

# Aula 4 - Análise textual de uma amostra do Twitter

José Devezas <[jld@fe.up.pt](mailto:jld@fe.up.pt)>

Faculdade de Engenharia da Universidade do Porto

Gestão da Informação em Redes Sociais

Mestrado em Ciência da Informação



# Conteúdos

- Introdução
- Conceitos básicos sobre análise de texto
- Tutorial básico de KNIME
- 2ª avaliação



# Introdução

- Nesta aula vamos esquecer por algum tempo a rede e focar apenas na análise textual (*text mining*).
- Em particular, vamos identificar:
  - Entidades mencionadas em grupos de documentos (*named entity recognition*);
  - Bem como o sentimento dominante em grupos de documentos (*sentiment analysis*).
- Para isso, vamos usar o KNIME, software open source para análise de dados, baseado em programação visual
  - Os “Workflows” serão fornecidos, apenas requerendo pequenas alterações.
- ~~Segundo trabalho individual focado na análise de texto.~~



# Conceitos básicos sobre análise de texto

## Pré-processamento

- Ler ficheiro e remover ruído.
  - Ruído é tudo aquilo que é irrelevante para o nosso objetivo de análise.
  - Para a análise textual do Twitter, removemos:
    - i. URLs;
    - ii. Menções (p.e., @elonmusk);
    - iii. Inícios de tweet com “RT”;
    - iv. Caracteres que representam símbolos ou emoji (p.e., ☹️ 🙌 🤖).
  - Os elementos de i. a iii. são melhor analisados com recurso ao grafo (p.e., i.) link analysis para URLs, ii.) grafo de menções, iii.) grafo de retweets).
- Converter o texto de cada tweet para uma lista de palavras (*tokenization*).



# Conceitos básicos sobre análise de texto

## Reconhecimento de entidades mencionadas

- Entidades são sequências de palavras que se referem a pessoas, organizações, localizações, etc.
- A deteção de entidades depende geralmente da criação de modelos (por exemplo aprendidos automaticamente com *machine learning*, ou criados com recurso a gramáticas), que dependem da língua e de um conjunto de dados previamente anotado.
  - As anotações identificam as fronteiras das entidades no texto, identificando a sua classe. Por exemplo:  
“<person>José Devezas</person> trabalha na <organization>FEUP</organization>, que fica na cidade do <location>Porto</location>.”
- Uma forma rápida de caracterizar o conjunto de dados é identificar as palavras-chave ou entidades mais frequentes, visualizando-as sobre a forma de uma *tag cloud*.



# Conceitos básicos sobre análise de texto

## Análise de sentimento

- Uma das formas mais simples de análise de sentimento é a classificação da polaridade de um texto:
  - **Positivo;**
  - **Neutro;**
  - **Negativo.**
- Para isso é geralmente necessário aprender modelos com técnicas de *machine learning* e com base num conjunto de textos previamente anotados com a classe de polaridade.
- Durante a aprendizagem, o modelo vai sendo melhorado conforme vão sendo fornecidos mais textos anotados com a polaridade.
  - **Intuição:** quanto mais textos positivos, melhor sabemos se determinada palavra é usada em textos com sentimento positivo; palavras mais positivas implicam textos mais positivos (e vice-versa).

# Tutorial básico de KNIME



# Visão geral

O KNIME tem vários nós que executam tarefas específicas. Os nós podem ter entradas e saídas. Existem nós, como o “File Reader”, que apenas têm uma saída e servem para ler ficheiros do sistema.

Workflows-exemplo (EXAMPLES)  
e workflows criados pelo  
utilizador (LOCAL)

Descrição detalhada  
sobre o nó selecionado.

The screenshot displays the KNIME software interface. On the left, the 'KNIME Explorer' pane shows a tree view with 'LOCAL (Local Workspace)' expanded, containing 'GIRS - Polarity Classification' and 'GIRS - Tag Cloud'. The 'Workflow Coach' pane shows 'Node recommendations only available'. The 'Node Repository' pane shows a search bar and a list of categories including 'IO', 'Manipulation', 'Views', 'Analytics', 'Database', 'Other Data Types', 'Structured Data', 'Scripting', 'Tool Integration', 'Community Nodes', 'KNIME Labs', 'Workflow Control', 'Social Media', 'Reporting', 'Chemistry', and 'ChemAxon / Infocom'. The main workspace shows a workflow with three sections: 'Loading and Filtering' (File Reader, Exercício, String Replacer, Strings To Document, Column Filter), 'Named Entity Recognition' (OpenNLP NE tagger, OpenNLP NE tagger, OpenNLP NE tagger), and 'Named Entity Frequency Calculation' (Tag Filter, Bag of Words Creator, Tags to String, TF, GroupBy). The 'File Reader' node is selected, and its description is shown in the 'Node Descri...' pane on the right. The 'Console' pane at the bottom shows a list of error and warning messages.

**Nós disponíveis**

# Configurar um nó

Cada nó pode ser ligado a outros nós (clcando numa entrada ou saída e arrastando), devendo ser previamente configurado (selecionando “Configure” com o botão direito do rato ou utilizando duplo clique).

The screenshot displays the KNIME Explorer interface with a workflow titled 'GIRS - Polarity Classification'. The workflow consists of several nodes: 'File Reader', 'String Replacer', 'Strings To Document', 'Column Filter', 'penNLP NE tagger', and 'GroupBy'. A context menu is open over the 'File Reader' node, showing options such as 'Execute', 'Configure...', 'Cancel', 'Reset', 'Edit Node Description...', 'New Workflow Annotation', 'Collapse into Metanode', 'Encapsulate into Wrapped Metanode', 'Compare Nodes', 'Show Flow Variable Ports', 'Cut', 'Copy', 'Paste', 'Undo', 'Redo', 'Delete', and 'File Table'. The 'Configure...' option is highlighted. On the right side, the 'Node Description' panel for the 'File Reader' node is visible, providing detailed instructions on how to configure the node for reading data from an ASCII file or URL.

**File Reader**

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats. When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely but only the first

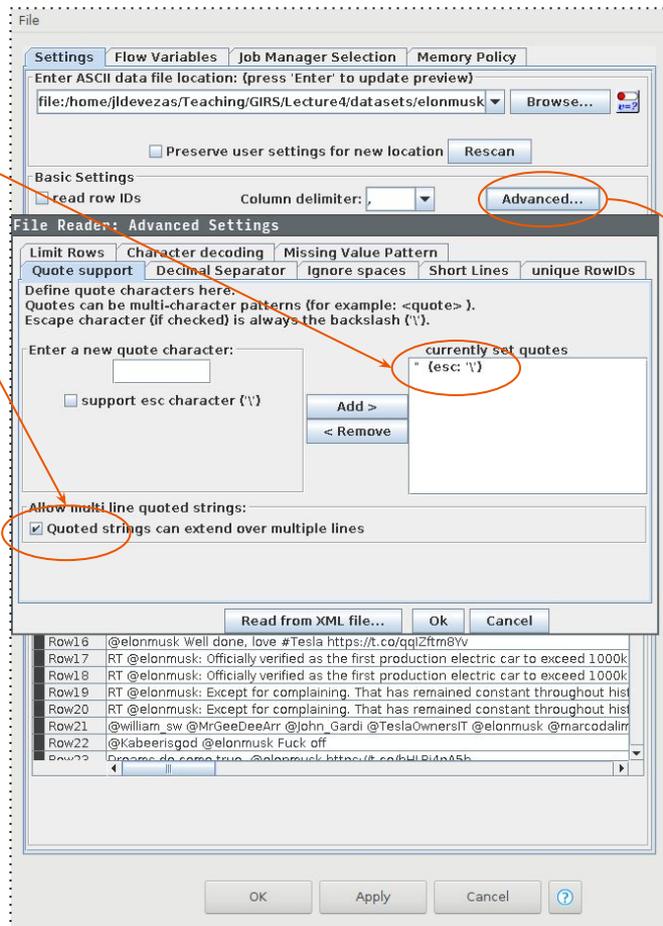
**KNIME Console**

```
ERROR Split Collection Column 3:288 Temp folder "/tmp/knime_08_Sentiment_CL79
ERROR Random Forest Predictor 3:302 Temp folder "/tmp/knime_08_Sentiment_CL79
ERROR Scorer 3:280 Temp folder "/tmp/knime_08_Sentiment_CL79431
ERROR Scorer 3:280 Temp folder "/tmp/knime_08_Sentiment_CL79431
WARN Row Filter 2:307 No row filter specified
WARN Joiner 2:305:295 Memory is low. I have no chance to free memo
ERROR Vocabulary Extractor (deprecated) 2:306:293 Unable to clone input data at p
WARN KNIMEApplication$3 Potential deadlock in SWT Display thread det
WARN Joiner 2:306:299 Please define at least one joining column pa
WARN Joiner 2:306:299 Memory is low. I have no chance to free memo
ERROR Vocabulary Extractor (deprecated) 2:306:293 Unable to clone input data at p
```

# Configurar o “File Reader”

O nó “File Reader” permite carregar dados a partir de ficheiros de texto. No nosso caso, vamos carregar um CSV com o conjunto de tweets recolhidos a partir das contas do @realDonaldTrump e do @elonmusk.

Importante!



# Executar o Workflow

Para executar o Workflow, deve clicar-se com o botão direito no nó para o qual queremos obter um resultado e clicar em “Execute”. O nó deve passar para o estado verde ou então devolver um aviso de erro.

The screenshot displays the KNIME software interface. The main workspace shows a workflow with several nodes: File Reader (Load Twitter data), String Replacer (Exercício), Stings To Document, Column Filter, Named Entity Recognition (OpenNLP NE tagger, OpenNLP), Tag Filter, and Bag of Words Creator. The 'String Replacer' node is highlighted with a yellow box, and its context menu is open, showing the 'Execute' option. The KNIME Console at the bottom shows error messages:

```
KNIME Console
ERROR Split Collection Column 3:298 Temp folder */tmp/knime_08_Sentiment_CL79
ERROR Random Forest Predictor 3:302 Temp folder */tmp/knime_08_Sentiment_CL79
ERROR Scorer 3:280 Temp folder */tmp/knime_08_Sentiment_CL79431
ERROR Scorer 3:280 Temp folder */tmp/knime_08_Sentiment_CL79431
WARN Row Filter 2:307 No row filter specified
WARN Joiner 2:305:295 Memory is low. I have no chance to free memo
ERROR Vocabulary Extractor (deprecated) 2:306:299 Unable to clone input data at p
WARN KNIMEApplications$3 Potential deadlock in SWT Display thread det
WARN Joiner 2:306:299 Please define at least one joining column pa
WARN Joiner 2:306:299 Memory is low. I have no chance to free memo
ERROR Vocabulary Extractor (deprecated) 2:306:299 Unable to clone input data at p
```

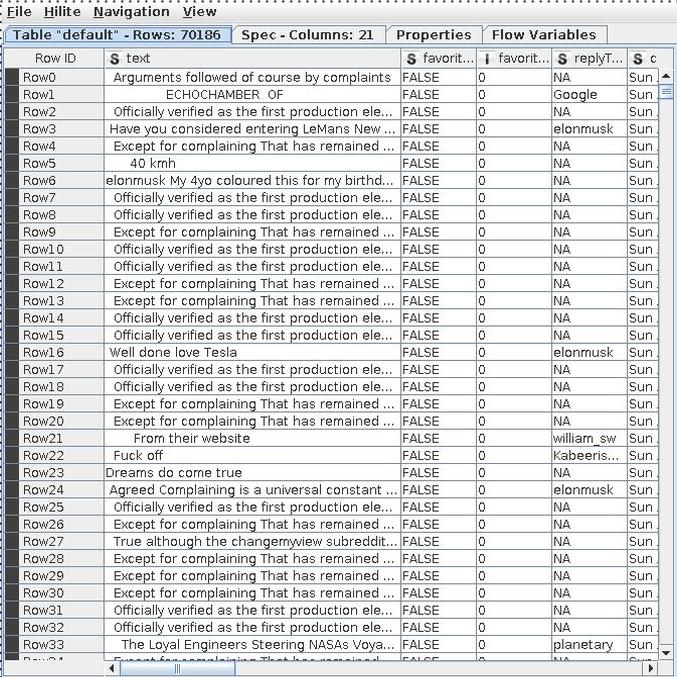
# Visualizar os resultados

Para visualizar os resultados de um nó, deve clicar-se com o botão direito e selecionar uma das opções disponíveis, geralmente com um ícone de lupa+tabela. As opções disponíveis dependem do nó em particular.

The screenshot displays the KNIME software interface. The main workspace shows a workflow with several nodes: 'File Reader', 'String Replacer', 'Strings To Document', 'Column Filter', 'Named Entity Recognition', 'OpenNLP NE tagger', 'OpenNLP', 'Named Entity Frequency Calculation', 'Tag Filter', and 'Bag of Words Creator'. A context menu is open over the 'String Replacer' node, listing various actions such as 'Configure...', 'Execute', 'Cancel', 'Reset', 'Edit Node Description...', 'New Workflow Annotation', 'Collapse into Metanode', 'Encapsulate into Wrapped Metanode', 'Compare Nodes', 'Show Flow Variable Ports', 'Cut', 'Copy', 'Paste', 'Undo', 'Redo', and 'Delete'. The 'String Replacer' node's configuration dialog is also visible on the right, showing options for 'Target column' and 'Pattern type'. The bottom panel shows the 'KNIME Console' with various error and warning messages.

# Visualizar os resultados para o “String Replacer”

O “String Replacer” remove ruído do texto (URLs, menções a contas do Twitter, expressões de retweet, e caracteres não-alfanuméricos). Ao clicar em “Input with replaced values” podemos ver o resultado deste processamento.



Row ID	text	favorit...	favorit...	replyT...	S c
Row0	Arguments followed of course by complaints	FALSE	0	NA	Sun
Row1	ECHOCHAMBER OF	FALSE	0	Google	Sun
Row2	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row3	Have you considered entering LeMans New ...	FALSE	0	elonmusk	Sun
Row4	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row5	40 kmh	FALSE	0	NA	Sun
Row6	elonmusk My 4yo coloured this for my birthd...	FALSE	0	NA	Sun
Row7	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row8	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row9	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row10	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row11	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row12	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row13	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row14	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row15	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row16	Well done love Tesla	FALSE	0	elonmusk	Sun
Row17	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row18	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row19	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row20	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row21	From their website	FALSE	0	william_sw	Sun
Row22	Fuck off	FALSE	0	Kabeeris...	Sun
Row23	Dreams do come true	FALSE	0	NA	Sun
Row24	Agreed Complaining is a universal constant ...	FALSE	0	elonmusk	Sun
Row25	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row26	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row27	True although the changemvew subreddit...	FALSE	0	NA	Sun
Row28	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row29	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row30	Except for complaining That has remained ...	FALSE	0	NA	Sun
Row31	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row32	Officially verified as the first production ele...	FALSE	0	NA	Sun
Row33	The Loyal Engineers Steering NASAs Voya...	FALSE	0	planetary	Sun



## 2ª avaliação

- Individualmente (à vez, caso não haja computadores para todos):
  - Vamos selecionar dois atributos (colunas no CSV) e decidir valores para filtrar essas colunas (p.e., `retweetCount > 10000`).
  - Alunos no mesmo computador não podem escolher as mesmas colunas.
  - Vamos explorar a *tag cloud* de entidades mencionadas em grupos de tweets restringidos por um e pelo outro filtro.
  - Vamos explorar a polaridade nos mesmos grupos de tweets, aplicando alternadamente ambos os filtros.
- Individualmente:
  - Cada aluno deve escrever um mini relatório, analisando manualmente de forma crítica os resultados obtidos (p.e., citando alguns tweets identificados correta ou incorretamente como positivos ou negativos).
- Entrega:
  - Pode ser feita até à próxima aula.