

# Ferramentas de Mineração de Dados

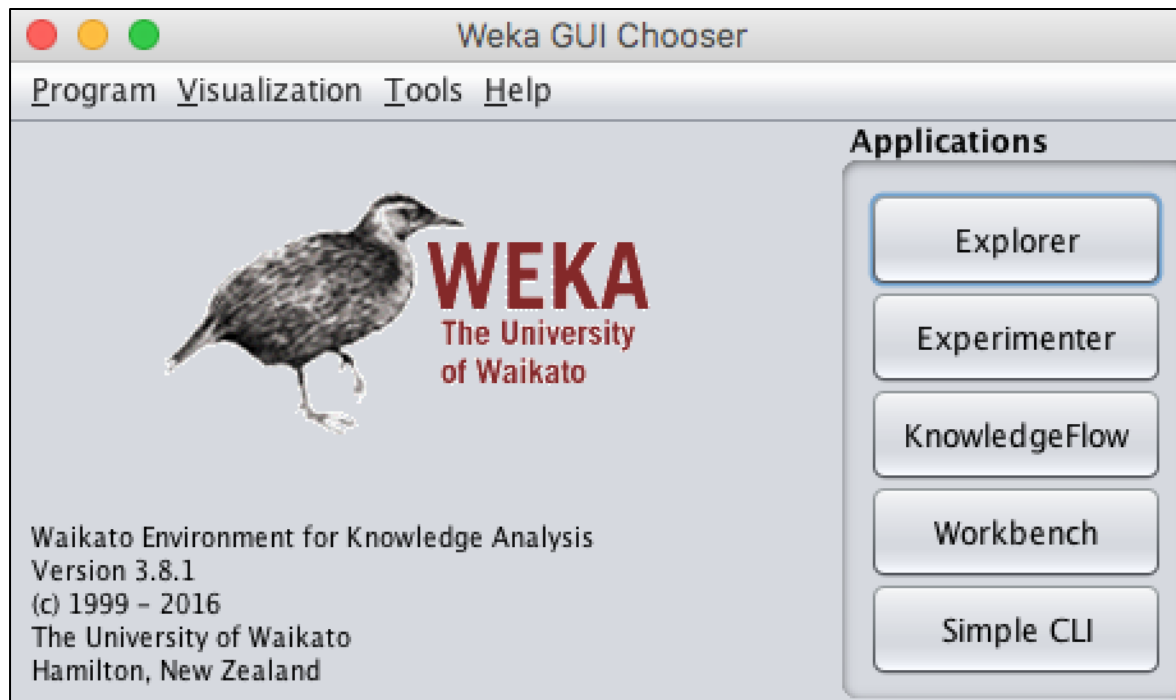
Prof<sup>a</sup>: Solange Oliveira Rezende

# Ferramentas

- Existem várias ferramentas para Mineração de Dados ou Textos
- Nesta aula:
  - Weka
    - Pré-processamento, Classificação, Agrupamento e Regras de Associação
  - Apriori
    - Regras de Associação
  - Torch
    - Pré-processamento, Agrupamento de Textos e Visualização
  - Outras ferramentas

# Weka – Waikato Environment for Knowledge Analysis

- Software popular e código aberto de aprendizagem de máquina.
- Pode ser utilizado através da interface gráfica, linha de comando ou Java API.



# Weka – Instalação

- Instalação simples.
- Versões para Windows, Linux e Mac.
- Site: <https://www.cs.waikato.ac.nz/ml/weka/>



Machine Learning Group at the University of Waikato

Project

Software

Book

Publications

People

Related

## Weka 3: Data Mining Software in Java

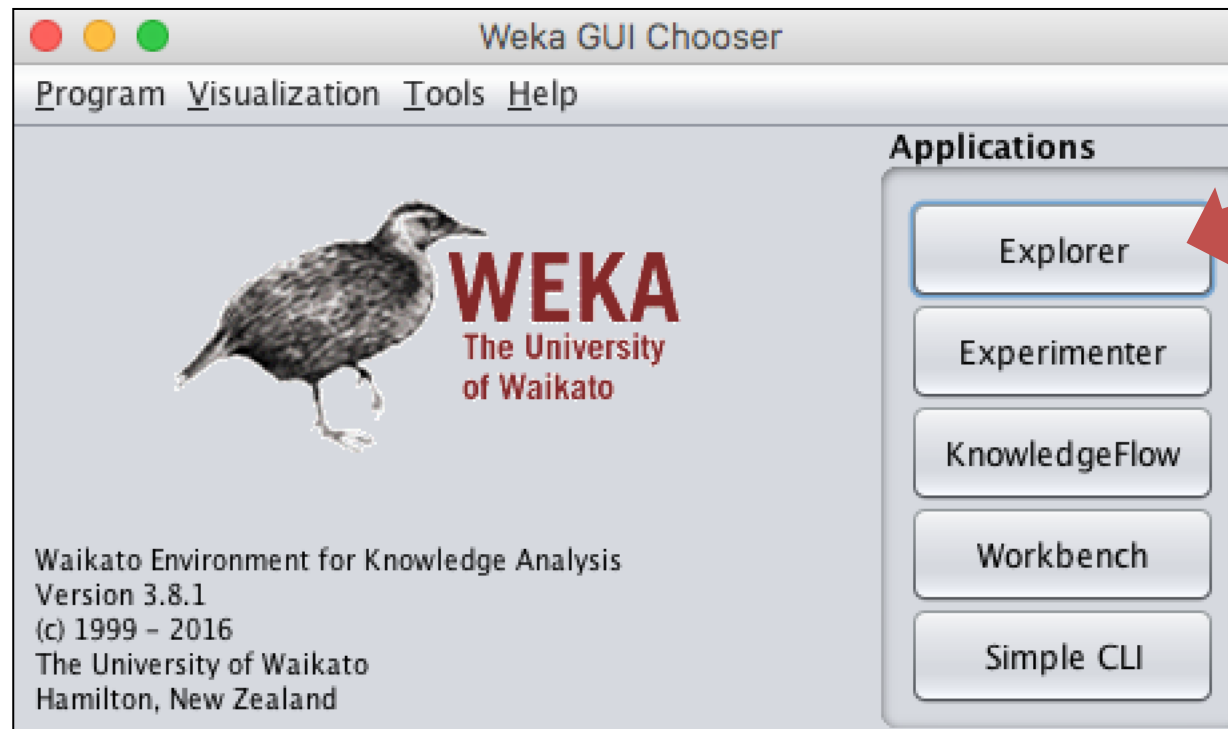
Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

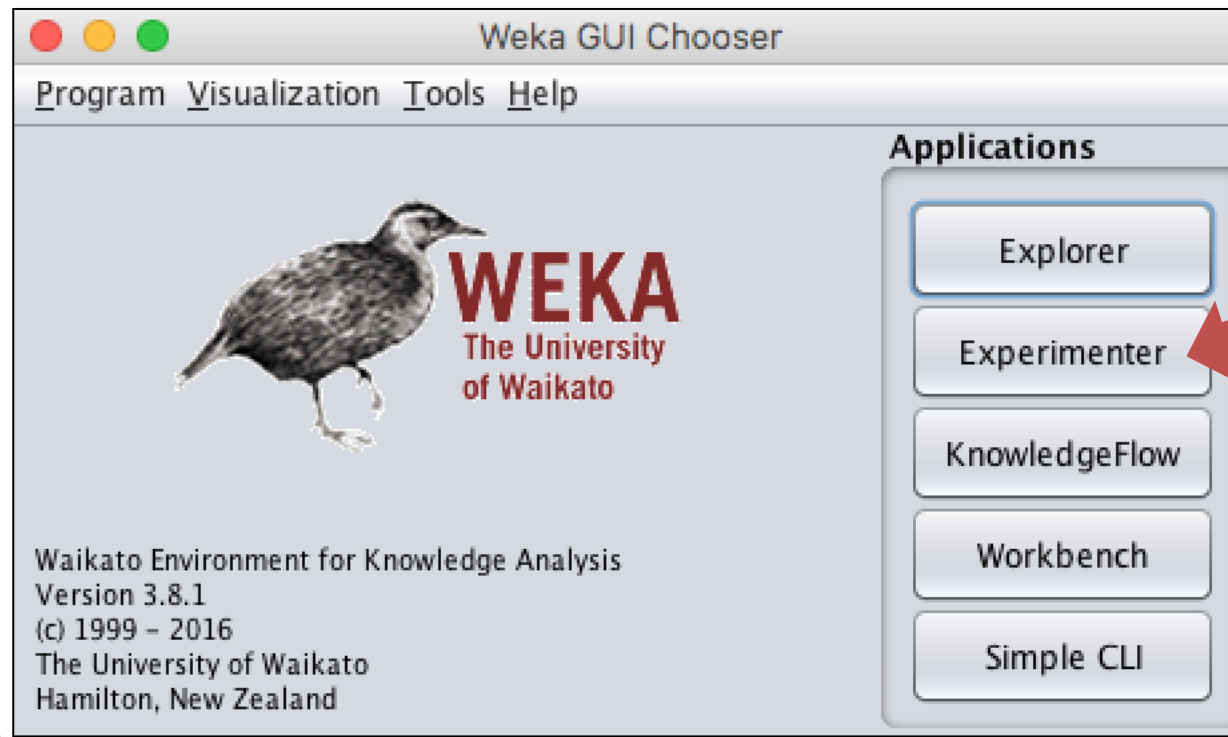
# Weka

- Explorer (aplicação principal)
  - Explorar o conjunto de dados, visualizar dados, aplicar filtros;
  - Realizar classificação, agrupamento, regras de associação e seleção de atributos.



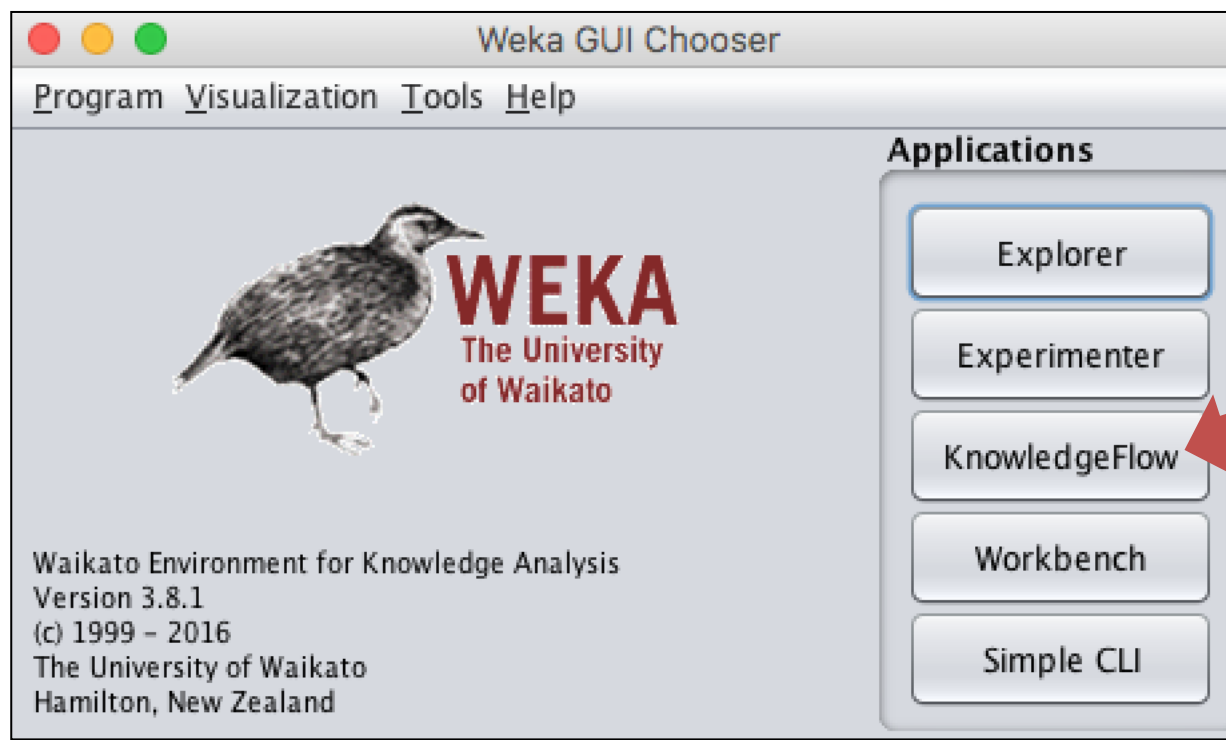
# Weka

- Experimenter
  - Permite a avaliação de desempenho de algoritmos diferentes em base de dados diferentes.



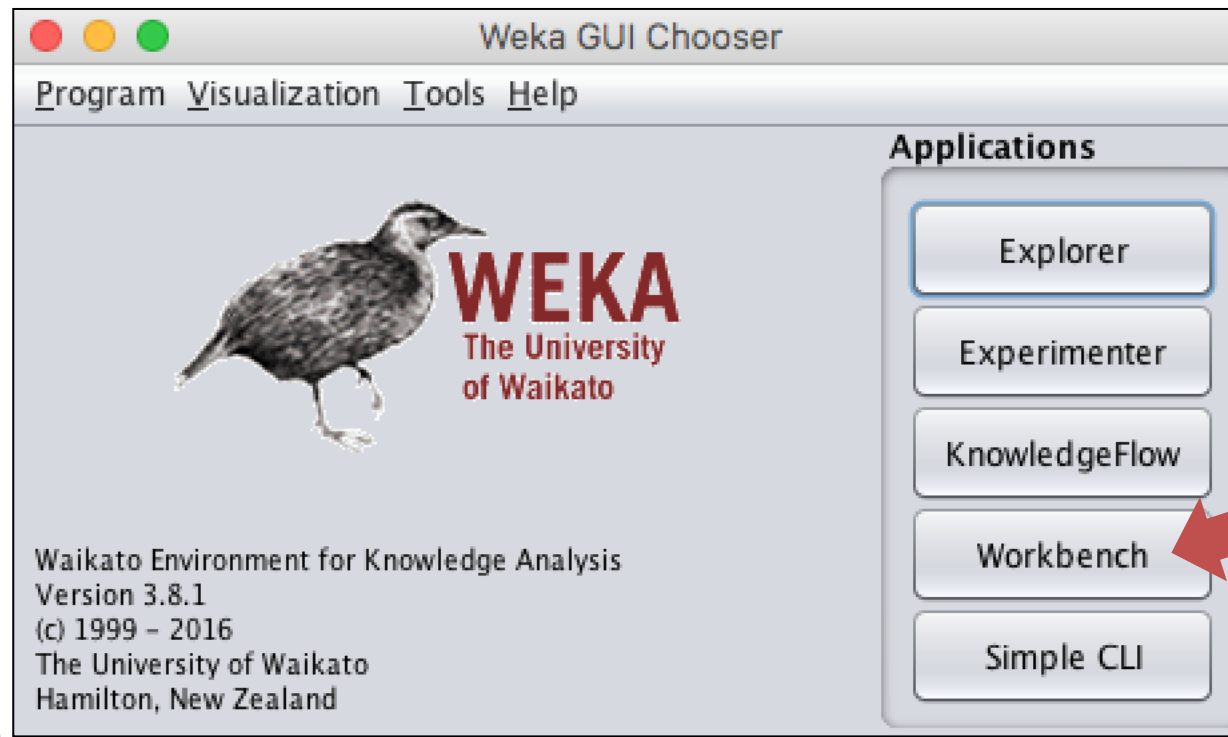
# Weka

- KnowledgeFlow
  - Permite construir o processo de aprendizado de máquina na forma de fluxograma, arrastando e configurando atividades. Essencialmente o mesmo que o Experimenter, porém com interface que permite arrastar e soltar.



# Weka

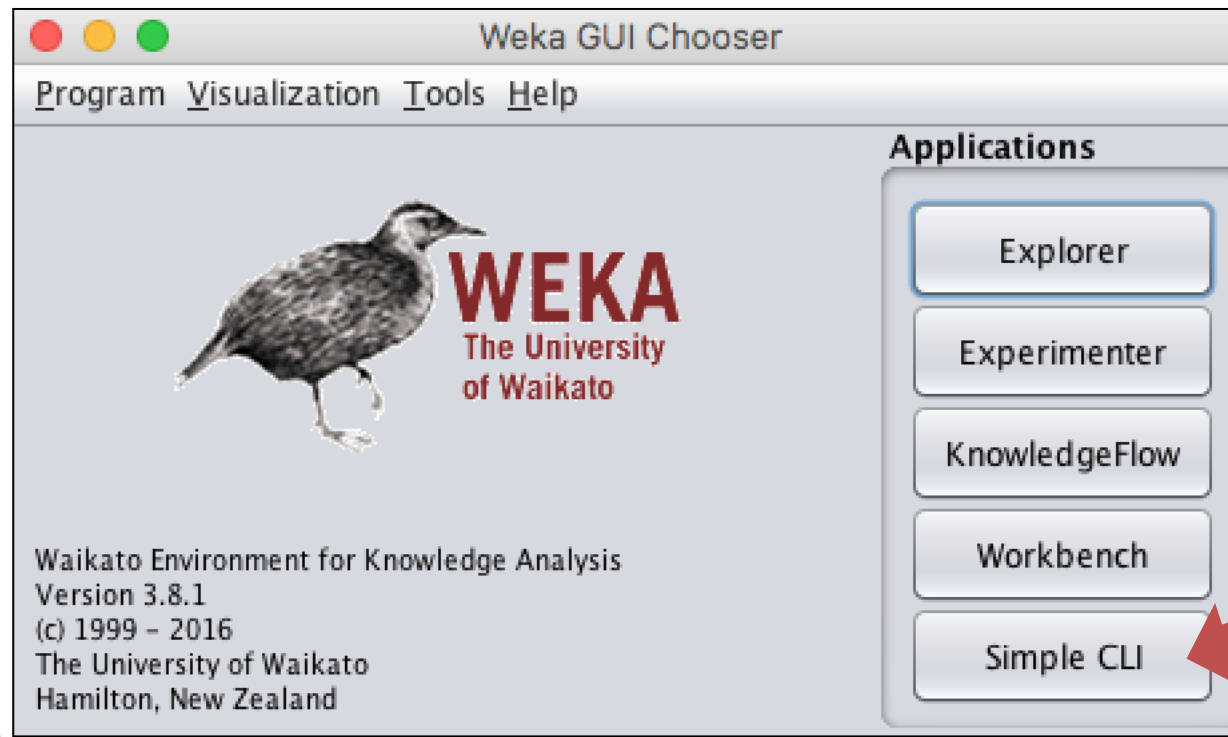
- Workbench
  - Aplicação *all-in-one* que combina toda, porém o usuário é livre para configurar a interface.





# Weka

- Simple CLI
  - Permite utilizar o Weka através de linha de comando.



- Funcionalidades
  - Pré-processamento
  - Classificação
  - Regressão
  - Agrupamento
  - Regras de Associação
  - Visualização
- Distribuído sob a licença GPL

# Weka

- Os dados podem estar nos formatos: ARFF, CSV, entre outros.
- O formato de arquivo de dados próprio da ferramenta WEKA é o **ARFF** (Attribute-Relation File Format)

# Weka – Exemplo de ARFF

```
%Exemplo do formato ARFF
@relation alunos_graduacao

@attribute nome string
@attribute idade numeric
@attribute sex {fem,masc}
@attribute notaP1 numeric
@attribute class {aprovado,reprovado}

@data
Roberta,25,fem,10.0,aprovado
Pedro,20,masc,8.0,aprovado
Maria,22,fem,?,reprovado
Joana,25,fem,9.0,aprovado
```

# Weka – Carregamento dos Dados

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose  Apply

**Current relation**

Relation: iris Attributes: 5  
Instances: 150 Sum of weights: 150

**Attributes**

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> sepallength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

**Selected attribute**

Name: sepallength Type: Numeric  
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

**Status**

OK Log x0

# Weka – Pré-processamento

- As ferramentas de pré-processamento do Weka são chamadas de filtros
  - Discretização
  - Normalização
  - Seleção de atributos
  - Reamostragem

# Weka – Pré-processamento

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

**Filter**

**Current relation**

Relation: iris      Attributes: 5  
Instances: 150      Sum of weights: 150

**Attributes**

All | None | Invert | Pattern

No.		Name
1	<input checked="" type="checkbox"/>	sepalength
2	<input type="checkbox"/>	sepalwidth
3	<input type="checkbox"/>	petallength
4	<input type="checkbox"/>	petalwidth
5	<input type="checkbox"/>	class

**Selected attribute**

Name: sepalength      Type: Numeric  
Missing: 0 (0%)      Distinct: 35      Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)

**Status**

OK  x 0

# Weka - Classificação

The screenshot shows the Weka Explorer application window. The 'Classify' tab is selected and circled in red. Below the tabs, the 'Classifier' section shows a 'Choose' button circled in blue with a red arrow pointing to it, and the classifier name 'J48 -C 0.25 -M 2' displayed next to it. The 'Test options' section on the left has 'Supplied test set' selected. The 'Classifier output' section on the right displays performance metrics and a confusion matrix.

**Classifier**

Choose **J48 -C 0.25 -M 2**

**Test options**

- Use training set
- Supplied test set
- Cross-validation Folds
- Percentage split %

**Classifier output**

```

=== Detailed Accuracy by Class ===
                TP Rate  FP Rate  Precisi
                0.854    0.629    0.761
                0.371    0.146    0.521
Weighted Avg.   0.710    0.484    0.689

=== Confusion Matrix ===
  a  b  <-- classified as
598 102 |  a = good
198 111 |  b = bad
    
```



# Weka - Agrupamento

The screenshot shows the Weka Explorer application window. At the top, there are five tabs: Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. The 'Cluster' tab is highlighted with a red circle. Below the tabs, the 'Clusterer' section is active. It features a 'Choose' button and a text field containing 'SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000'. A red arrow points to the 'Choose' button. Below this, the 'Cluster mode' section has four radio buttons: 'Use training set' (selected), 'Supplied test set' (with a 'Set...' button), 'Percentage split' (with a '%' sign and a '66' input field), and 'Classes to clusters evaluation' (with a '(Nom) class' dropdown menu). A checked checkbox 'Store clusters for visualization' is also present. On the right, the 'Clusterer output' section displays 'Run information' with the following details: Scheme: weka.clu, Relation: german\_c, Instances: 1000, Attributes: 21, and a list of attribute names including checking, duration, credit, and purpose.

# Weka – Regras de Associação

The screenshot shows the Weka Explorer interface with the 'Associate' tab selected. The 'Associator' section displays the 'Apriori' algorithm with the following command line: `Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1`. A red arrow points to the 'Choose' button. The 'Associator output' section shows the following results:

```

minimum support: 0.10 (1094 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44
Size of set of large itemsets L(2): 380
Size of set of large itemsets L(3): 910
Size of set of large itemsets L(4): 633
    
```

The 'Result list (right-click...)' section shows a single entry: '01:12:41 - Apriori'.

# Weka – Seleção de Atributos

**Weka Explorer**

Preprocess Classify Cluster Associate **Select attributes** Visualize

**Attribute Evaluator**

Choose **CfsSubsetEval -P 1 -E 1**

**Search Method**

Choose **BestFirst -D 1 -N 5**

**Attribute Selection Mode**

Use full training set

Cross-validation Folds  Seed

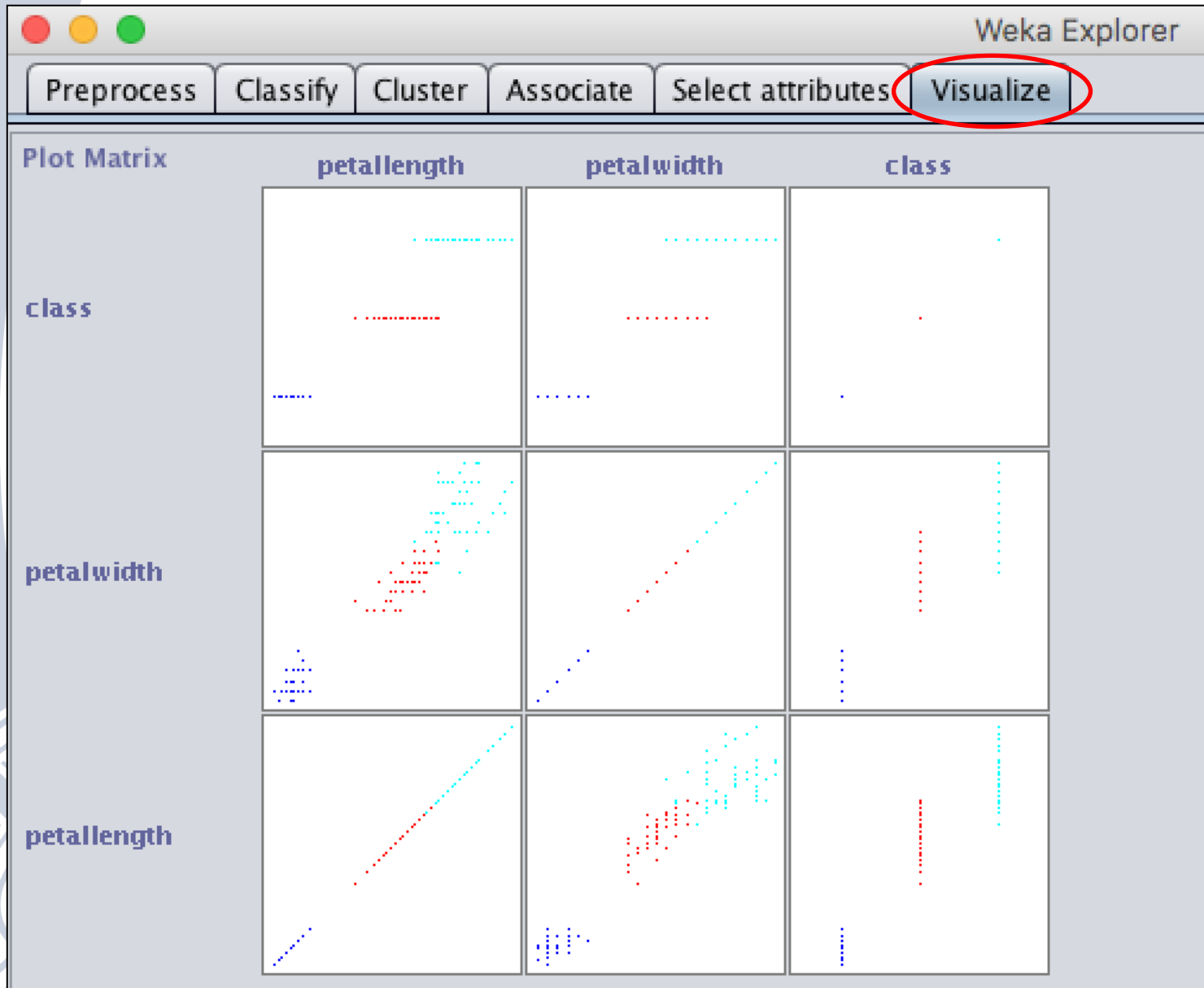
(Nom) total

**Attribute selection output**

```

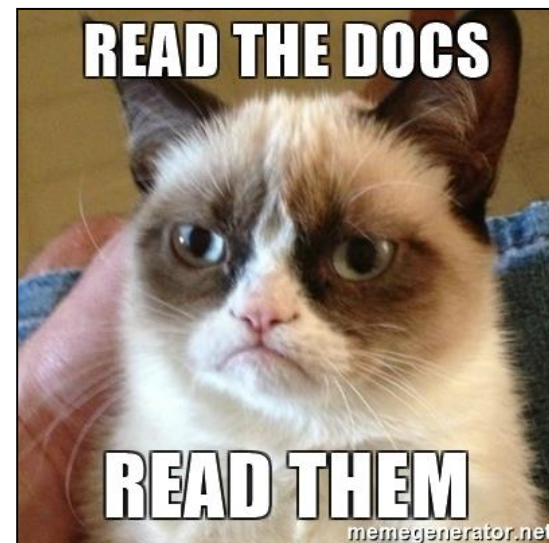
evaluator:      weka.attributeSe
Search:         weka.attributeSe
Relation:       supermarket
Instances:      4627
Attributes:     217
                [list of attribu
Evaluation mode: evaluate o
    
```

# Weka – Visualização dos Dados



# Weka – Documentação

- <https://www.cs.waikato.ac.nz/ml/weka/documentation.html>



Machine Learning Group at the University of Waikato

Project

Software

Book

Publications

People

Related

## Documentation

For an overview of the techniques implemented in Weka, and the software itself, you may want to consider taking a look at the **data mining book**. However, there is a large amount of freely available information as well. Weka has extensive help facilities built in and comes with a comprehensive manual.

# Linguagens mais utilizadas

- R (<http://www.r-project.org/>)
  - dplyr, plyr e data.table (manipulação de dados)
  - stringr (manipulação de strings)
  - zoo (time-series)
  - ggvis, lattice e ggplot2 (gráficos)
  - caret (Machine Learning) Site para download e manuais:
- Python (<https://www.python.org>)
  - SciPy / NumPy (computação científica)
  - Pandas (manipulação de dados)
  - Matplotlib (gráficos)
  - Scikit-learn (Machine Learning)

# Python ou R?

- R ou Python para análise de dados?
  - <http://www.cienciaedados.com/r-ou-python-para-analise-de-dados/>
- Discussão dos prós e contras de cada um:
  - <https://www.kdnuggets.com/2015/05/r-vs-python-data-science.html>
- Infográfico interessante:
  - <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>

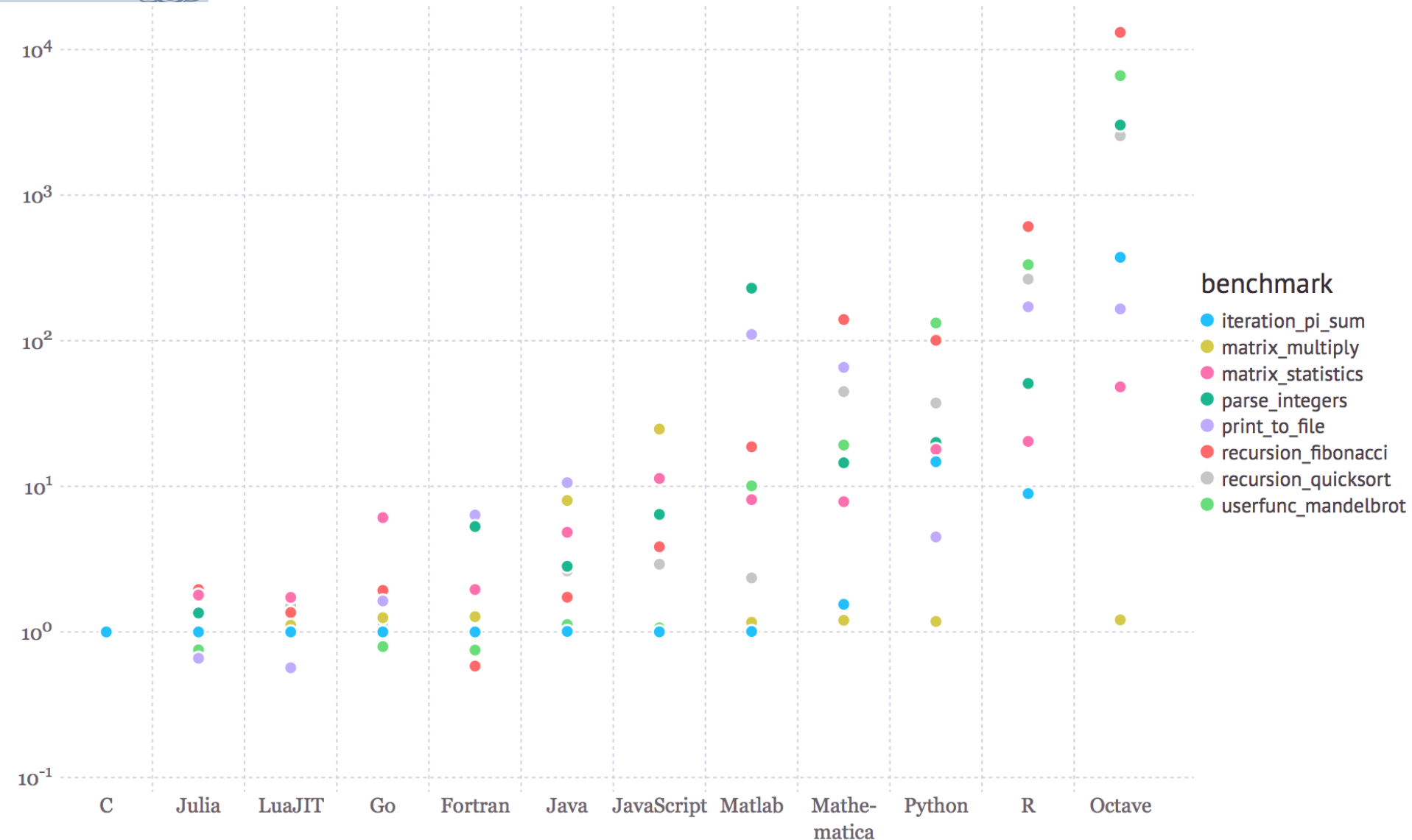
# Exemplo – Mineração de Tweets



# Julia?

- Julia é uma linguagem *open-source* de alto nível e alta performance para computação numérica.
  - Compilada *just-in-time* (JIT);
  - Sintaxe simples (parecida com Python);
  - Tipagem dinâmica (também permite especificar o tipo);
  - Bibliotecas do Python, C e Fortran podem ser utilizadas;
  - Permite metaprogramação;
  - Gerenciamento automático de memória.
- Link : <http://julialang.org/>

# Julia - benchmark



# Apriori

- Implementação do algoritmo Apriori em linguagem C desenvolvida por Cristian Borgelt
- Encontra
  - regras de associação
  - itemsets frequentes
  - itemsets máximos
  - itemsets fechados

# Apriori

- Página para download e instruções para execução:

<http://www.borgelt.net/apriori.html>

- Também é disponibilizado no site uma interface gráfica para visualização de regras desenvolvida em linguagem Java, o ARView
- Outras ferramentas também são disponibilizadas

# Apriori – Formato de Entrada

```
transacoes x
cafe pao manteiga
leite cerveja pao manteiga
cafe pao manteiga
leite cafe pao manteiga
cerveja
manteiga
pao
feijao
arroz feijao
arroz|
```

# Apriori - Execução

- Exemplo de linha de comando

```
./apriori -tr -o -c70.0 -s30.0 transacoes saida
```

- Exemplo do arquivo de saída

```
manteiga <- cafe (30, 100)  
cafe <- manteiga (40, 75)  
pao <- cafe (30, 100)  
pao <- manteiga (40, 100)  
manteiga <- pao (50, 80)
```

# Apriori – Parâmetros de Execução

Parâmetro	Descrição	Padrão
-t#	tipo de alvo (s: itemsets frequentes, c: fechados, m: máximos, g: geradores, r: regras de associação)	s
-m#	número mínimo de itens por conjunto/regra	1
-n#	número máximo de itens por conjunto/regra	sem limite
-s#	suporte mínimo para um conjunto/regra	10%
-S#	suporte máximo para um conjunto/regra (positivo: porcentagem, negativo: número absoluto)	100%
-o	uso da definição de suporte da regra original	cabeça e corpo
-c#	confiança mínima de uma regra	80%
-e#	medida de avaliação adicional	nenhum
-a#	modo de agregação para medida de avaliação	nenhum
-d#	limiar para medida de avaliação adicionada	10%
-i#	melhora mínima de medida de avaliação	sem limite
-z	ignora avaliação abaixo do suporte esperado	avaliar todos
-p#	(tamanho mínimo para) poda com avaliação ( $< 0$ : para frente ("fraco"), $> 0$ para frente ("forte"), $= 0$ : poda para traz)	sem poda
-q#	ordenar itens considerando a frequência (1: ascendente, -1: descendente, 0: não ordenar, 2: ascendente, -2: descendente considerando a soma do tamanho das transações)	2

# Apriori – Parâmetros de Execução

Parâmetro	Descrição	Padrão
-u#	filtrar itens não usados das transações (0: não filtra itens, <0: fração de itens removidos para filtragem, >0: considera taxas de vezes de execução)	0.01
-x	não podar com extensões perfeitas	podar
-y	poda a-posteriori de conjuntos de itens	
-T	não organizar transações com árvores de prefixo	
-Z	imprimir estatísticas dos conjuntos de itens	
-k#	separador de itens para saída	" "
-l#	sinal de implicação das regras de associação	< -
-v#	formato da informação de saída dos conjuntos/regras	(%S)
-l#	ordenar itens na saída pelo tamanho (< 0: descendente, > 0: ascendente)	sem ordenar
-r#	separador de registros/transações	\n
-f#	separadores de itens/campos	\t
-C#	caracteres de comentários	\#
-!	imprimir informação de opção adicional	
infile	arquivo de transações	[obrigatório]
outfile	arquivo de saída	[opcional]
appfile	arquivo definindo itens selecionados	[opcional]



# Torch

- Conjunto de ferramentas para
  - Pré-processamento
  - Agrupamento de Textos (vários algoritmos)
  - Classificação Hierárquica e Visualização de Hierarquias de Tópicos
- Desenvolvida por Ricardo Marcacini – LABIC-ICMC-USP
- Página principal da Torch:  
<http://sites.labic.icmc.usp.br/torch/>

# Torch - JPreText

- Ferramenta para pré-processamento de textos
- Principais funcionalidades:
  - Remoção de stopwords
  - Stemming
  - Criação da bag-of-words
- Download e instruções de uso:

<http://sites.labic.icmc.usp.br/torch/msd2011/jpretext/>

# Torch - JPreText

- Entrada: diretório com um documento em cada arquivo (.txt)
  - “nome\_classe.ID\_ARQUIVO.txt”
- Caso a classe de cada arquivo seja conhecida, ela pode ser usada para avaliação dos agrupamentos gerados pela Torch
- Execução:

```
java -Xmx2G -cp jpretext.jar pretext.Main  
./config.ini
```

# Torch - JPreText

- Arquivo de configurações (config.ini):

```
#Text Collection (e.g. Reuters)
Text Source: ./MinhaColecaoDeTextos

# Preprocessing
# Stemming language options are English, Portuguese or None
Stem Language: Portuguese
Max. Keywords: 20
Stopwords File: ./stopwords.txt

# Term Selection using Document Frequency
Min. DF: 2

# Term weighting can be TF or TFIDF
Term Weighting: TFIDF
Normalization: Yes

# Output
CSV Data File: MinhaColecaoDeTextos.csv
```

# Torch – TopHClust

- Algoritmo para agrupamento hierárquico (Bisecting K-means)
- Página para download e instruções:

<http://sites.labic.icmc.usp.br/torch/msd2011/tophclust/>

- Execução:

```
java -cp tophclust.jar cluster.XSectingKmeans  
./config.ini
```

# Torch - ClusterMap

- Classificação Hierárquica e Visualização de coleções em uma hierarquia de tópicos
- Download e instruções:

<http://sites.labic.icmc.usp.br/torch/msd2011/clustermap/>

- Execução:

```
java -Xmx1G -cp clustermap.jar  
torch.clustermap.Explorer ./config.ini
```

# Torch - ClusterMap

The screenshot displays the Torch - ClusterMap application interface. On the left, a hierarchical tree structure is visible, with the following items expanded:

- Root
  - Rossi, Schumacher, Honda, Munder, Terceira, Min, Campeonato
    - Rali, Min, Prova, Munder, Dakar, Vence, Franca
    - Wirdheim, Formula, Vence, Pedrosa, Campeonato, Sperfico, Ter
    - Hamilton, Raikkonen, Corrida, Alonso, Ponto, Finlandesa, McLaren
    - Rossi, Honda, Barrou, Gibernau, Italiano, Motogp, Espanhola
    - Schumacher, Barrichello, Pole, Alemao, Vaias, Montoya, Position
    - Bate, Laconi, Toseland, Acidente, Belga, Chile, Calendario
  - Olimpica, Ficou, Atenas, Pan, Jogou, Americano, Conquistou
    - Regata, Salto, Ficou, Ponto, Olimpica, Atenas, Lugar
    - Tocha, Olimpica, Luta, Categoria, Cidade, Judo, Atenas
    - Livre, Medalhista, Recorde, Revezamento, Medley, Thiago, Conqui
    - Maratona, Vanderlei, Pan, Kumite, Peruana, Jurandir, Cintia
    - Pan, Futsal, Americano, Dominicana, Rio, Aquatico, Inclusao
    - Atletismo, Modalidade, Pan, Santo, Paisagem, Patinacao, Delega
    - Pan, Reuniao, Modalidade, Cob, Americano, Doping, Odepa
    - Santo, Pan, Hudson, Souza, Herda, Domingo, Americano
    - Santo, Luta, Agosto, Diogo, Dominicana, Domingo, Antoine
    - Dominicana, Pan, Vila, Presidente, Boxe, Santo, Americano
    - Pan, Santo, Castroneves, Americano, Patinacao, Domingo, Medal
      - Pan, Medalhista, Santo, Atletismo, Desempenho, Metade, Winn
      - Castroneves, Patinacao, Norte, Americano, Santo, Conquista, A
    - Pentatlo, Moderno, Americano, Santo, Samantha, Daniele, Pan
    - Pan, Cerimonia, Dominicana, Republica, Santo, Aberto, Xiv
  - Setimo, Dupla, Derrotou, Titulo, Finalista, Kuerfen, Parceria
  - Goleiro, Minuto, Futebol, Selecao, Paraguai, Placar, Ronaldo
  - Handebol, Masculino, Selecao, Feminina, Hungria, Jogou, Derrotou

On the right, the search results are displayed as a list of five items:

- [1. Sam Hornish Jr. vence em sua estréia na Penske - Helio Castroneves fica em 2º Muita emoção na etapa de abertura da Indy Racing League-IRL 2004, realizada no último domingo \(29/02\) em Homestead, Miami....](#)  
 File: 164.txt (0kb) Score: 0.6123069509510087  
 Keywords: [equip, gil, final, fic, volt, prov, ult, castronev, disput, americ, estre, venc, nort, brasil, etap]
- [2. Mayra Ramos conquista a medalha de bronze na Patinação Artística em Santo Domingo A patinadora brasileira Mayra Ramos conquistou nesta terça-feira \(12/08\) a medalha de bronze na Patinação Artística do...](#)  
 File: 330.txt (0kb) Score: 0.5724744154995098  
 Keywords: [fic, conqu, unic, medalh, doming, nort, brasil, sant, patinaca, americ, bronz, prat, republ, emocion, heath, acredit]
- [3. Marcel Stürmer conquista a medalha de ouro para o Brasil na Patinação Artística A apresentação do brasileiro Marcel Stürmer no programa longo livre da Patinação Artística deu ao Brasil mais uma medalh...](#)  
 File: 329.txt (0kb) Score: 0.23750331345294895  
 Keywords: [fic, result, daniel, dominic, quart, marcel, medalh, livr, doming, long, pont, brasil, apoi, feir, jog, patinaca, competica, americ, maurici]
- [4. Armstrong vence a Volta da França pela 6ª vez seguida e entra para a história O ciclista norte-americano Lance Armstrong entrou para a história neste domingo \(25/07\) como o primeiro ciclista a vencer ...](#)  
 File: 413.txt (1kb) Score: 0.11536041681811393  
 Keywords: [min, volt, montanh, consecu, lanc, armstrong, dia, titul, nort, australi, etap, entr, americ, cicl, hist, venc, franc, segu]
- [5. Behar e Shelda perdem para Walsh e May e ficam com a prata no Vôlei de Praia As brasileiras](#)

# Outras Ferramentas

- Orange
  - Uma ferramenta para visualização e análise de dados tanto para usuários iniciantes quanto especialistas
  - Componentes para aprendizado de máquina e consequentemente mineração de dados
  - Permite selecionar dados de treino e teste visualmente
  - Requer Python instalado
  - Download <http://orange.biolab.si/download/>



# Outras Ferramentas

- Tanagra
  - Ferramenta para mineração de dados de propósito acadêmico
  - Métodos para análise exploratória de dados, aprendizado estatístico e aprendizado de máquina
  - Projeto open-source
  - Download: <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>

# Outras Ferramentas

- C4.5
  - Implementação do algoritmo C4.5 de Ross Quinlan
  - Inclui procedimento para amostragem na etapa de avaliação do algoritmo
  - Download:  
<http://www.rulequest.com/Personal/>

# Outras Ferramentas

- MCL++
  - Biblioteca de classes em C++ para aprendizado de máquina
  - supervisionado
  - Domínio público
  - Download: <http://www.sgi.com/tech/mlc/>

# Outras Ferramentas

- MineSet
  - Provê 5 ferramentas para exploração visual de dados e para resultados de mineração de dados
  - Facilita o entendimento de grandes quantidades de dados
  - Download:  
<http://www.dcc.uchile.cl/~rbaeza/cursos/visual/sg/index.html>

# Outras Ferramentas

- Clementine
  - Suíte comercial popular para mineração de dados
  - Embarcado em vários sistemas de mineração de dados
  - Download: <http://www.spss.com/clementine>

# Outras Ferramentas

- Knime
  - Desenvolvido em Java e assim como o Weka, sua biblioteca pode ser facilmente incorporada em outros códigos
  - Permite definir um fluxo de dados
  - Download: <http://www.knime.org>

# Outras Ferramentas

- Keel
  - Ferramenta open-source desenvolvida em Java
  - Desenvolvido para fins de pesquisa e educação
  - Implementa algoritmos evolutivos, fuzzy e demais algoritmos para regressão, classificação, agrupamento, etc.
  - Análise de resultados utilizando testes de significância estatística
  - Inclui técnicas para pré-processamento dos dados
  - Download: <http://www.keel.es/>

# Outras Ferramentas

- Cluto
  - Software de agrupamento para conjuntos dados de alta dimensionalidade
  - Análise de características dos grupos
  - Disponibiliza uma interface gráfica e uma interface web
  - Download:  
<http://glaros.dtc.umn.edu/gkhome/views/cluto>



# Agradecimentos

- Material desenvolvido com ajuda de
  - Rafael Geraldeli Rossi
  - Roberta Akemi Sinoara
  - Camila Vaccari Sundermann
  - Felipe Coutinho