Universidade Federal do Ceará
Centro de Ciências
Departamento de Computação

# CKP9011 e CK0268 – Introdução à Ciência de Dados

**2024.1**

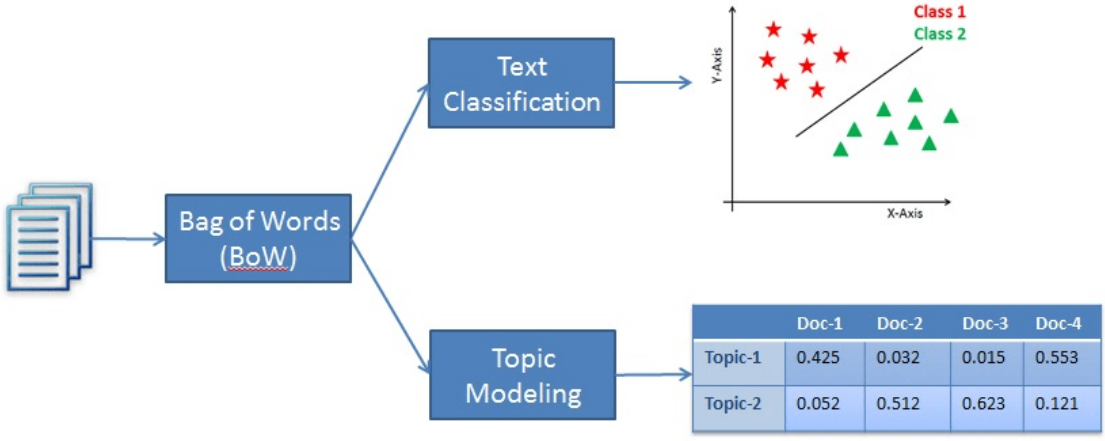
**Lista 9**

Exercício: Regras de Associação
Objetivos: Exercitar os conceitos referente a Algoritmos de Agrupamento.
Data da Entrega: 29/09/2024

OBS 1: Exercício Individual.
OBS 2: A entrega da lista deverá ser executada utilizando-se o SIGAA.

## Questão 1

A modelagem de tópicos é uma tarefa de Processamento de Linguagem Natural (PLN) que visa identificar estruturas semânticas, também chamadas de tópicos, compartilhadas entre textos que compõem um conjunto chamado corpus, a fim de determinar quais eventos, conceitos ou assuntos estão sendo discutidos. Em outras palavras, a modelagem de tópicos consiste em analisar textos e agrupá-los em tópicos conforme as palavras-chave presentes nesses textos. Essa tarefa parte da premissa de que cada texto pode ser representado por uma mistura de tópicos, sendo cada tópico composto por palavras que melhor o define. Vale ressaltar que a caracterização de tópicos é uma tarefa complexa, uma vez que requer um avaliador humano para identificar qual é o assunto a partir das palavras mais representativas de um determinado tópico.




| | Doc-1 | Doc-2 | Doc-3 | Doc-4 |
|---|---|---|---|---|
| Topic-1 | 0.425 | 0.032 | 0.015 | 0.553 |
| Topic-2 | 0.052 | 0.512 | 0.623 | 0.121 |

Entre os principais algoritmos dedicados à tarefa de modelagem de tópicos está o Latent Dirichlet Allocation (LDA), um algoritmo de aprendizado não supervisionado que tenta descrever um conjunto de textos como uma coleção de categorias (tópicos, grupos ou clusters) distintas. Os tópicos são aprendidos como uma distribuição de probabilidade sobre as palavras que ocorrem no corpus. Assim, o corpus pode ser descrito como uma combinação dos tópicos encontrados. Outros algoritmos importantes são: Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM), Pseudo-document based Topic Model (PTM) e o BERTopic.

No nosso caso particular, o corpus é formado por um conjunto de mensagens compartilhadas no WhatsApp. Utilizando os dados referente a postagens no WhatsApp, descubra os tópicos presentes nessas mensagens.

A avaliação experimental deverá considerar:

a) O algoritmo de modelagem de tópicos: LDA, GSDMM, PTM e BERTopic.

"A Educação, qualquer que seja ela, é sempre uma teoria do conhecimento posta em prática".

**Paulo Freire**