewer

Info

Tokens: n/a

Types: n/a

Matching documents: 44/44

Matches: n/a

Search features

Filter...

- C ATU Topic
- S Title
- S Abstract
- S Content
- S ATU Numerical
- C ATU Type

Display features

Filter...

- C ATU Topic
- S Title
- S Abstract
- S Content
- S ATU Numerical
- C ATU Type

RegExp Filter:

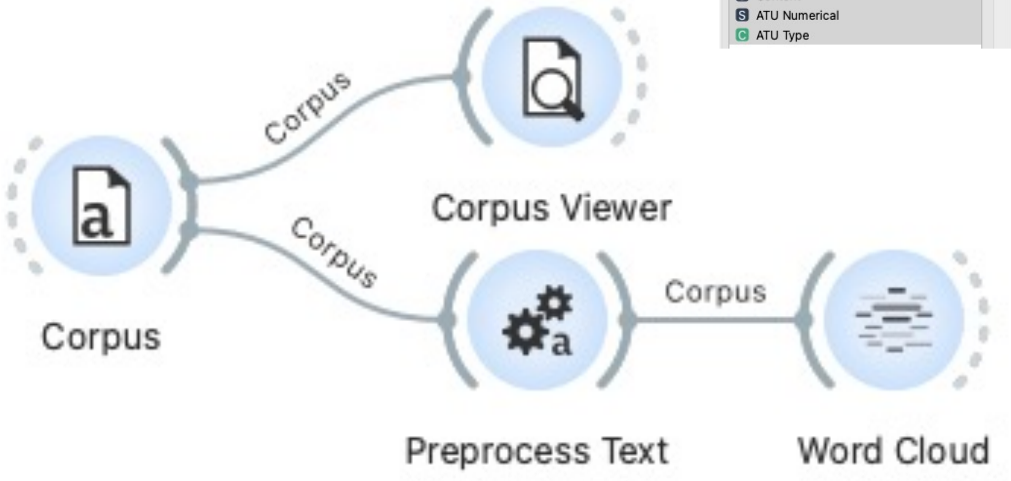
1	A Tale About the Boy Wh...
2	Brier Rose
3	Cat and Mouse in ...
4	Cinderella
5	Hansel and Gretel
6	Herr Korbes
7	Jorinda and Jorindel
8	Little Red Riding Hood
9	Mother Holle
10	Old Sultan
11	Pack of Scoundrels
12	Rapunzel
13	Rumpelstiltskin
14	Snow White
15	The Blue Light

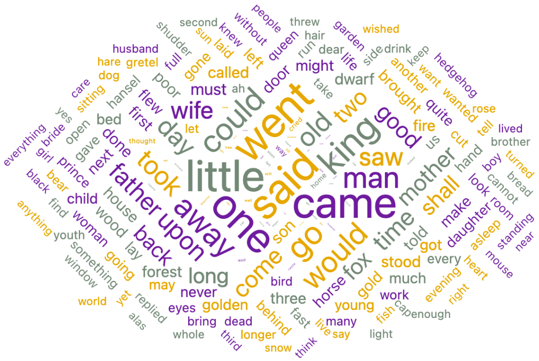
ATU Topic: Tales of Magic

Title: A Tale About the Boy Who Went Forth to Learn What Fear Was

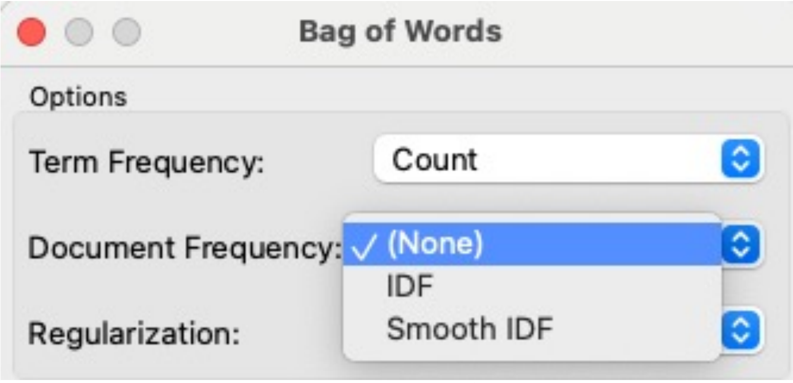
Abstract: A simple boy who just wants to be frightened.

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





Hierarchical Clustering



Bag of Words

Options

Term Frequency: Count

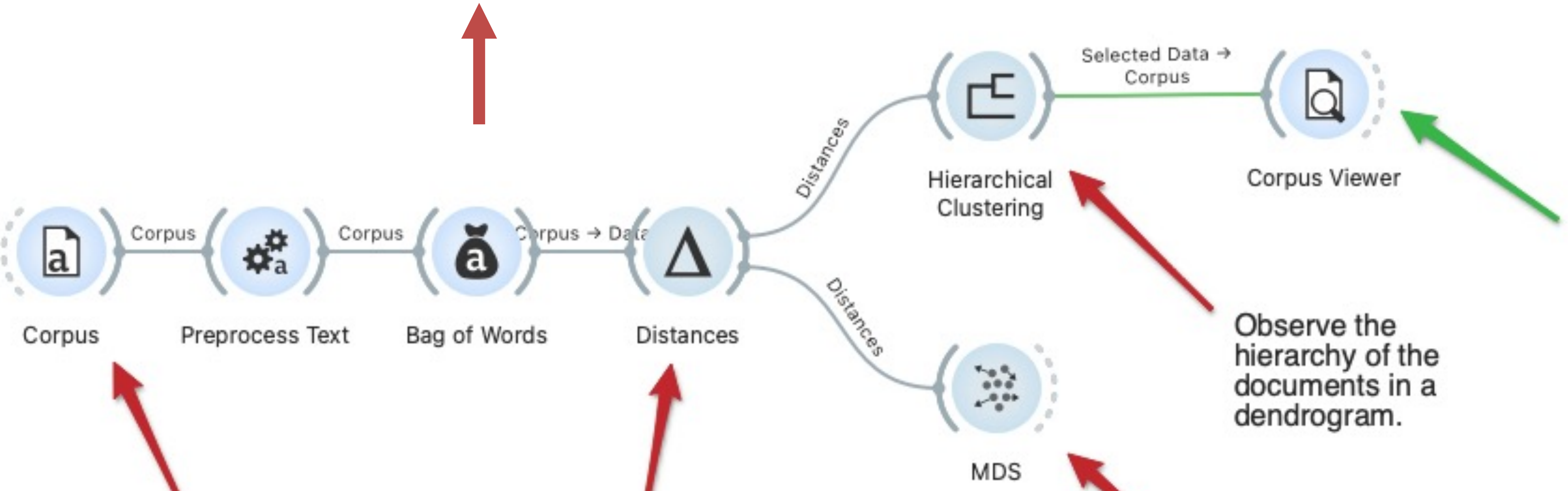
Document Frequency: ✓ (None)

IDF

Smooth IDF

Regularization:

Bag of Words model creates a corpus with word counts for each data instance (document). The count can be either absolute, binary (contains or does not contain) or sublinear (logarithm of the term frequency). Bag of words model is required in combination with Word Enrichment and could be used for predictive modelling.



Select a cluster in Hierarchical Clustering and observe selected documents in Corpus Viewer.

Observe the hierarchy of the documents in a dendrogram.

Load 'grimm-tales-selected.tab'.

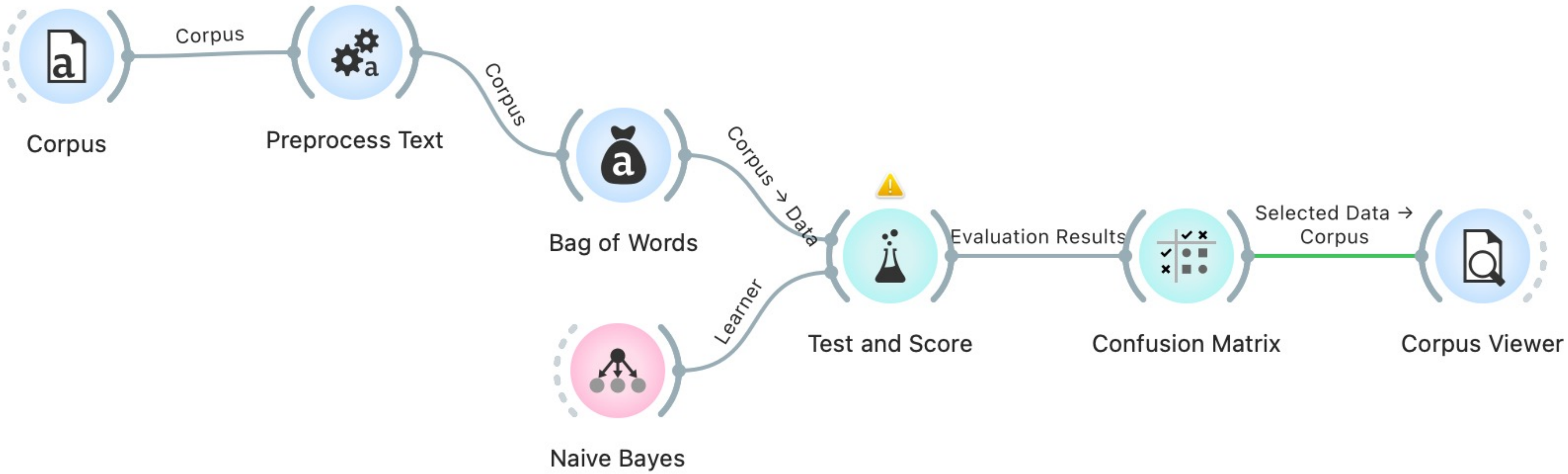
Compute distances between documents with cosine distance. The smaller the distance, the more similar two document are.

Plot the data in 2D. MDS projects multidimensional data, such as BoW, to a plane where similar documents lie closer together.

Example Orange3

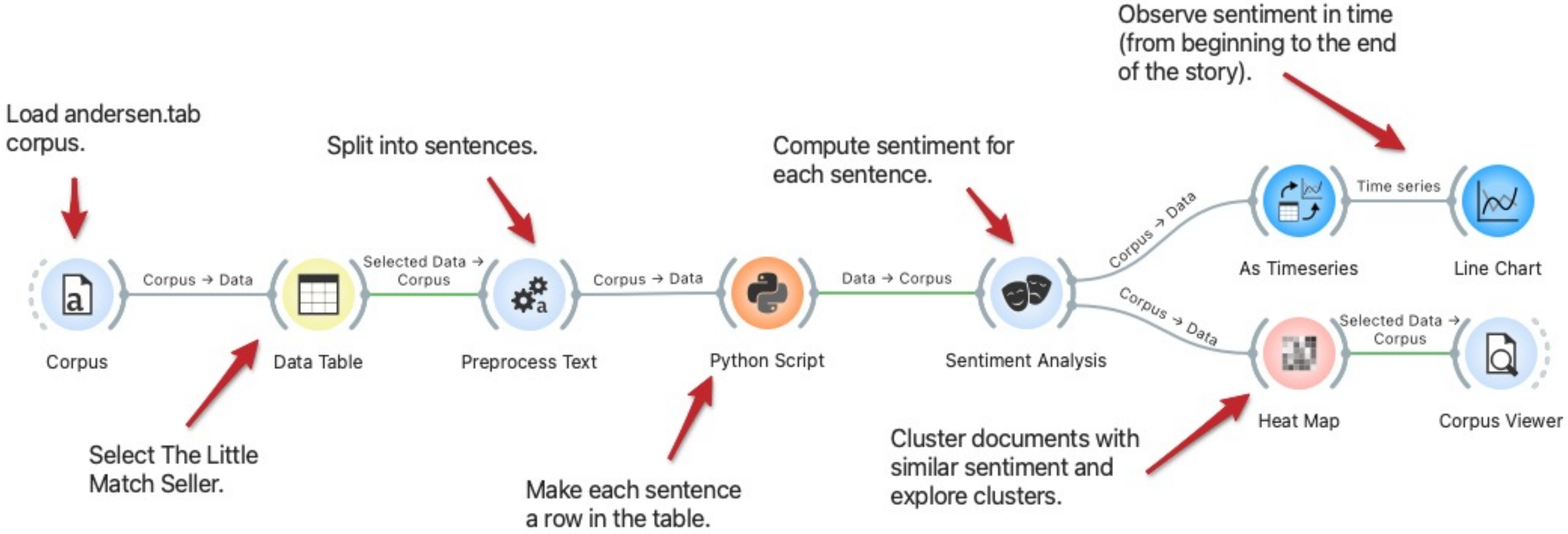
Text Mining

- Analytics in Text Mining



Story Arcs

1. Select the story from the corpus of Andersen tales.
2. Create a table where each sentences of the tale into a separate row.
3. Sentiment analysis to compute the sentiment of each sentence, then we observe the emotional arcs through the story
4. Observe sentences with similar scores in the Heat Map and Corpus Viewer


Load andersen.tab corpus.

Split into sentences.

Compute sentiment for each sentence.

Observe sentiment in time (from beginning to the end of the story).

Select The Little Match Seller.

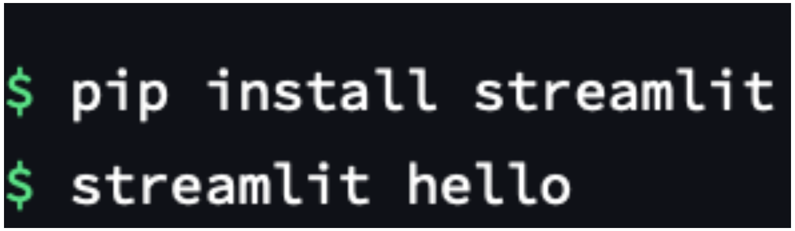
Make each sentence a row in the table.

Cluster documents with similar sentiment and explore clusters.

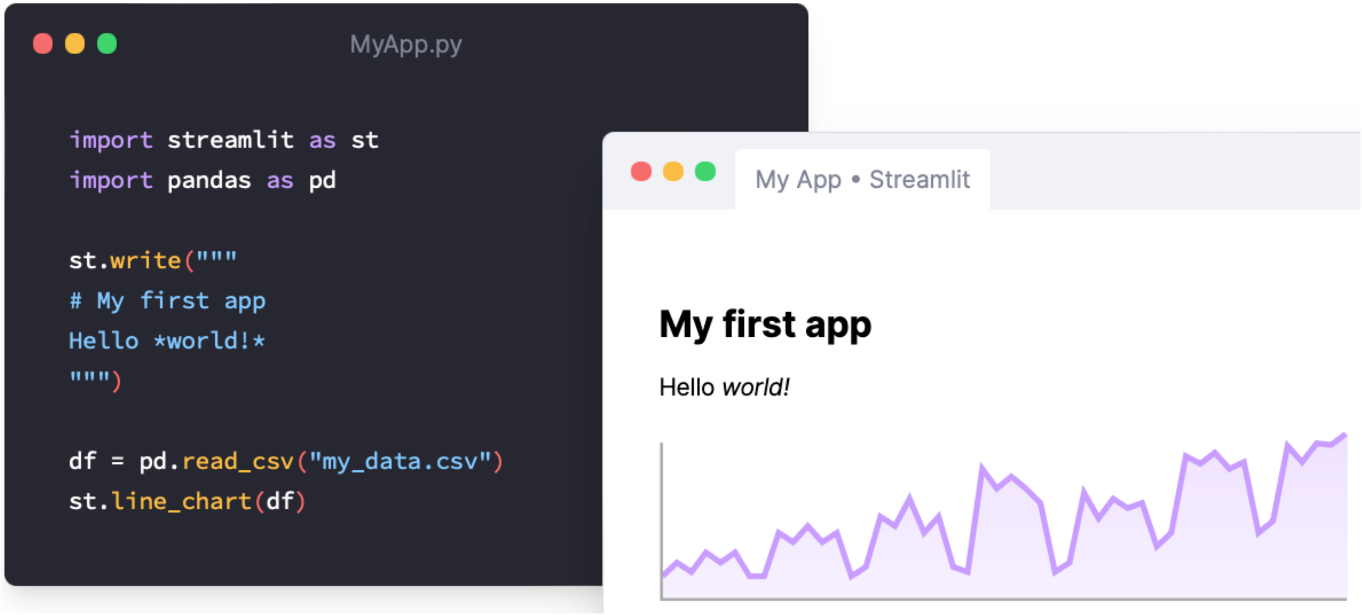
https://orangedatamining.com/examples/?tag=Text+Mining

Deployment

- Web Services
- Python Application
- Open Source



```
$ pip install streamlit
$ streamlit hello
```

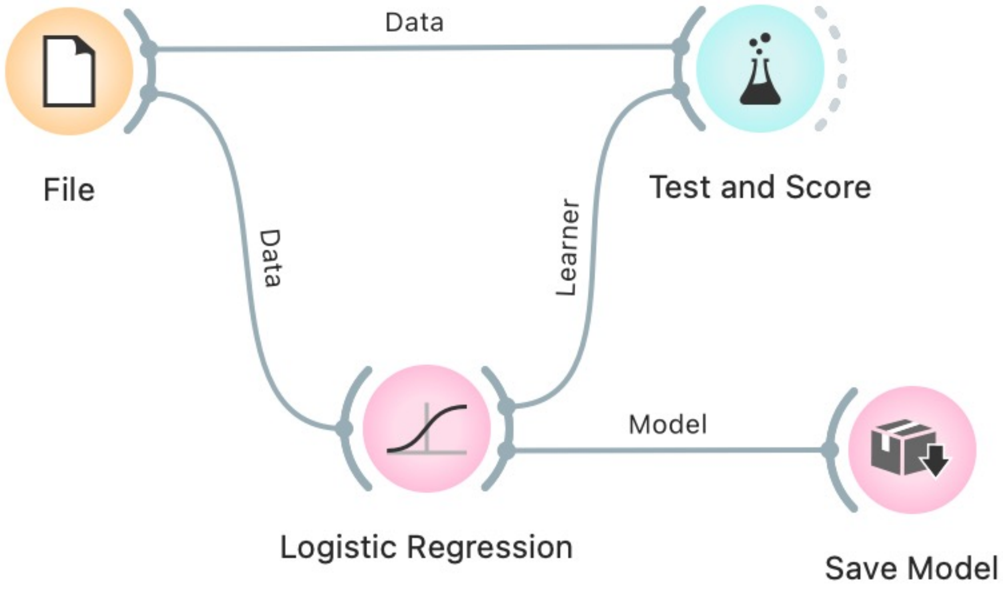


```
import streamlit as st
import pandas as pd

st.write("""
# My first app
Hello *world!*
""")

df = pd.read_csv("my_data.csv")
st.line_chart(df)
```

My first app

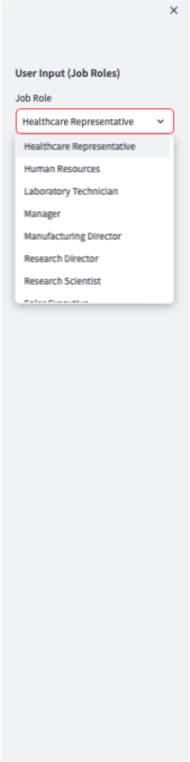
Hello *world!*

Deployment


Data
File
Test and Score
Data
Learner
Model
Logistic Regression
Save Model

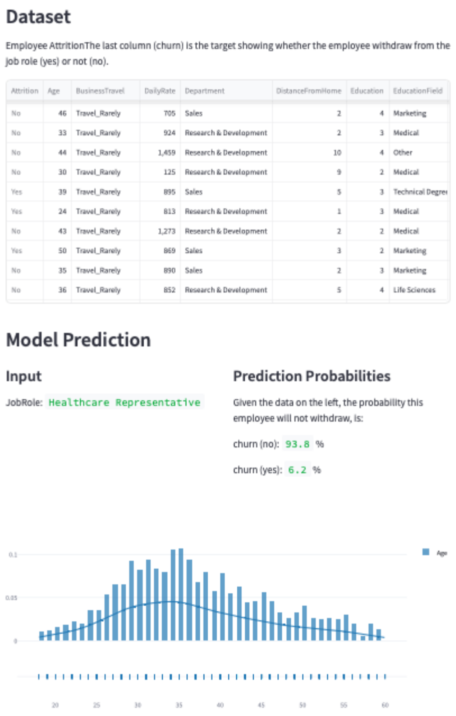
Deployment

1. Create folder & Copy source files
2. Run ”streamlit run files.py”
3. Type ctrl+c to stop



User Input (Job Roles)

Job Role: Healthcare Representative



Dataset

Employee AttritionThe last column (churn) is the target showing whether the employee withdraw from the job role (yes) or not (no).

Attrition	Age	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField
No	46	Travel_Rarely	705	Sales	2	4	Marketing
No	33	Travel_Rarely	924	Research & Development	2	3	Medical
No	44	Travel_Rarely	1,459	Research & Development	10	4	Other
No	30	Travel_Rarely	125	Research & Development	9	2	Medical
Yes	39	Travel_Rarely	895	Sales	5	3	Technical Degre
Yes	24	Travel_Rarely	813	Research & Development	1	3	Medical
No	43	Travel_Rarely	1,273	Research & Development	2	2	Medical
Yes	50	Travel_Rarely	869	Sales	3	2	Marketing
No	35	Travel_Rarely	890	Sales	2	3	Marketing
No	36	Travel_Rarely	852	Research & Development	5	4	Life Sciences

Model Prediction

Input

JobRole: `Healthcare Representative`

Prediction Probabilities

Given the data on the left, the probability this employee will not withdraw, is:

churn (no): `93.8` %

churn (yes): `6.2` %

Any Question ?

- https://orangedatamining.com/examples/
- Streamlit.io