

Aula 8

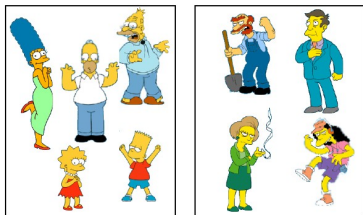
Análise de Agrupamento - Parte II

Introdução ao Agrupamento Hierárquico

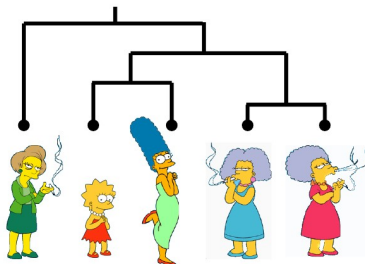
- Método de agrupamento hierárquico criam uma hierarquia de partições
- A hierarquia de partições é representada em um formato de árvore
- Níveis mais baixos da hierarquia representam grupos mais gerais e níveis mais altos representam grupos mais específicos

Introdução ao Agrupamento Hierárquico

Particional



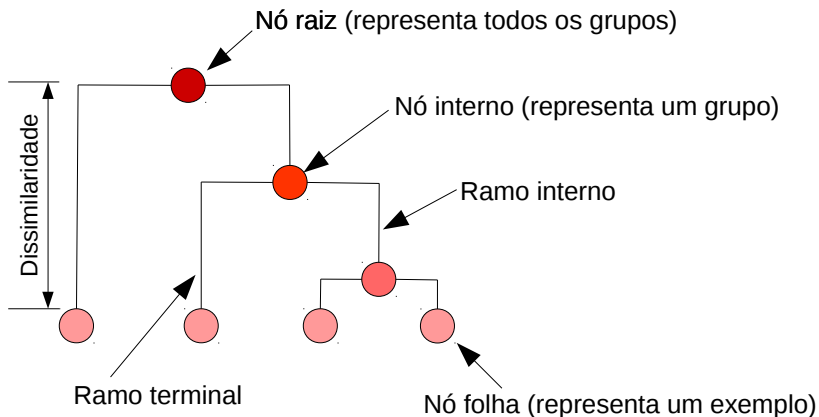
Hierárquico



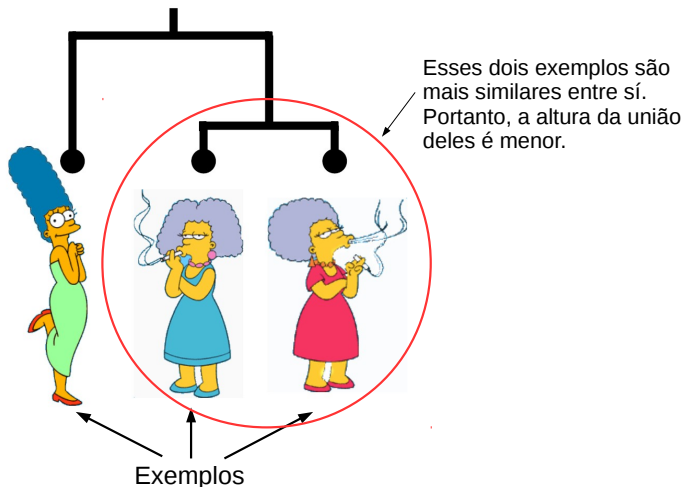
Introdução ao Agrupamento Hierárquico

- Usualmente o resultado de um procedimento de agrupamento hierárquico é representado no formato de uma **dendrograma**
- No dendrograma, além de apresentar uma **estrutura de árvore** que representam os **grupos e subgrupos** que foram unidos durante o procedimento de agrupamento hierárquico, também apresenta a **dissimilaridade entre os grupos**, a qual é determinada pela “**altura**” da união dos grupos

Introdução ao Agrupamento Hierárquico



Introdução ao Agrupamento Hierárquico



Introdução ao Agrupamento Hierárquico

Pedro (português)

Petros (grego), Peter (inglês), Piotr (polonês), Peadar (irlandês), Pierre (francês), Peder (dinamarquês), Peka (havaiano), Pietro (italiano), Piero (italiano alternativo), Petr (tcheco), Pyotr (russo)

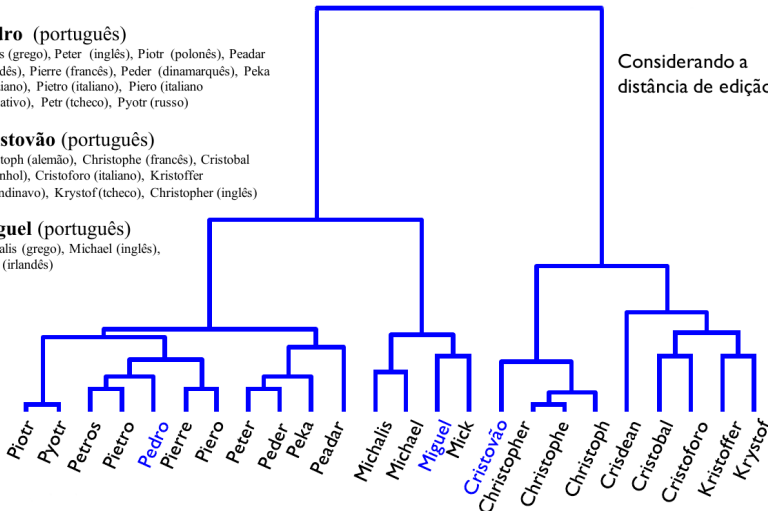
Cristovão (português)

Christoph (alemão), Christophe (francês), Cristobal (espanhol), Cristoforo (italiano), Kristoffer (escandinavo), Krystof (tcheco), Christopher (inglês)

Miguel (português)

Michalis (grego), Michael (inglês), Mick (irlandês)

Considerando a
distância de edição

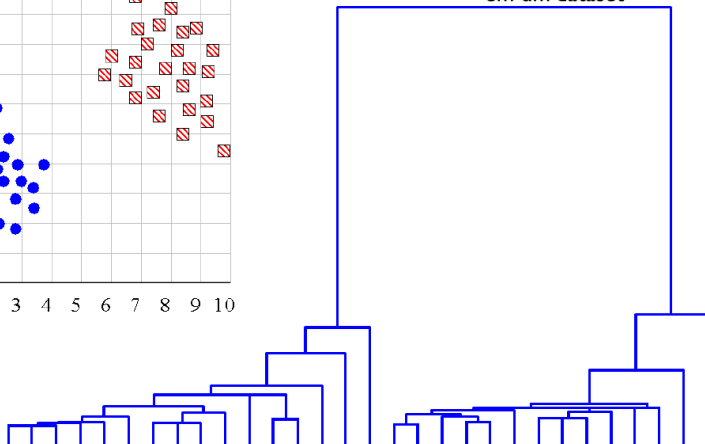
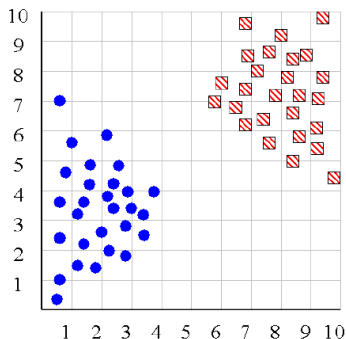


Introdução ao Agrupamento Hierárquico

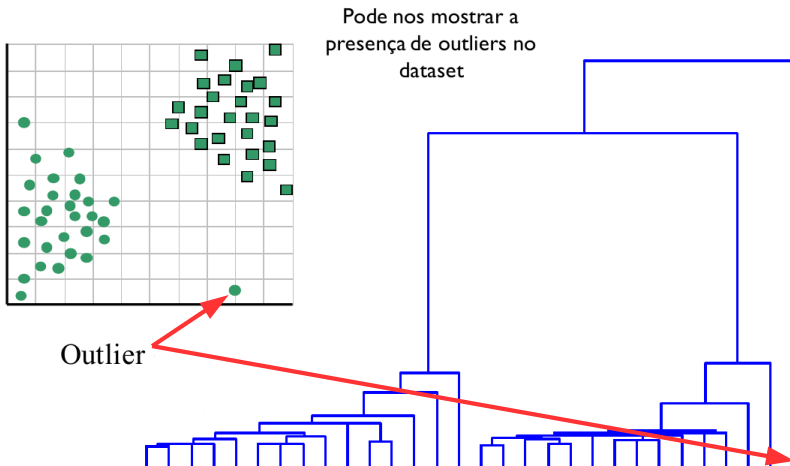
- Além da hierarquia de grupos e as dissimilaridades entre os grupos, o dendrograma pode nos fornecer outras informações úteis:
 - Presença de *outliers* no conjunto de dados
 - Número natural de grupos no conjunto de dados

Introdução ao Agrupamento Hierárquico

Pode nos mostrar o
número correto de clusters
em um dataset



Introdução ao Agrupamento Hierárquico



Introdução ao Agrupamento Hierárquico

- Se no agrupamento particional uma análise combinatorial para gerar todos os possíveis grupos de objetos já era inviável...
- No agrupamento hierárquico gerar todas as possíveis hierarquias de grupos também é intratável...
- **Solução: heurísticas / algoritmos gulosos**

Introdução ao Agrupamento Hierárquico

- Existem duas abordagens para se gerar as hierarquias de grupos
 - **Aglomerativa** (*bottom-up*)
 - **Divisiva** (*top-down*)

Agrupamento Hierárquico Aglomerativo

- Passos

- 1 Inicialmente cada objeto é um grupo
- 2 Faça até haver um único grupo
 - 1 Unir o par de grupos mais próximo

Agrupamento Hierárquico Aglomerativo











- Os algoritmo aglomerativos realizam o processo de agrupamento baseados em uma matriz de proximidades (usualmente a proximidade é uma distância)
- Essa matriz possui uma dimensão $|\mathcal{D}| \times |\mathcal{D}|$, na qual \mathcal{D} representa o conjunto de exemplos de base de dados
- As células dessa matriz armazenam a proximidade entre os grupos
- Usualmente considera-se uma matriz triangular para economizar memória

Agrupamento Hierárquico Aglomerativo

Exemplo de uma matriz de dissimilaridade inicial
(considerando cada objeto como grupo)

$$D(\text{Marge Simpson}, \text{Lisa Simpson}) = 8$$

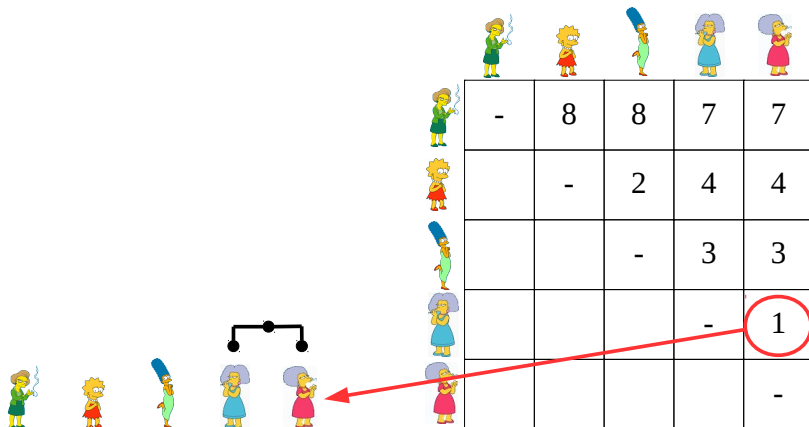
$$D(\text{Maggie Simpson}, \text{Bart Simpson}) = 1$$

					
	-	8	8	7	7
		-	2	4	4
			-	3	3
				-	1
					-

Agrupamento Hierárquico Aglomerativo

- A cada passo, o algoritmo vai escolher qual o “melhor” par de grupos a serem unidos
- Esse “melhor” par de grupos vai ser sempre escolhido com base nas informações da matriz de proximidade

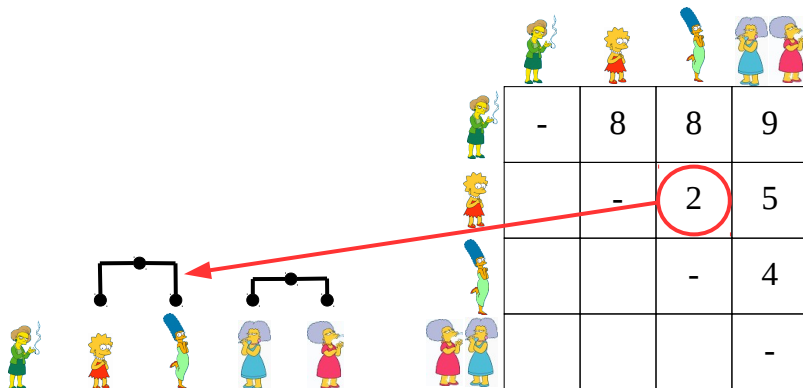
Agrupamento Hierárquico Aglomerativo



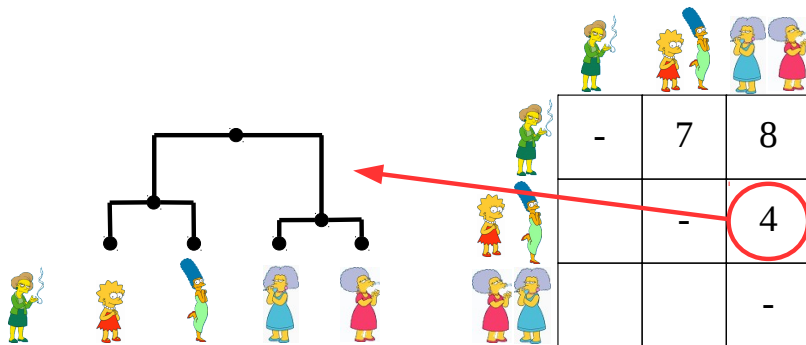
Agrupamento Hierárquico Aglomerativo

- Novamente, são escolhidos dois pares de grupos para serem unidos
- Entretanto, observe agora que têm-se 4 grupos ao invés de 5
- A matriz de distância deve refletir essa atualização
- Deve-se agora considerar a distância dos outros grupos para o novo grupo

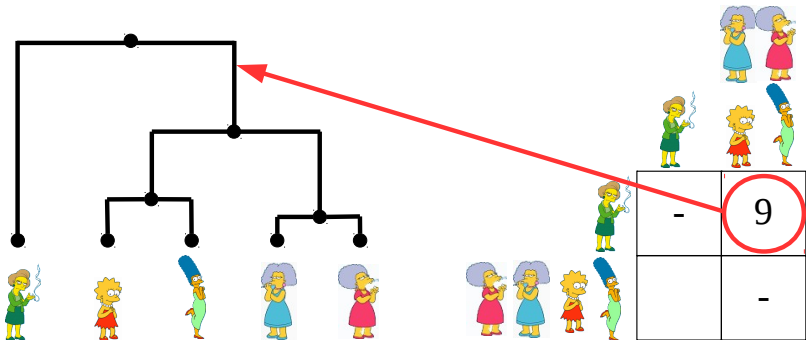
Agrupamento Hierárquico Aglomerativo



Agrupamento Hierárquico Aglomerativo



Agrupamento Hierárquico Aglomerativo



Agrupamento Hierárquico Aglomerativo

- Já foi mostrado nesta disciplina como calcular a proximidade entre dois objetos

Porém, como calcular a proximidade entre grupos de objetos?

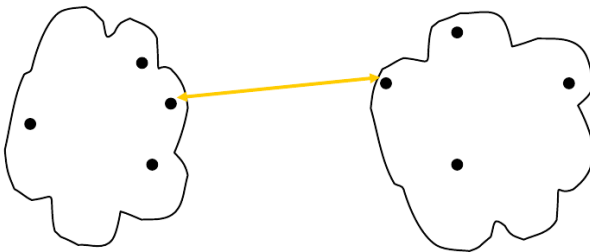


Agrupamento Hierárquico Aglomerativo

- A forma de como calcular a proximidade entre os grupos irá definir o algoritmo de agrupamento hierárquico aglomerativo
- Os algoritmos mais tradicionais são:
 - *Single-linkage* (Min-distance)
 - *Complete-linkage* (Max-distance)
 - *Average-linkage*











Single-Linkage

- A distância entre dois grupos é dada pela menor distância entre um objeto x pertencente a um grupo C_a e um objeto y pertencente a um grupo C_b

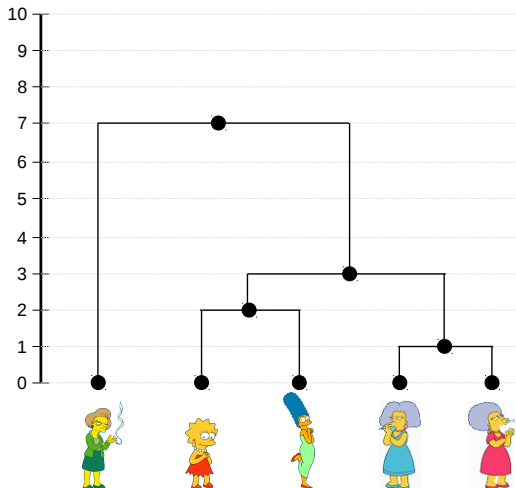


$$D(C_a, C_b) = \min(d(x, y)), x \in C_a, y \in C_b$$

Single-Linkage

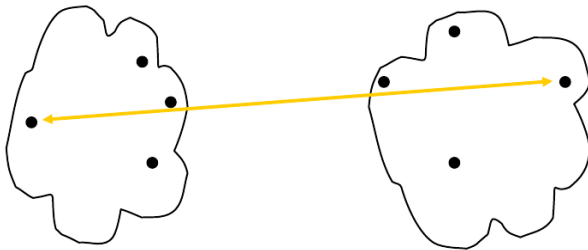
					
	-	10	9	7	8
		-	2	5	4
			-	3	3
				-	1
					-

Single-Linkage













Complete-Linkage

- A distância entre dois grupos é dada pela maior distância entre um objeto x pertencente a um grupo C_a e um objeto y pertencente a um grupo C_b

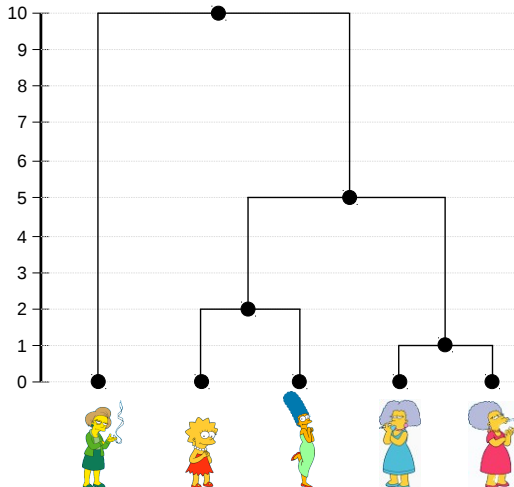


$$D(C_a, C_b) = \max(d(x, y)), x \in C_a, y \in C_b$$

Single-Linkage

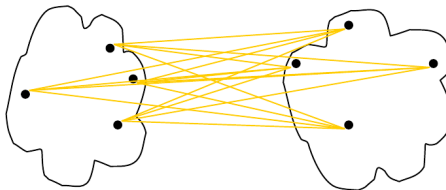
					
	-	10	9	7	8
		-	2	5	4
			-	3	3
				-	1
					-

Single-Linkage













Average-Linkage

- A distância entre dois grupos é dada pela média da distância entre cada par de objetos (x, y) tal que $x \in C_a$ e $y \in C_b$ –
Unweighted Pair Group Method with Arithmetic mean (UPGMA)

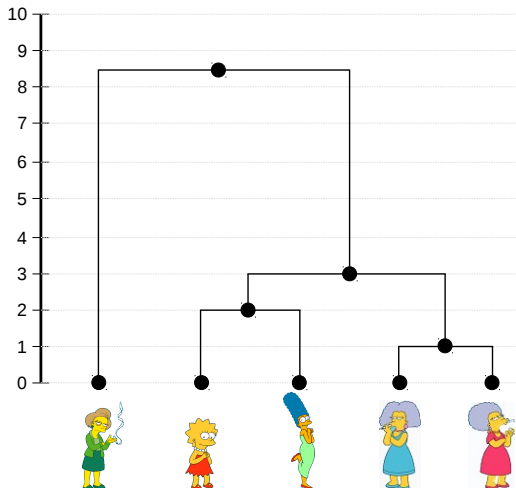


$$D(C_a, C_b) = \frac{\sum_{x \in C_a} \sum_{y \in C_b} d(x, y)}{|C_a| * |C_b|}$$

Average-Linkage

					
	-	10	9	7	8
		-	2	5	4
			-	3	3
				-	1
					-

Average-Linkage

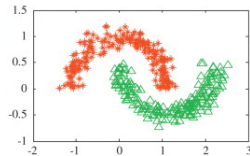


Considerações sobre Agrupamento Hierárquico Aglomerativo

- Complexidade de Espaço: $O(|\mathcal{D}^2|)$
- Complexidade de Tempo: $O(|\mathcal{D}^3|)$ ou $O(|\mathcal{D}^2 \log(\mathcal{D})|)$

Single-Linkage

- Capaz de capturar grupos em formatos não globulares



Complete-Linkage

- Tende a formar grupos globulares
- Tende a quebrar grupos grandes

Average-Linkage

- Meio termo entre *Single-Linkage* e *Average Linkage*
- A tendência é que forme grupos em formatos globulares

Agrupamento Hierárquico Divisivo

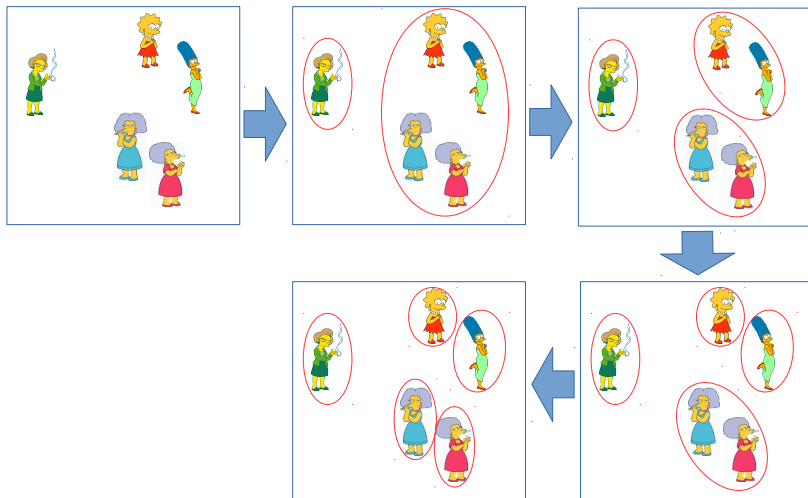
- **Passos**

- ➊ Inicialmente todos os objetos são alocados em um único grupo
- ➋ Faça até haver um objeto por grupo ou até atingir um número de grupos desejado
 - ➊ Dividir um grupo existente em dois grupos

Bisecting k-Means

- O algoritmo mais tradicional é o *Bisecting k-Means*
- Corresponde à aplicação do algoritmo k-Means com $k = 2$ a cada grupo, gerado assim sucessivas divisões e, portanto, uma hierarquia de divisões

Bisecting k-Means



Bisecting k-Means

Como escolher qual grupo vai ser dividido?



Bisecting k-Means

- Grupo com maior tamanho
- Grupo com maior erro quadrático médio

Considerações sobre o Agrupamento Hierárquico Divisivo

- Complexidade de espaço é menor que o do agrupamento aglomerativo: $|D| + |C|$
- Complexidade de tempo também é menor $|D|^2 \times |C|$
- Vale ressaltar que o algoritmo divisivo pode ser parado antes
→ pode-se definir um número de grupos
- Vale ressaltar [2] que assim como no *k-Means*, idealmente há de se executar múltiplas inicializações de centroides e escolher aquela que provê o melhor resultado

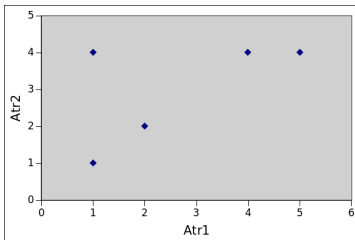
Dendrograma

- Altura do dendrograma
 - Distância entre os centroides dos grupos
 - Erro quadrático do grupo

Exercício

Execute o algoritmo *Bisecting k-Means* para o conjunto de dados apresentado abaixo. Considere a distancia Euclidiana como medida de proximidade. Considere os centroides da primeira divisão $c_1 = \{1, 3\}$ e $c_2 = \{4, 3\}$, e os centroides da segunda divisão como sendo $c_1 = \{2, 1\}$ e $c_2 = \{1, 3\}$. Por fim, construa o dendrograma considerando o erro quadrático como altura das junções.

ID	Atr 1	Atr2
1	1	1
2	2	2
3	1	4
4	4	4
5	5	5



Material Complementar

- Hierarchical clustering

https://en.wikipedia.org/wiki/Hierarchical_clustering

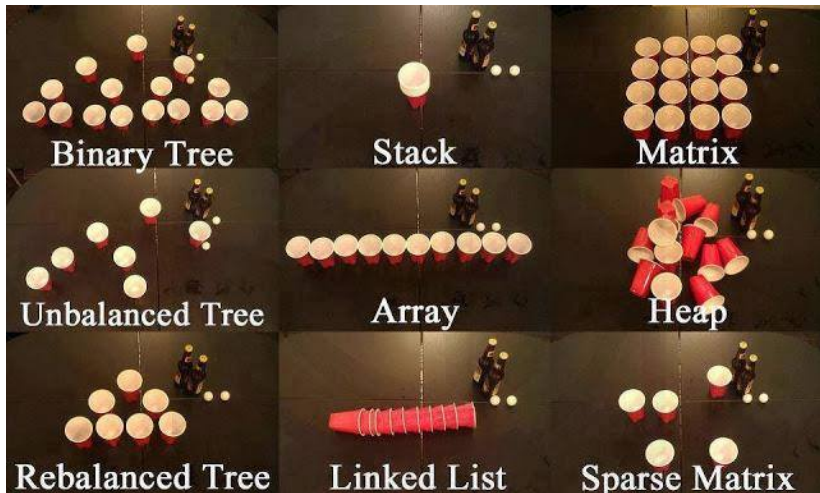
- Hierarchical Clustering

<https://www.kdnuggets.com/2019/09/hierarchical-clustering.html>

- Começando com Orange 05: Clustering Hierárquico

<https://www.youtube.com/watch?v=dJ5z2SRwzgs&v1=pt>

Imagem do Dia



Inteligência Artificial
<http://lives.ufms.br/moodle/>

Rafael Geraldeli Rossi
rafael.g.rossi@ufms.br

Slides baseados no material do Prof. Bruno Nogueira

Referências Bibliográficas I