Classificação de Texto

E

Clusterização de Texto

Luciano Ignaczak
Sócio-Fundador
ignaczak@icybersec.com

Márcio Garcia Martins
Sócio-Fundador
marciogm@icybersec.com

Introdução

- **Aprendizado não-supervisionado**: este modelo não exige dados para treinamento, podendo ser aplicado para qualquer tipo de texto sem necessidade de esforço manual.

- **Aprendizado supervisionado**: este modelo geralmente é suportado por aprendizado de máquina que, baseado em um conjunto de dados usados para treinamento, aprende uma função de classificação que pode ser usada para calcular predições de novos dados.

Fonte: AGGARWAL CHARU C. and ZHAI, CHENGXIANG . **Mining Text Data**. 2012.

Classificação de Texto

- Classificação de Texto (Text Classification) também é conhecida como categorização de texto (Text Categorization);

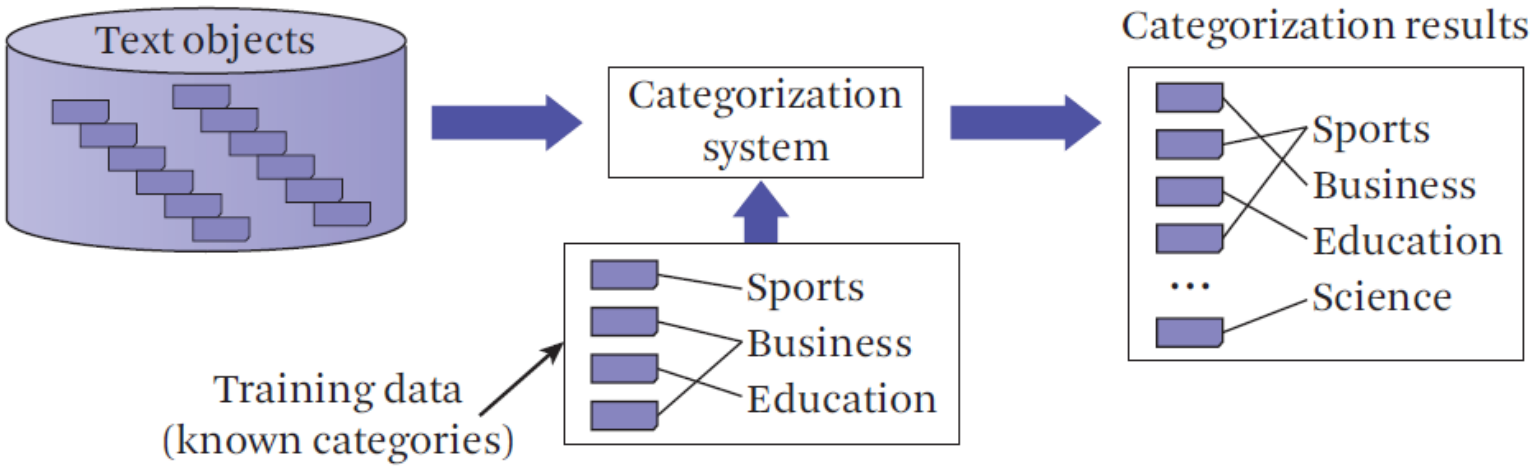
- A classificação de texto é baseado no modelo de aprendizado supervisionado. Este modelo é utilizado para classificar textos de acordo com um conjunto de categorias previamente definidos.

- Para realizar a classificação de texto, é necessário obter uma amostra para treinamento. Esta amostra será utilizada pelo modelo para "compreender" quais as características de cada texto correspondem a uma determinada categoria.

Fonte: ANANDARAJAN, MURUGAN, et al. **Practical Text Analytics:** Maximizing the Value of Text Data. 2019.

Classificação de Texto

- Por exemplo, uma agência de notícias pode ter interesse em classificar artigos em uma ou mais categorias como: tecnologia, esportes, política, entretenimento, etc.



Fonte: ZHAI, CHENGXIANG and MASSUNG, SEAN. **Text Data Management and Analysis**: A Practical Introduction to Information Retrieval and Text Mining. 2016.

Classificação de Texto

- **Multilabel Classification:** neste tipo de classificação há uma sobreposição entre as categorias previamente definidas e um texto pode pertencer a mais de uma categoria.

- **Single-label Classification:** neste tipo cada texto pertence a apenas uma categoria.

- **Binary Classification:** nesta classificação existem apenas duas categorias e o texto é classificado em uma delas.

- **Multi-class:** neste tipo, um texto é classificado em apenas uma classe, considerando a existência de três ou mais classes sem sobreposição.

Fontes: FELDMAN, RONEN and SANGER, JAMES. **The Text Mining Handbook**: Advanced Approaches in Analyzing Unstructured Data. 2007.
SOKOLOVA, MARINA and LAPALME, GUY. **A systematic analysis of performance measures for classification tasks**. 2009.

Aplicação de Classificação de Texto

- Classificação de email e filtragem de spam;
- Organização e filtragem de notícias;
- Organização de documentos;
- Análise de opiniões.

Fonte: AGGARWAL CHARU C. and ZHAI, CHENGXIANG . **Mining Text Data**. 2012.

Clusterização de Texto

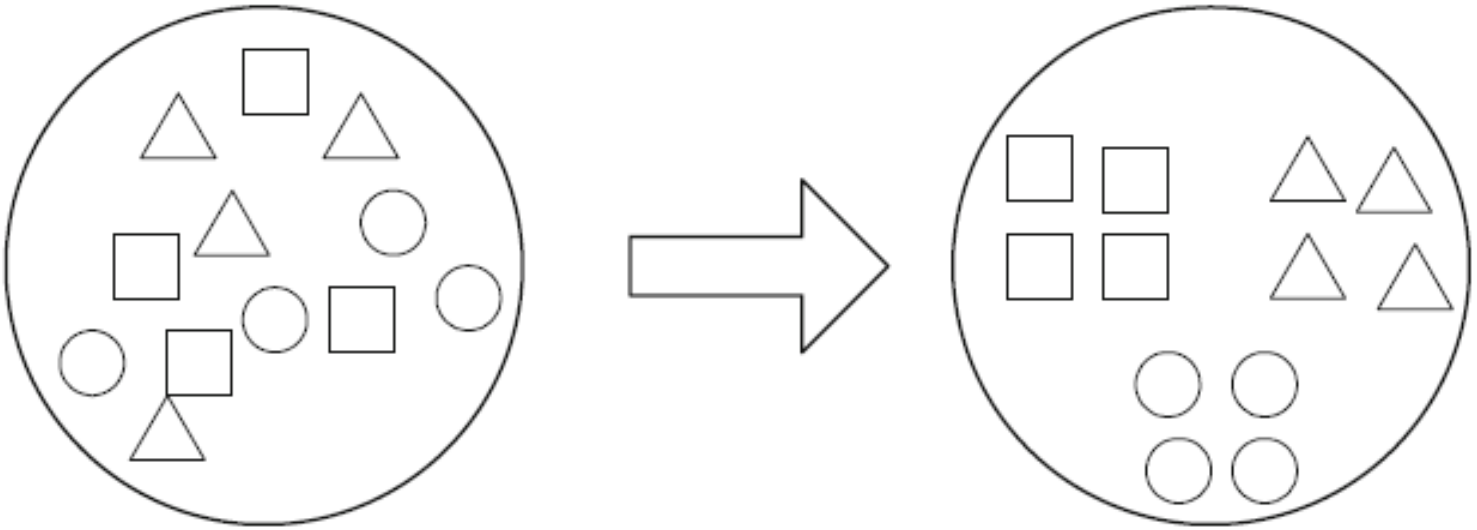
- Clusterização é o processo de segmentar um conjunto de itens de tipos diversos em subconjuntos contendo itens do mesmo tipo;

- Clusterização é uma tarefa normalmente baseada em aprendizado não-supervisionado, portanto não utiliza uma etapa de treinamento para decidir para qual subconjunto um item deve ser atribuído;

- A decisão de qual subconjunto atribuir um item é baseada em medidas de similaridade.

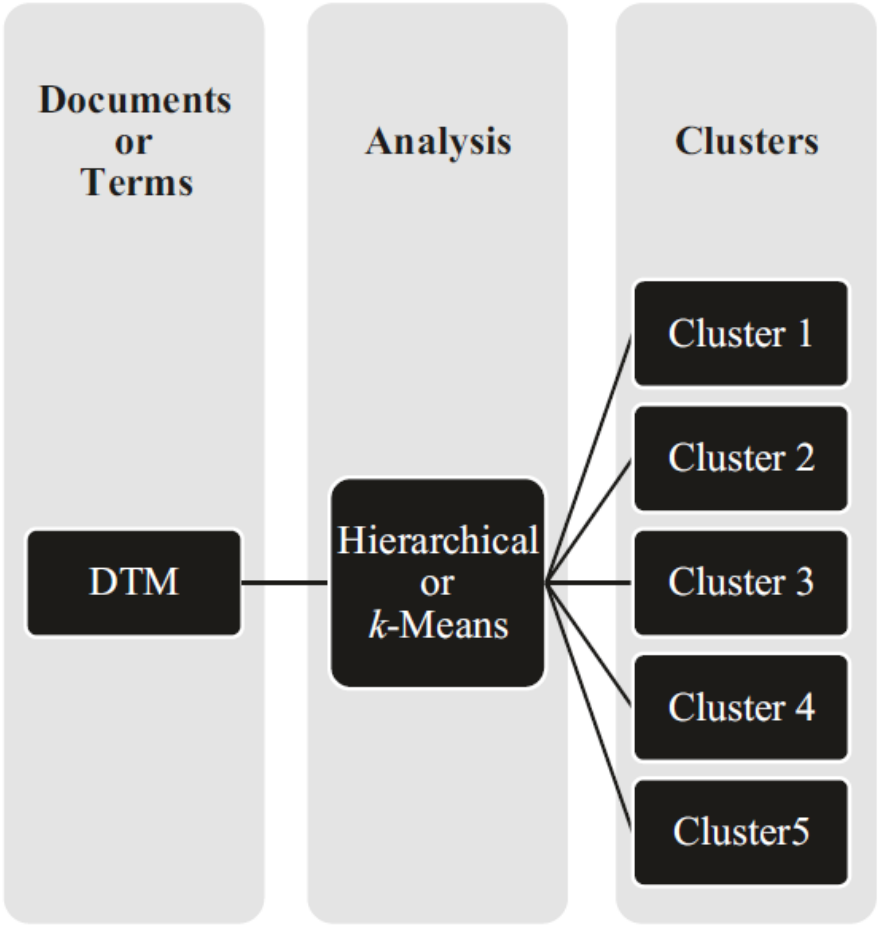
Fonte: JO, TAEHO. **Text Mining:** Concepts, Implementation, and Big Data Challenge. 2018.

Clusterização de Texto



Fonte: JO, TAEHO. **Text Mining:** Concepts, Implementation, and Big Data Challenge. 2018.

Clusterização de Texto



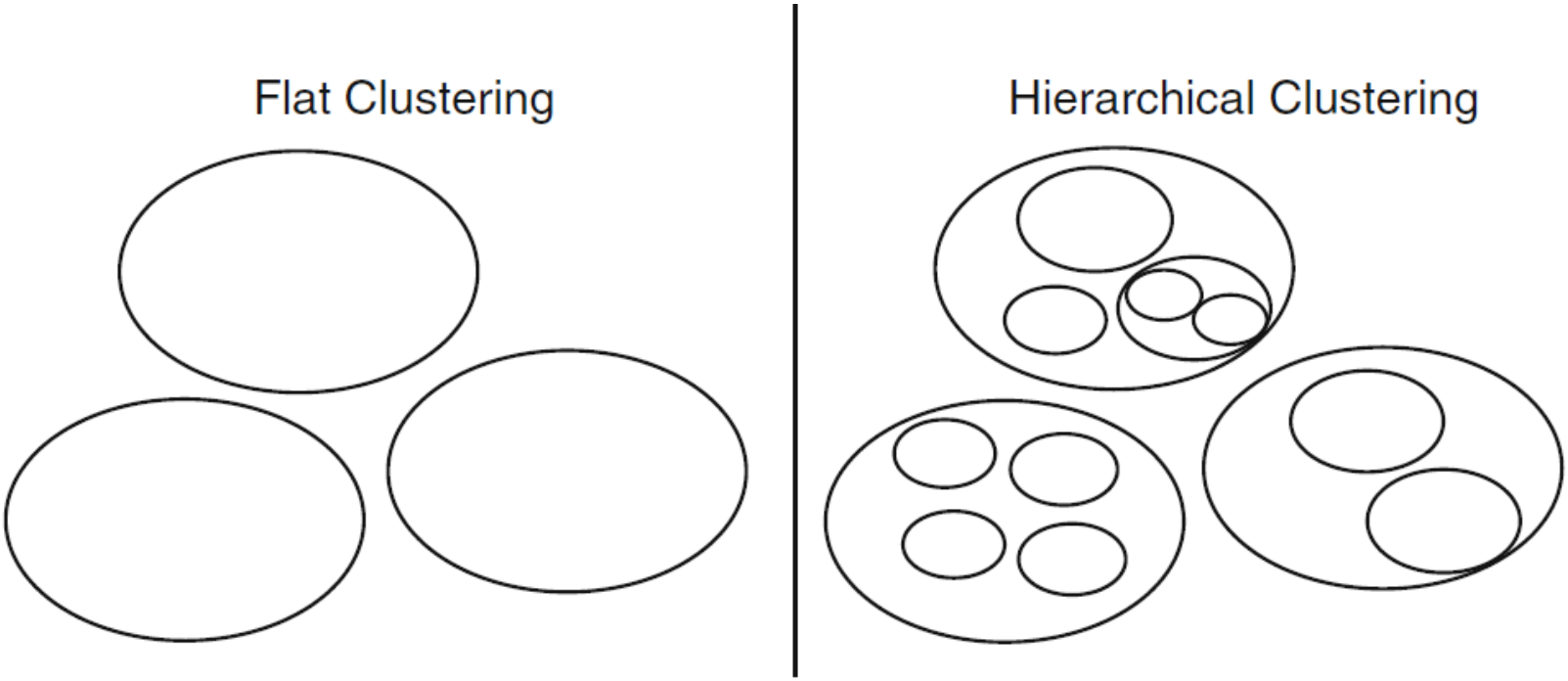
Fonte: ANANDARAJAN, MURUGAN, et al. **Practical Text Analytics:** Maximizing the Value of Text Data. 2019.

Clusterização de Texto

- **Hard clustering**: nesta abordagem, cada item é atribuído para um subconjunto específico;
- **Soft clustering**: abordagem que possibilita que um item seja atribuído para mais de um subconjunto;
- **Flat clustering**: abordagem que organiza todos clusters em uma lista única;
- **Hierarchical clustering**: abordagem que organiza os clusters em formato de árvore.

Fonte: JO, TAEHO. **Text Mining:** Concepts, Implementation, and Big Data Challenge. 2018.

Clusterização de Texto



Fonte: JO, TAEHO. **Text Mining:** Concepts, Implementation, and Big Data Challenge. 2018.

Aplicação de Clusterização de Texto

- Navegação e organização de documentos;
- Classificação de documentos;

Fonte: AGGARWAL CHARU C. and ZHAI, CHENGXIANG . **Mining Text Data**. 2012.