

Visualização de Informação com a Ferramenta Orange

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

22 / 08 / 2019

Eric Macedo Cabral
cabral.eric@usp.br

Sobre os Desenvolvedores

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

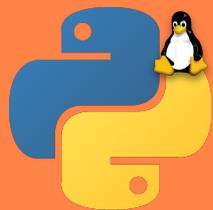


University of Ljubljana
Faculty of Computer and
Information Science

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





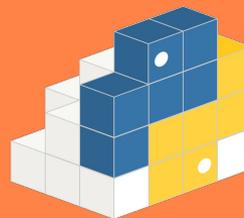
Instalação - Linux

Python 3 ou superior



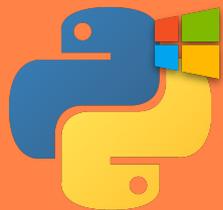
Ubuntu

```
sudo apt install python3-pip
sudo apt install python3-pyqt5.*
```



Pip

```
pip3 install orange3 --user
Execute:
python3 -m Orange.canvas ou
orange-canvas
```



Instalação - Windows

Não necessita de permissão de administrador



Anaconda
[Download](#)

```
conda config --add channels conda-forge
conda install orange3
conda install -c defaults pyqt=5 qt
```

Standalone
Installer
[Download](#)





Roteiro

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

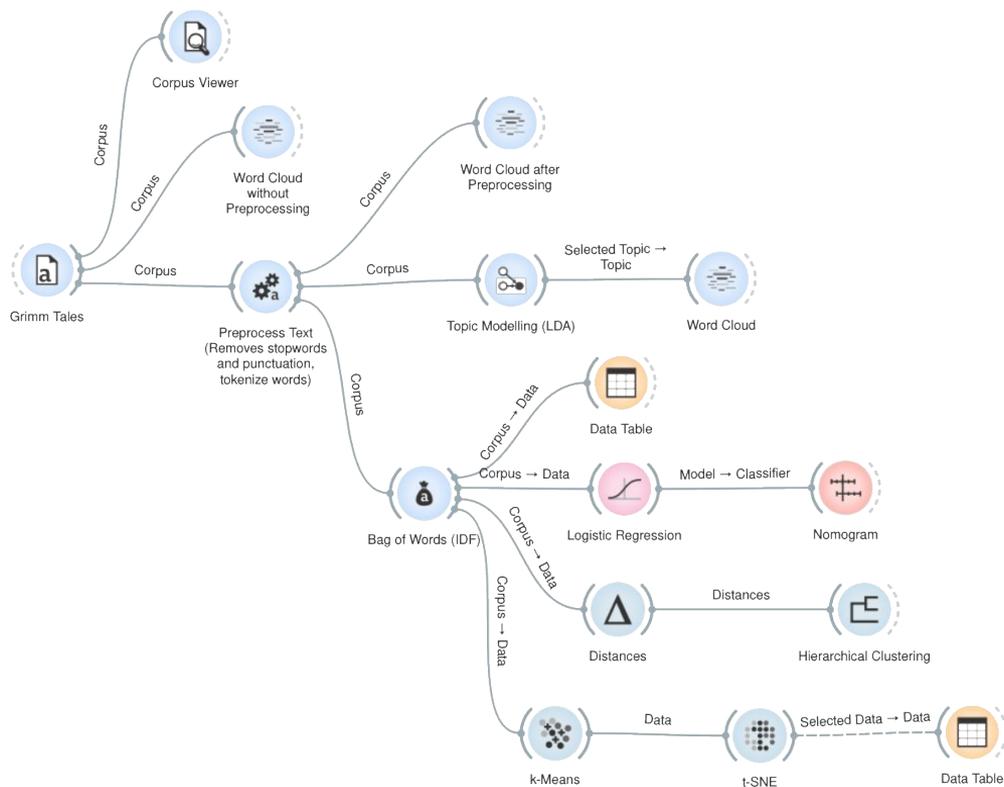


1. Conceitos Básicos

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

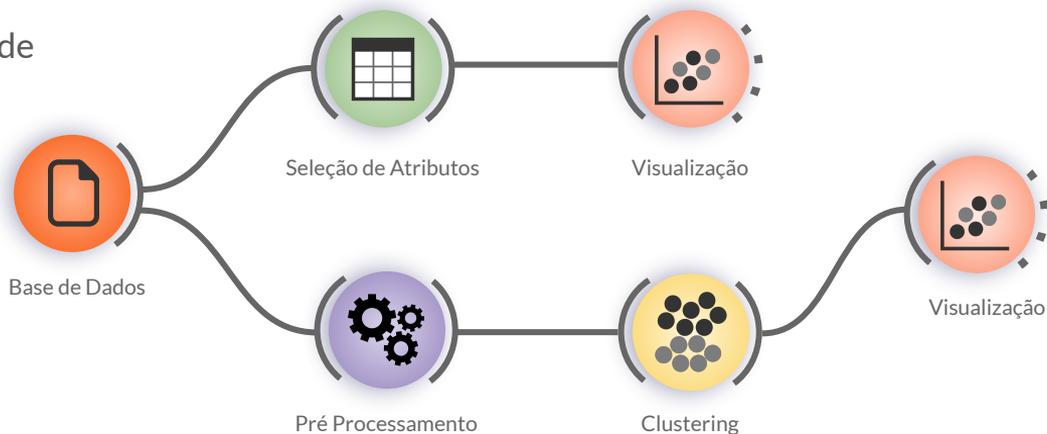
Arquivo 01_edu.ows

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



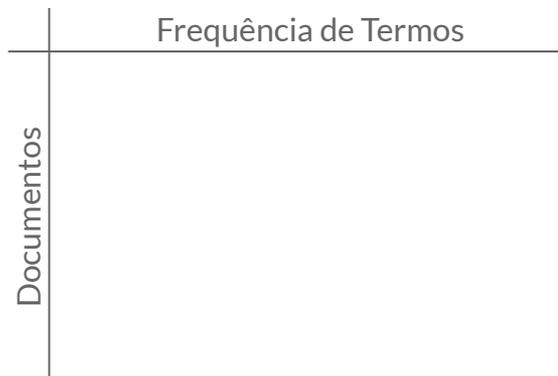
A decorative horizontal bar consisting of a dark blue segment on the left and an orange segment on the right.

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)
 - Term Frequency - Inverse Document Frequency (TF-IDF)





Pipeline de Texto II

- Clustering
 - K-Means (K-Médias)
 - Clustering Hierárquico (Dendograma)
- Redução de dimensionalidade
 - Principal Component Analysis (PCA)
 - t-Distributed Stochastic Neighbor Embedding (t-SNE)



Pipeline de Texto III

- Modelagem de Tópicos
 - Latent Dirichlet Allocation (LDA)

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

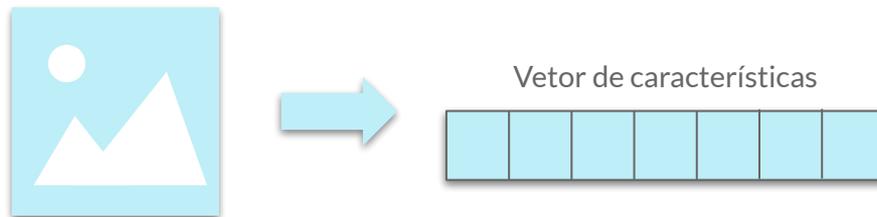
Arquivo `03_text.ows`

Fazer download do add-on "Text"



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"





3. Estendendo a Ferramenta

1. Scripting

Scripting

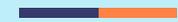
- Widget “Python Script”
- Orange.data.Table
 - X
 - Y
 - W
 - Metas
- Orange.data.Domain
 - Attributes
 - Class_var,
 - Metas
 - Source



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
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