Instituto Superior de Engenharia de Coimbra

DEPARTAMENTO DE ENG. INFORMÁTICA E SISTEMAS
LICENCIATURA EM ENGENHARIA INFORMÁTICA

INTELIGÊNCIA COMPUTACIONAL 23/24

AULA LABORATORIAL N.º 5

TEXT-MINING PARA CLASSIFICAÇÃO DE DOCUMENTOS

1. Objectivos

Pretende-se analisar diferentes técnicas de pré-processamento e classificação de documentos usando a ferramenta "Orange".
O Orange é uma ferramenta open-source para "data mining", incorporando as principais técnicas de análise de dados, visualização, classificação, regressão clustering e avaliação.

A biblioteca de text mining contém ferramentas específicas para o pré-processamento de texto, tais como construção de "tokens", remoção de "stop words" ou algoritmos de "stemming". A Figura 1 apresenta o "pipeline" associado ao processo de texto-mining.

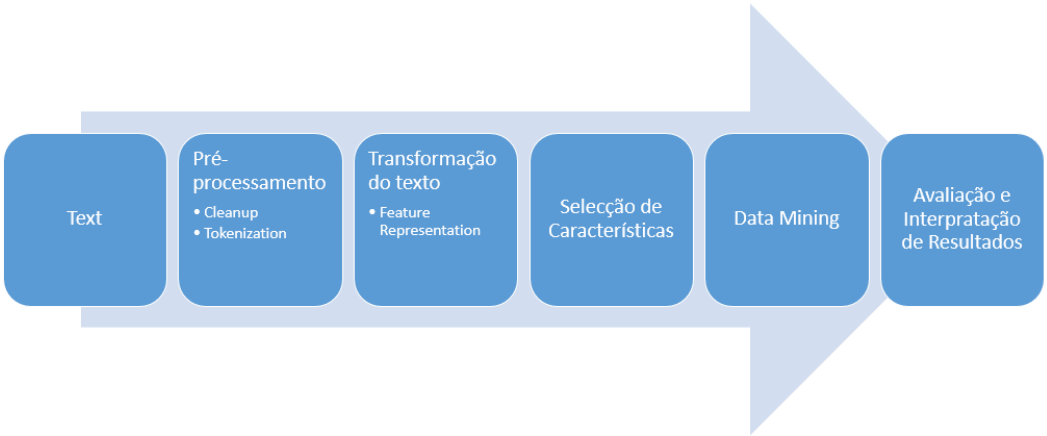


Figura 1a – Processo de Text-Mining.

A representação para um conjunto de documentos consiste usualmente em três etapas [1]:

- ***Tokenization*** - Dividir cada documento em "tokens" – usualmente palavras.

- ***Construção* de vocabulário** – formar um vocabulário com todas as palavras que aparecem no conjunto, ordenado por exemplo alfabeticamente – processo designado de "bag of words".
- ***Codificação*** - Para cada documento, determinar a frequência com que as palavras do vocabulário aparecem no referido documento e no corpus (conjunto de todos os documentos).

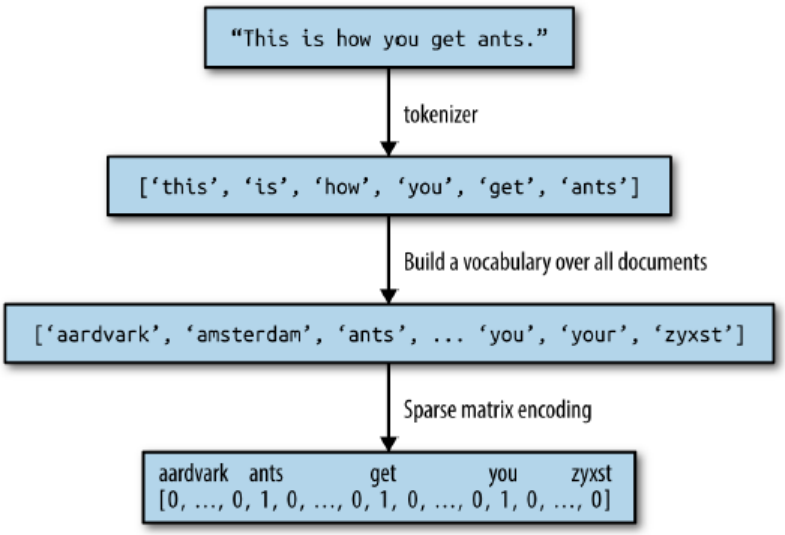


Figura 1b – Processo de Text-Mining.

2. Orange

O Orange [2] disponibiliza os mais importantes algoritmos para pré-processameto de dados, aprendizagem automática (machine learning), seleccção de características (feature selection) e avaliação de desempenho (performance evaluation). Os algoritmos são representados por "widgets" e organizam-se num "workflow" que representa a sequência de operações (processo) e o fluxo dos dados (dataset).

- A partir do navegador do Anaconda instale a ferramenta Orange3:

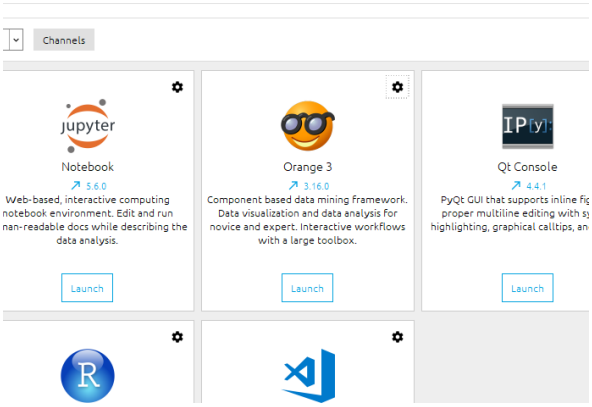


Figura 1c – Anaconda.

- Execute o Tutorial "online" que pode ser invocado a partir da opção "Help". Também poderá consultar o manual em formato .pdf que descreve todos os algoritmos e interface gráfico.

3. Exercício – Iris Dataset

Como primeiro exemplo de criação de um processo, na aplicação Orange, crie um workflow para analisar o **dataset "iris":**

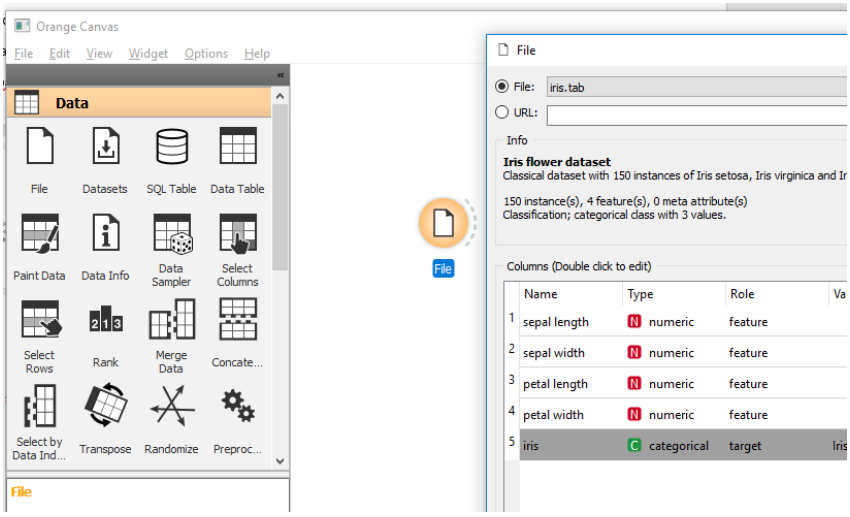


Figura 2 – Definição do dataset com o widget "File".

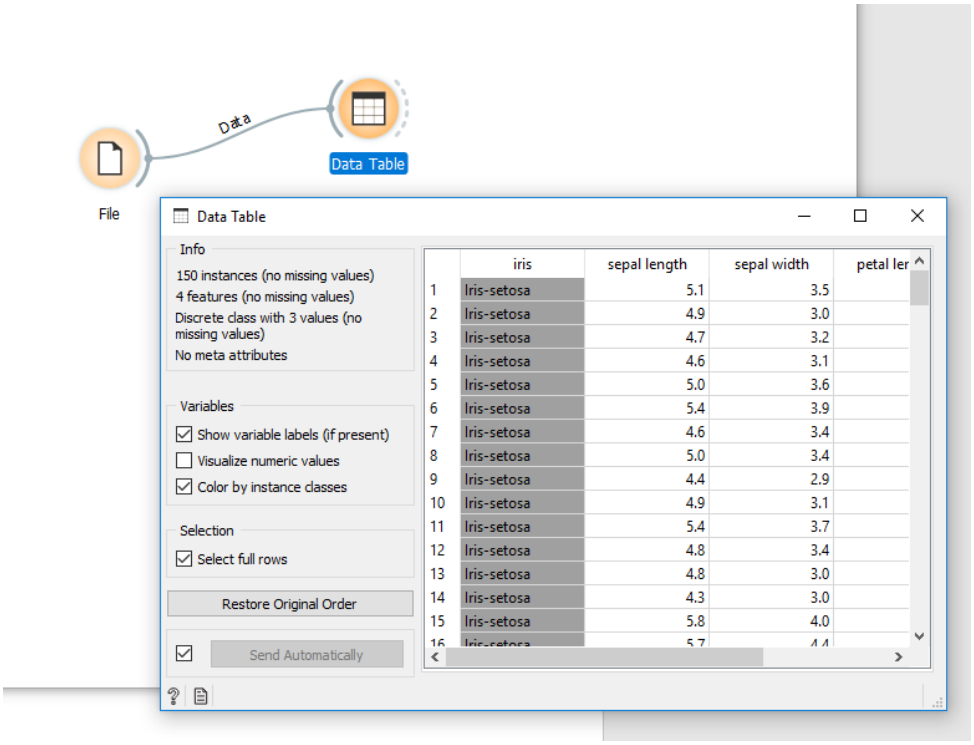


Figura 3 – Visualização do dataset com o widget "Data Table"

- Visualize os dados com "scatter plot" e procure as características mais determinantes para classificar as três classes.

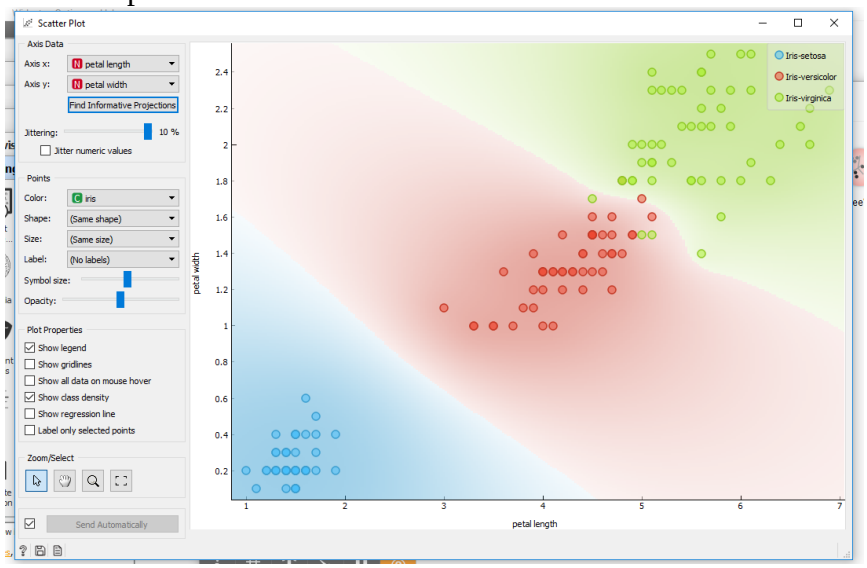
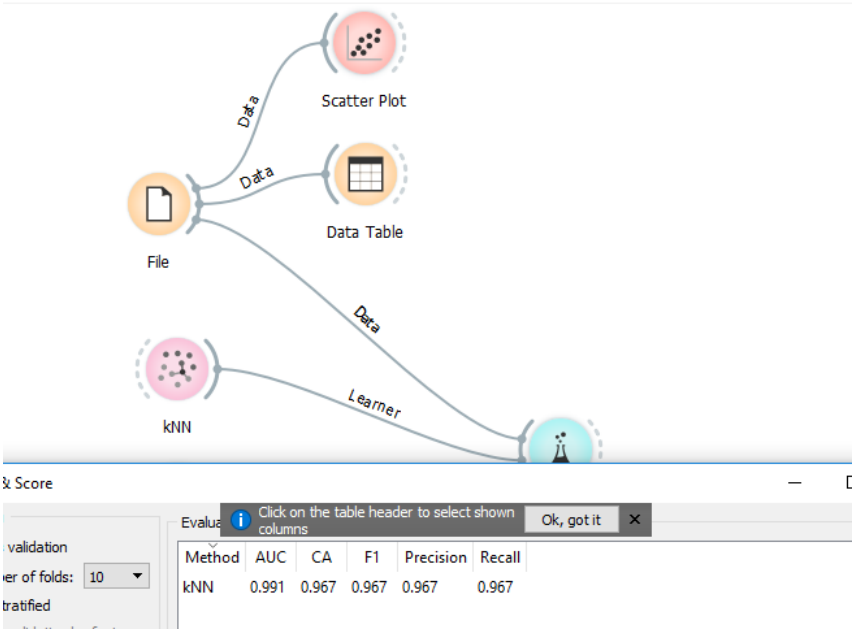


Figura 4 – Visualização do dataset com o widget "Scatter Plot".

Avalie agora a capacidade de três classificadores:

- K-nearest neigbours - KNN
- Máquinas de vetores de suporte - SVM
- Redes Neuronais MLP

Identifique os respetivos widgets, use cross-validation de 10 para treino e apresente a respetiva matriz de confusão. Apresente como resultado o classificador com melhor desempenho. Que conclusões pode retirar?



F

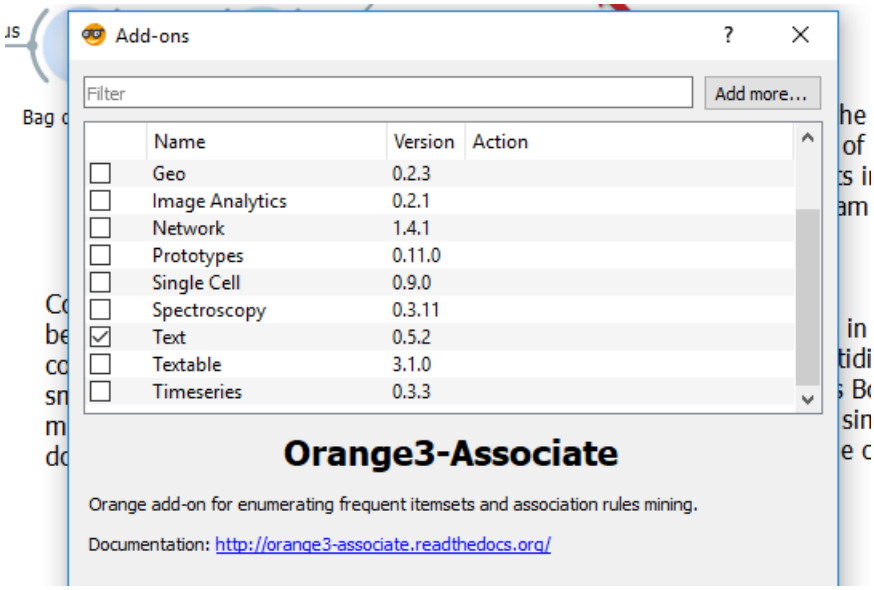
Figura 5 – workflow com o widget "KNN".

Analise e descreva o problema de classificação. Quantos exemplos, atributos e classes existem?

Defina um modelo para classificar corretamente os exemplos do conjunto de dados. Inclua o operador para aplicação do modelo "Apply model" e avalie os resultados. Quantos exemplos foram corretamente classificados?

4. Classificação de Documentos

Em "add-ons", instale a biblioteca para Text Mining e construa o workflow apresentado na (Figura 6).



F

Figura 6a – Instalação de Add-ons.

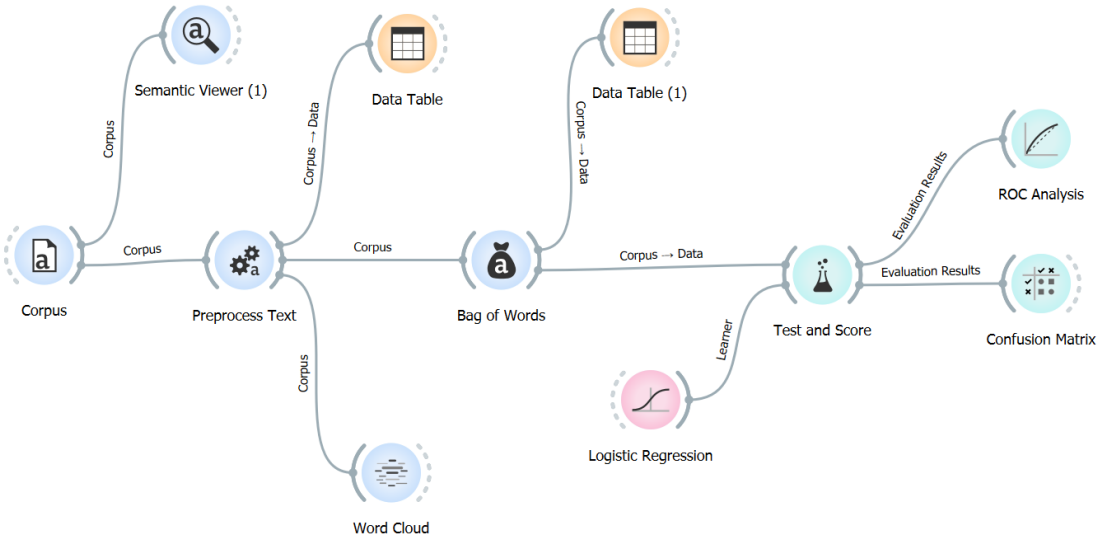


Figura 6b. Workflow para classificação de texto.

Considere como corpus o "dataset" de contos de Grimm, incluído no Orange (Figura 7). Pretende-se classificar os documentos em duas categorias "Animal Tales" e "Magic Tales".

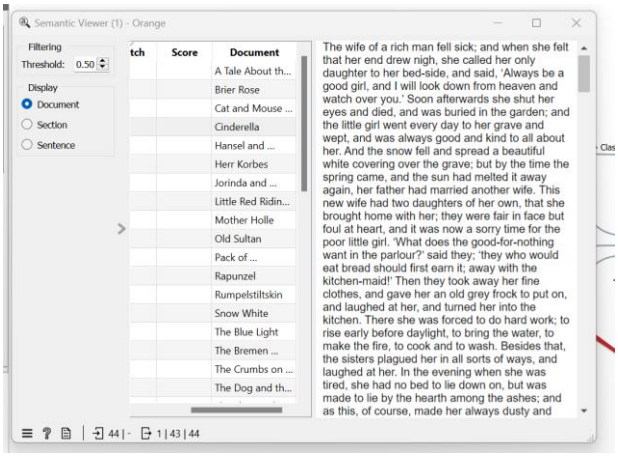


Figura 7. Visualização do corpus a classificar.

1. Analise o processo apresentado e responda às questões:
    a. A primeira etapa consiste na representação de documentos por tokens. Implemente o seguinte processo de pré-processamento (Figura 8):
        i. Transformation: Lower case;
        ii. Tokenization: Manter apenas palavras;
        iii. Filtering: Eliminar "stopwords" + 100 tokens mais frequentes.
        iv. Normalization: "Snowball Stemmer".

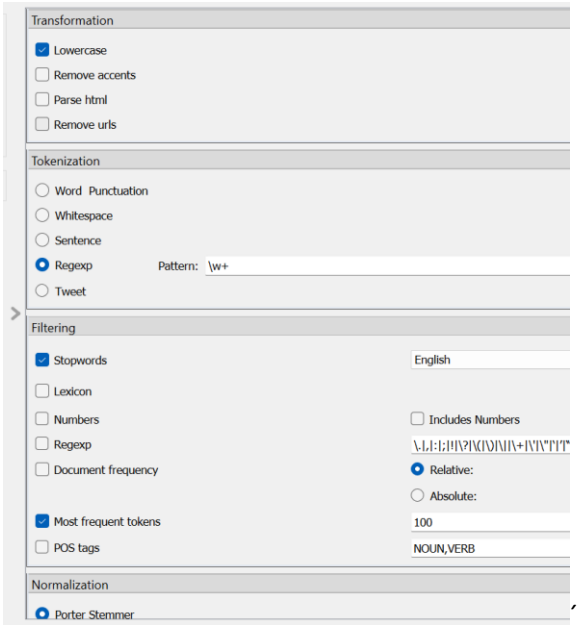


Figura 8. Pré-processamento.

b. Visualize os "tokens" criados - use o widget "word cloud".
c. Descreva o funcionamento do operador "bag of words". Considere como conjunto de features o número de ocorrência de uma palavra - "token" no documento.
d. Construa um modelo baseado numa rede neuronal capaz de classificar corretamente em duas classes.
    i. Apresente os resultados com crossvalidation de 5-folds.
    ii. Forme um conjunto de teste com 20% das amostras. Avalie o resultado para o conjunto de teste.
e. Avalie os resultados obtidos. Ajuste os parâmetros de pré-processamento e o classificador de forma a otimizar o desempenho do classificador.

5. "Sentiment analysis Dataset"

Pretende-se classificar um conjunto de documentos relativos a crítica de filmes, disponível em http://www.cs.cornell.edu/People/pabo/movie-review-data/
O dataset contém 2000 documentos, metade deles atribuídos a cada uma de duas classes: crítica negativa ou crítica positiva.

A Partir de http://www.cs.cornell.edu/people/pabo/movie-review-data/ faça o download do dataset do "polarity dataset V2.0".

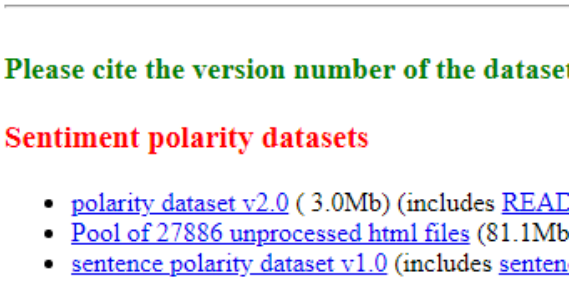


Figura 9 – Dataset.

- Importe os dados para um novo fluxo e use o operador "Import Documents" para ler os ficheiros. Forme um conjunto reduzido de exemplos para facilitar a definição de parâmetros.

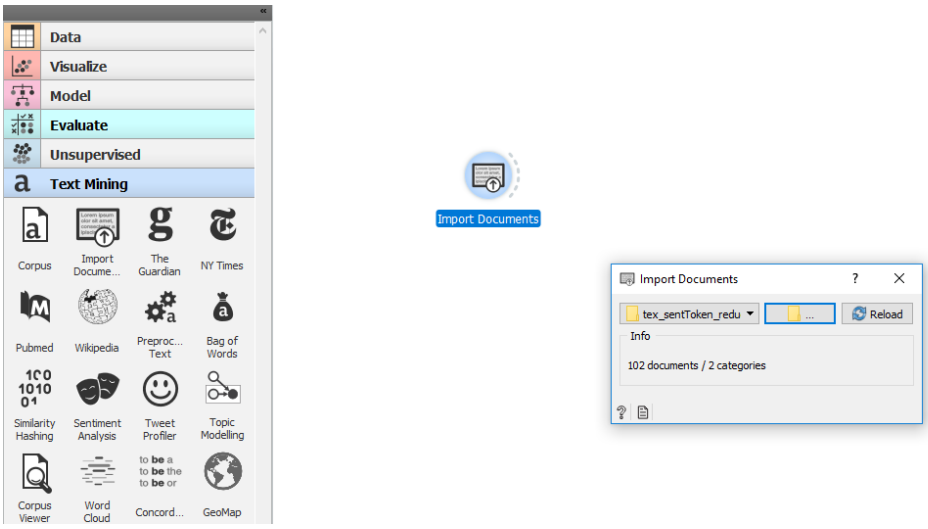


Figura 10 – Exemplo com um dataset reduzido com 102 exemplos.

- Inclua os seguintes parâmetros para pré-processamento:

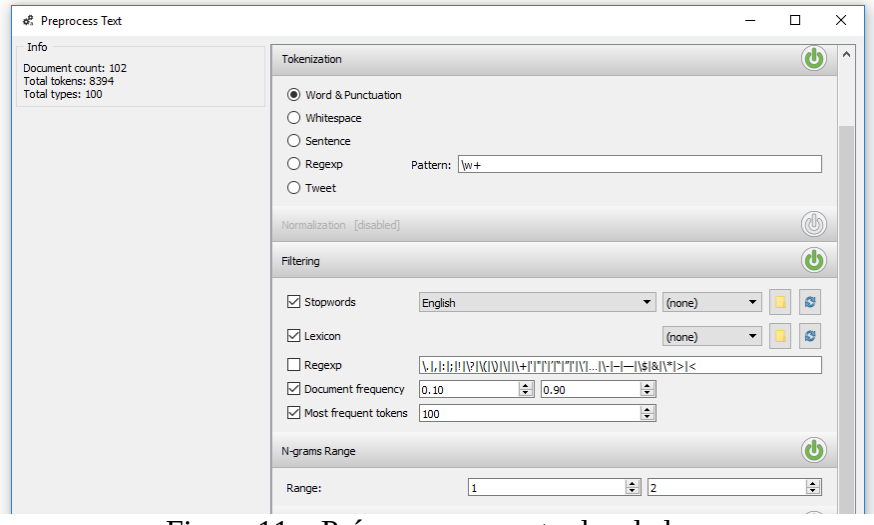


Figura 11 – Pré-processamento dos dados.

- Inclua os widgets “bag of words”, “word cloud” e avalie os resultados. Quais os termos mais relevantes?

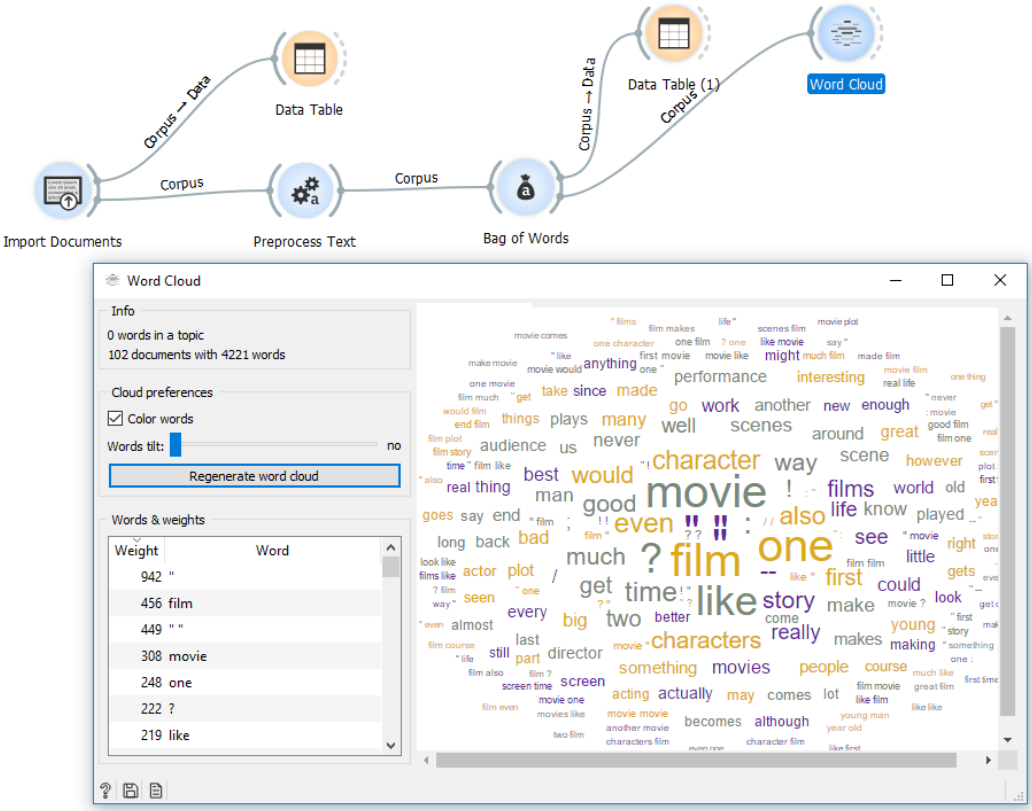


Figura 12 – word cloud.

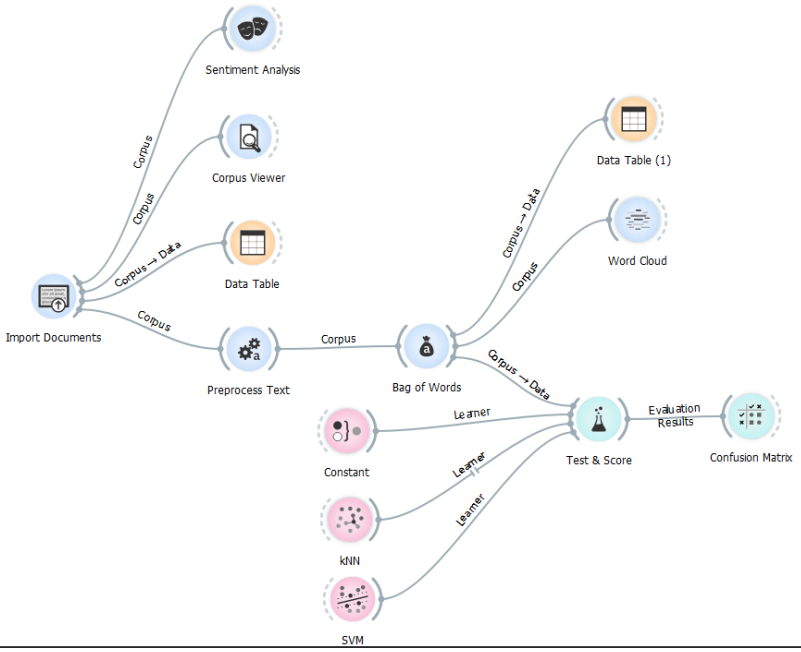


Figura 13 – work flow com três classificadores e matriz de confusão.

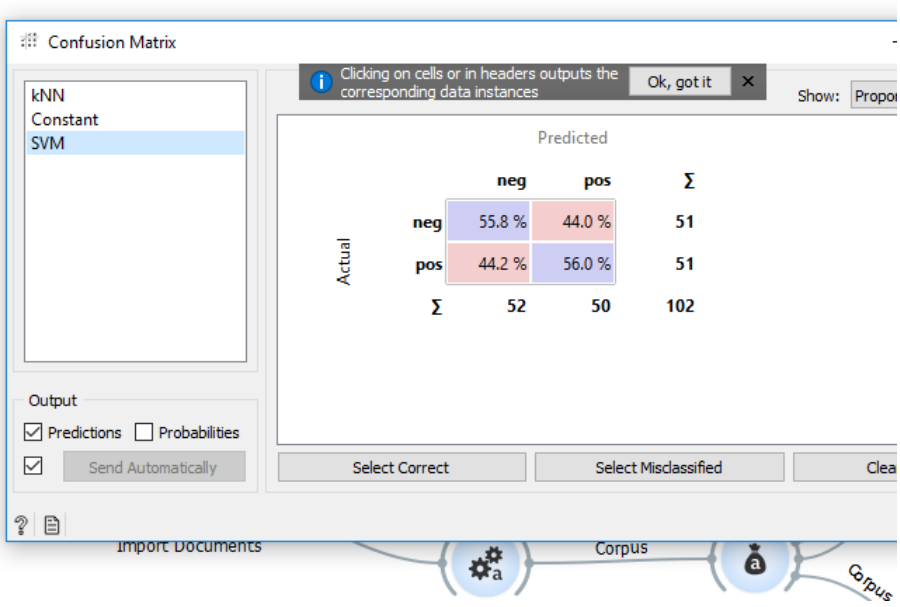


Figura 14. Matriz de confusão com SVM.

Trabalho experimental

i. Modifique os parâmetros associados à técnica de processamento "bag of words" e procure melhorar os resultados. Quais os operadores de pré-processamento de texto mais adequados a este problema?

ii. Aplique o widget "word embedding" e avalie os resultados. Este operador permite uma representação eficiente e densa onde termos semelhantes têm uma codificação semelhante. Os parâmetros são o resultado de treino de uma rede neuronal.

iii. Aplique o melhor classificador ao mesmo problema com o *dataset* completo. Forme um conjunto independente de teste.
No estudo comparativo, tenha em consideração a dimensão do espaço de características e a complexidade da solução (número de neurónios ou vetores de suporte).
Apresente a solução com o melhor valor de AUC e menor número de *features*.

Referências

- [1] Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.".
- [2] https://orangedatamining.com/