Sdco

/ CANALSANDECO

TEXT MINING

/CANALSANDECO

TEXT MINING

CONHECIDA TAMBÉM COMO MINERAÇÃO DE DADOS TEXTUAIS E SEMELHANTE À ANÁLISE TEXTUAL, REFERE-SE AO PROCESSO DE OBTENÇÃO DE INFORMAÇÕES IMPORTANTES DE UM TEXTO

CORPUS

COLEÇÃO DE TEXTOS A SEREM

ANALISADOS EM ***TEXT MINING***

BAG OF WORDS

REPRESENTAÇÃO SIMPLIFICADORA

USADA NO PROCESSAMENTO DE

LINGUAGEM NATURAL E RECUPERAÇÃO

DE INFORMAÇÃO (IR). TAMBÉM

CONHECIDO COMO O MODELO DE ESPAÇO

VETORIAL

BAG OF WORDS

NESSE MODELO, UM TEXTO (COMO UMA FRASE OU UM DOCUMENTO) É REPRESENTADO COMO O “SACO” DE SUAS PALAVRAS, DESCONSIDERANDO A GRAMÁTICA E ATÉ A ORDEM DAS PALAVRAS, MAS MANTENDO A MULTIPLICIDADE

BAG OF WORDS

O MODELO BAG-OF-WORDS É COMUMENTE USADO EM MÉTODOS DE CLASSIFICAÇÃO DE DOCUMENTOS ONDE A OCORRÊNCIA (FREQUÊNCIA) DE CADA PALAVRA É USADA COMO UMA CARACTERÍSTICA PARA TREINAR UM CLASSIFICADOR

BAG OF WORDS

EXEMPLOS:

- EU TE AMO
- AMO TE ENCONTRAR
- FALA SÉRIO

BAG OF WORDS

EXEMPLOS:

- T1 = EU TE AMO
- T2 = AMO TE ENCONTRAR
- T3 = FALA SÉRIO

	TE	EU	AMO	ENCONTRAR	FALA	SERIO
T1	1	1	1	0	0	0
T2	1	0	1	1	0	0
T3	0	0	0	0	1	1

DISTÂNCIA ENTRE TEXTOS

	TE	EU	AMO	ENCONTRAR	FALA	SERIO
T1	1	1	1	0	0	0
T2	1	0	1	1	0	0
T3	0	0	0	0	1	1

$$DE(x,y) = \sqrt{\sum_{i}^{p}(x_i - y_i)^2}$$