

## Processamento de Linguagem Natural – Teste de avaliação

Nome completo:		RA:
----------------	--	-----

1. Considere o seguinte corpus (conjunto de treinamento) contendo duas classes ( $c_1$ ='pos' e  $c_2$ ='neg') associadas a um sistema de registro de descrições sobre filmes.

Classe	Texto
neg	just plain boring
neg	entirely predictable and lacks energy
neg	no surprises and very few laughs
pos	very powerful
pos	the most fun film of the summer

Calcule:

$P(c_1)$	
$P(c_2)$	
Vocabulário	
Tamanho do Vocabulário	

Classifique/categorize o documento **S** que contenha a frase: “**predictable with no fun**”, isto é, determine se **S** é da classe ‘pos’ ou ‘neg’.)

- Desconsidere as palavras que não estejam presente no vocabulário obtido no treinamento.
- Para facilitar: **Não** use logaritmo no cálculo da probabilidade.

2. Considere as seguintes probabilidades (likelihoods) para cada palavra obtida de um corpus de treinamento.

	<b>Classe 'Pos'</b>	<b>Classe 'Neg'</b>
I	0.09	0.16
always	0.07	0.06
like	0.29	0.06
foreign	0.04	0.15
films	0.08	0.11

Usando um classificador Bayesianos "ingênuo" a frase "I always like foreign films" deveria ser atribuída à classe Pos ou Neg?