Exercício 6

Data de entrega: 16/11, as 7:00 (da manha).

O diretorio ex6 contem 2 zip files. Ambos contem 5000 textos, um por arquivo. O arquivo fileR.zip gera um diretorio com os 5000 textos. O arquivo filesk.zip gera um diretorio de diretorios (com as classes) e os textos estao no subdiretorio apropriado. Esse parece ser o format mais util para a função sklearn.datasets.load_files

O arquivo category.tab contem a classe de cada documento.

Eu ja nao me lembro de onde sao os textos , mas sao parte de algum dataset de text mining com posts de tamanho medio de tecnologia.

Parte 1 - processamento de texto

Faça as tarefas usuais de processameno de textos:

- conversao de caracteres maiusculos para minusculos
- remoçao de pontuaçao
- remoçao de stop words
- steming dos termos
- remocao dos termos que aparecem em um so documento

Converta os textos processados acima em um bag of words no formato binario (0/1 se o termo aparece ou nao aparece no documento) e no formato de term frequency.

Em R o pacote tm faz a maioria se nao todo o preprocessamento e a conversao para uma matrix de termo/documento nos dois formatos. Este é um tutorial curto sobre o tm

Em Python, o sklearn tem funçoes para fazer a mesma coisa. Um tutorial sobre as funcoes no sklearn

Parte 2 - classificador multiclasse na matriz termo-documento original

O preprocessamento acima deve ser feito para todos os textos.

Divida o conjunto em 1000 documentos de teste e 4000 de treino aleatoriamente (pode ser estratificado ou nao).

Rode o naive bayes na matrix binaria. Qual a acuracia?

Rode o logistic regression na matrix binaria e de term frequency. Quais as acurácias.

Em Python use C=10000 em sklearn.linear_model.LogisticRegression para evitar que haja regularizacao.

Parte 3 - classificador multiclasse na matriz termo-documento reduzida

Rode o PCA e reduza o numero de dimensoes da matriz de term frequency para 99% da variancia original.

Rode pelo menos 2 algoritmos dentre SVM com RBF, gradient boosting e random forest na matrix com o numero de dimensoes reduzidas. Quais as acurácias?

Last modified: Mon Oct 31 19:08:16 BRST 2016