Exercício

Classificação de Texto usando o método Naive Bayes

Considere a seguinte situação. Temos 6 documentos D0 ... D5 como conjunto de exemplos de treinamento. Suponha que foram extraídas e consideradas somente 6 palavras de todos os documentos, de modo que estas formam o vocabulário para o problema em questão. Existem duas classes (categorias) de documentos: "terrorismo" e "diversão". Esses documentos foram preprocesssados e são apresentados na tabela a seguir. Os números na tabela correspondem à frequência de cada palavra nos documentos. Por exemplo, a palavra "matar" aparece 2 vezes no documento D0.

Documento	***matar***	***bomba***	***sequestro***	***música***	***cinema***	***TV***	**Classe (V)**
D0	2	1	3	0	0	1	terrorismo
D1	1	1	1	0	0	0	terrorismo
D2	1	1	2	0	1	0	terrorismo
D3	0	1	0	2	1	1	diversão
D4	0	0	1	1	1	0	diversão
D5	0	0	0	2	2	0	diversão

a) Calcule os seguintes termos, conforme ocorre na fase de apresendizagem do Naive Bayes:

i) |Vocabulario|

ii) $P(v_j)$: probabilidades a priori para cada classe

iii) n_i: número total de palavras (incluindo repetições) em cada classe

iv) $P(palavra_i \mid v_j)$: termos de probabilidade condicional de ocorrência das palavras, dada uma classe. Obs: para evitar o problema das frequências iguais a zero, utilize a estimativa de Laplace, assumindo a distribuição uniforme para a ocorrência de todas as palavras do vocabulário.

b) Classifique um novo documento de teste Dt, com a seguintes ocorrências de palavras:

Documento	***matar***	***bomba***	***sequestro***	***música***	***cinema***	***TV***	**Classe (V)**
Dt	2	1	2	0	0	1	?