Visualização de Informação com a Ferramenta Orange

Como processar, visualizar e interagir com dados na ferramenta de ciência de dados Orange ©

Eric Macedo Cabral
cabral.eric@usp.br

29 / 09 / 2020

Sobre os Desenvolvedores

- University of Ljubljana (https://fri.uni-lj.si/en)
  - Biolab (https://github.com/biolab)
- Orange (https://orange.biolab.si/)

University *of Ljubljana*
Faculty *of Computer and Information Science*

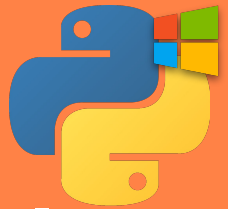
Sobre a Ferramenta

- Ferramenta Open source de Aprendizado de Máquina e Visualização de Dados
- Indicado para todos os níveis de conhecimento nas áreas anteriormente citadas e áreas derivadas
- Fluxos de trabalho (Pipelines, workflows)
- Interativo
- Extensível

Instalação - Linux

Python 3 ou superior

```
sudo apt install virtualenv build-essential python3-dev python3-pyqt5.*
virtualenv --python=python3 --system-site-packages orange3venv
source orange3venv/bin/activate
pip install orange3
python3 -m Orange.canvas
```



Instalação - Windows

Não necessita de permissão de administrador

Anaconda
Download

```
conda create python=3 --name orange3
conda config --add channels conda-forge
conda install orange3
conda install -c defaults pyqt=5 qt
conda install orange3-<addon name>
```

Standalone
Installer
Download

Roteiro

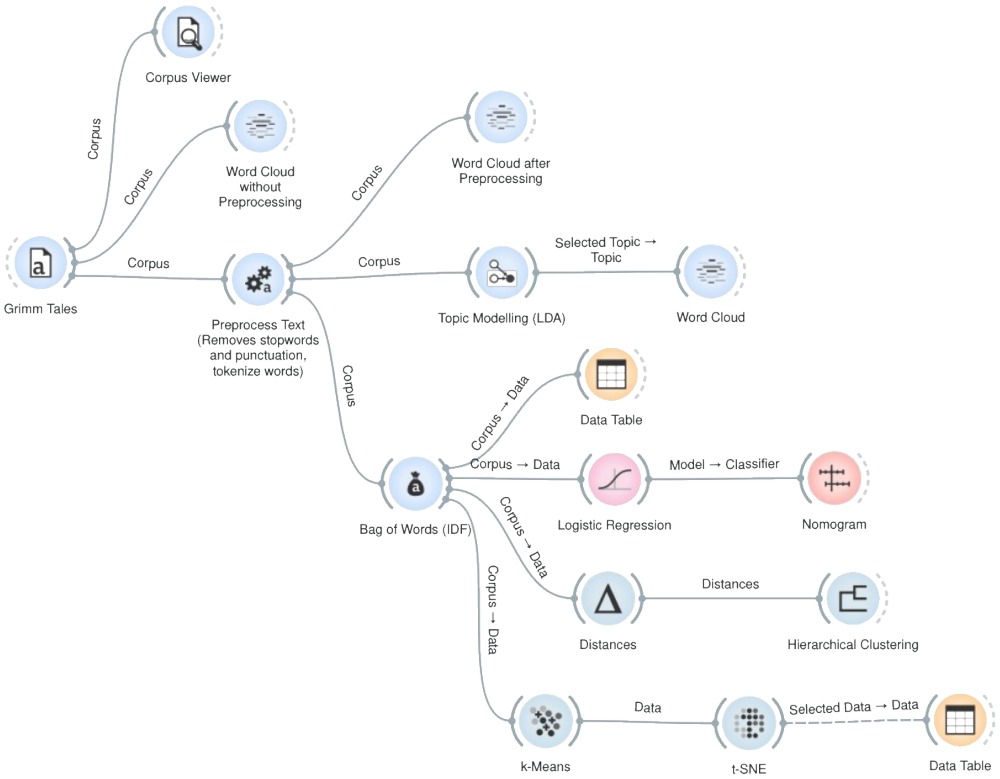
1. Conceitos Básicos
2. Pipelines
3. Estendendo a Ferramenta
4. Exercício

1. Conceitos Básicos

1. Canvas
2. Widgets
3. Workflows

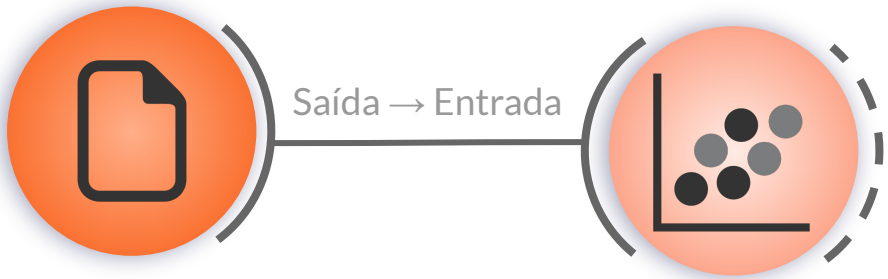
Canvas

- Quadro onde o usuário pode adicionar um fluxo de execução
- Mapa conceitual



Widgets

- Unidades computacionais
- Múltiplos propósitos
  - Ler dados
  - Visualizar
  - Processar dados
  - Script
  - Evento
- Entrada e saída



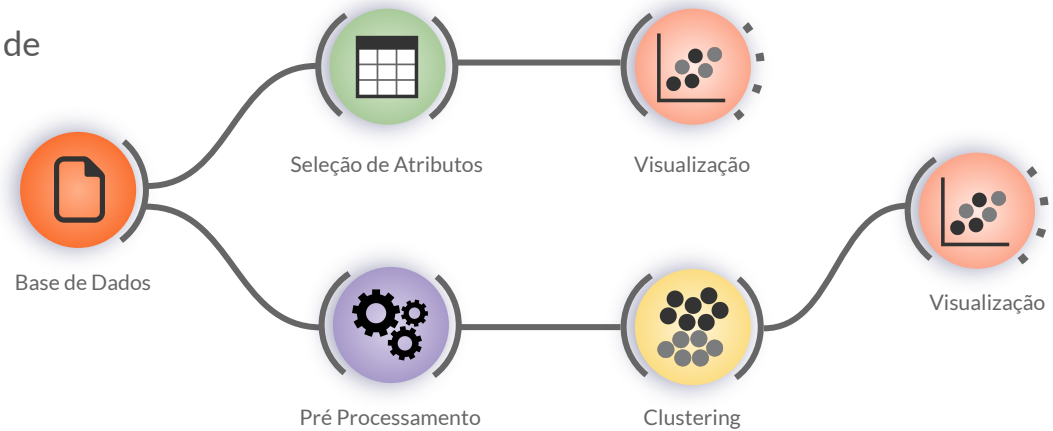
Exemplo 1: Interação com algoritmos de clustering

Arquivo 01_edu.ows

Fazer download do add-on “Educational”

Workflows

- Define uma ou mais sequências de operações e recursos
- Reutilizável
- Intuitivo



Exemplo 2: Interações com o conjunto de dados Iris

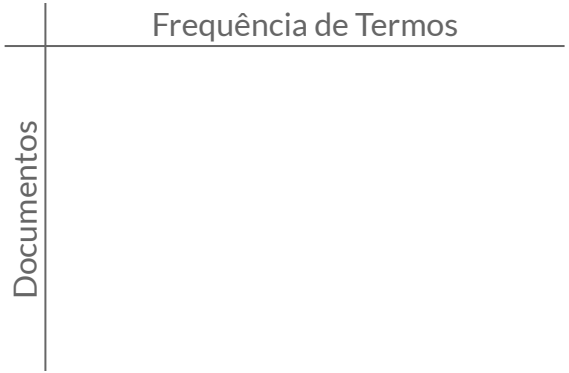
Arquivo 02_iris.ows

2. Pipelines

1. Texto
2. Imagens

Pipeline de Texto I

- Corpus
- Pré-processamento
  - Minúsculo
  - Remoção de Stopwords e caracteres não alfanuméricos
  - Lematização (Stemming) e Tokenização
- Processamento de texto
  - Bag of Words (Frequência de termos)
  - **T**erm **F**requency - **I**nverse **D**ocument **F**requency (**TF-IDF**)



Pipeline de Texto II

- Clustering
  - K-Means (K-Médias)
  - Clustering Hierárquico (Dendograma)
- Redução de dimensionalidade
  - **P**rincipal **C**omponent **A**nalysis (**PCA**)
  - **t**-Distributed **S**tochastic **N**eighbor **E**mbedding (**t-SNE**)

Pipeline de Texto III

- Modelagem de Tópicos
  - **L**atent **D**irichlet **A**llocation (**LDA**)

Exemplo 3: Análise de texto dos Contos dos Irmãos Grimm com TF-IDF e modelagem de tópicos

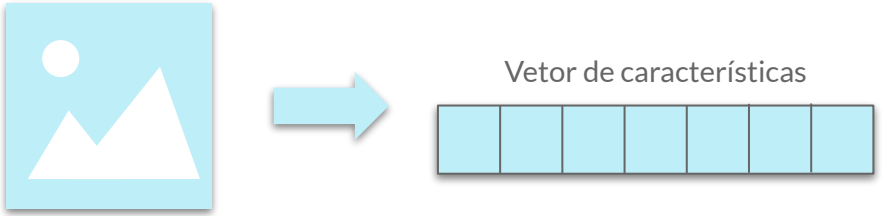
Arquivo 03_text.ows

Fazer download do add-on “Text”

O módulo *WordCloud* está apresentando problemas na versão 3.26 do Orange

Pipeline de Imagem I

- Redes neurais
- Descritores de imagens
- Processamento remoto



Pipeline de Imagem II

- K-Means
  - Distância do cosseno entre os vetores de características
- Redução de dimensionalidade
  - t-SNE

Exemplo 4: Classificação de imagens de animais

Arquivo 04_image.ows

Fazer download do add-on: “Image Analytics”
Fazer download do arquivo 10-animals.zip

Este workflow é particularmente pesado, então deve ser carregado com antecedência

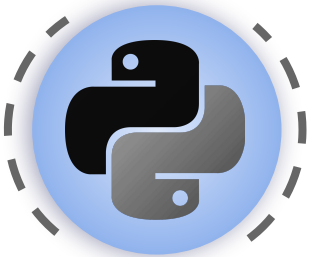
3. Estendendo a Ferramenta

1. Scripting

Scripting

- Widget "*Python Script*"
- Orange.data.Table
  - X
  - Y
  - W
  - Metas
- Orange.data.Domain
  - Attributes
  - Class_var,
  - Metas
  - Source

in_data, in_datas,
in_learner, in_learners,
in_classifier, in_classifiers,
in_object, in_objects



out_data,
out_learner,
out_classifier,
out_object

Exemplo 5: Manipulando dados do Human Development Report (HDR)

Arquivo 05_script.ows

4. Exercício

Exercício: Análise de texto em *NewsGroups*

1. Base de dados: 20NewsGroup-Train
2. Selecione os grupos:
   a. talk.politics.guns
   b. comp.sys.ibm.pc.hardware
   c. sci.space
   d. rec.motorcycles
3. Utilize os widgets que forem necessários para filtrar, pré-processar e demais tarefas
   a. Utilize o modelo de representação de documentos que desejar (Embeddings ou Bag-of-Words)
4. Encontre clusters de documentos (p. e. K-Means)
5. Faça uma projeção dos clusters encontrados (p. e. t-SNE)
6. Faça uma análise dos dados a partir da visualização e modifique seu pipeline se necessário

Referências e Links Úteis

- Documentação da ferramenta: https://orange.biolab.si/docs/
- Série de vídeos de introdução (Oficial): Playlist no YouTube
- Scikit-Learn: https://scikit-learn.org/
    - A biblioteca Scikit-Learn é amplamente utilizada na ferramenta Orange
    - Possui uma boa documentação e conteúdo teórico em Machine Learning e Data Science

Obrigado!

Dúvidas?

Contato: cabral.eric@usp.br