Tarefa: aprender a classificar textos pela abordagem NB

A partir do conjunto de dados fornecidos (IMDB), construir um classificador de textos segundo a abordagem Naïve Bayes (NB) descrita nas transparências.

Para a geração do modelo, siga a rotina LEARN_NAIVE_BAYES_TEXT(*Examples*, *V*).

Para a inferência, siga a rotina CLASSIFY_NAIVE_BAYES_TEXT(*Doc*), **mas utilizando a versão logarítmica para o discriminante**.

Os dados estão contidos em duas pastas, "pos" e "neg", referentes às duas classes de interesse. Dentro de cada pasta uma há 25k arquivos, nomeados com seu número dentro de cada pasta: o primeiro arquivo tem o nome "0.txt", o segundo "1.txt" e assim sucessivamente, até o último que chama "24999.txt". A partir destes dados, cada aluno deverá considerar apenas os 24k arquivos de cada pasta, cujos índices estão relacionados na tabela abaixo.

Organize o conjunto de dados para permitir a validação cruzada de 10 vezes (10-fold-cv). Construa 10 classificadores NB para cada subconjunto de treinamento e teste o seu desempenho preditivo sobre o respectivo subconjunto de teste.

Apresente a matriz de confusão média correspondente ao **teste**, juntamente com os respectivos desvios padrões. A partir da matriz de confusão média, calcule as seguintes métricas:

Precisão, Taxa de Verdadeiros Positivos (TVP), Taxa de Falsos Positivos (TFP), Medida-F.

A tarefa deve ser realizada usando qualquer linguagem de programação, mas **não** utilizando funções específicas para aprendizado bayesiano, como as disponíveis na ferramenta Weka, por exemplo.

Entregar relatório contendo detalhes da implementação e os resultados obtidos, conforme previsto na tarefa, bem como os códigos fonte e executável. **Prazo de entrega: 09/05/2016.**

Nome	positivos	negativos
Abner	0 – 23999	0 – 23999
Anderson	1 – 24000	1 – 24000
Arthur Ribacki	2 – 24001	2 – 24001
Artur Brum	3 – 24002	3 – 24002
Bruno Ferreira	4 –24003	4 – 24003
Bruno Marques	5 – 24004	5 – 24004
Carlos	6 – 24005	6 – 24005
Cristina	7 – 24006	7 – 24006
Diego	8 – 24007	8 – 24007
Fabian	9 – 24008	9 – 24008
Fabiano	10 – 24009	10 – 24009
Fabio	11 – 24010	11 – 24010
Frederico	12 – 24011	12 – 24011
Guilherme	13 – 24012	13 – 24012
Jéssica	14 – 24013	14 – 24013

Nome	positivos	negativos
Joaquim	15 – 24014	15 – 24014
Jorge	16 – 24015	16 – 24015
Liza	17 – 24016	17 – 24016
Lucas	18 – 24017	18 – 24017
Marcelo	19 – 24018	19 – 24018
Miller	20 – 24019	20 – 24019
Paula	21 – 24020	21 – 24020
Paulo	22 – 24021	22 – 24021
Rafael Machado	23 – 24022	23 – 24022
Rafael Gonzalez	24 – 24023	24 – 24023
Raul	25 – 24024	25 – 24024
Renan	26 – 24025	26 – 24025
Rogrigo	27 – 24026	27 – 24026
Rogers	28 – 24027	28 – 24027
Tiago	29 – 24028	29 – 24028
Vinicius	30 – 24029	30 – 24029