



DISCIPLINA: Programação e Inteligência Computacional

Professor: Felipe Appio

Essa atividade representa 3 pontos da prova P2 e poderá ser realizada em dupla.

Atividade P2

Análise de sentimento, mineração de opinião ou análise de subjetividade é uma das áreas dentro de Ciências da Computação que é responsável por classificar textos – mineração de texto. Em virtude do grande volume de informação que vem sendo gerado nas mídias sociais, vem recebendo bastante atenção nos últimos anos. Os textos são classificados geralmente em: positivo, negativo ou neutro. Por isso, é bastante comum utilizar métodos de Aprendizagem de Máquina Supervisionada para treinar o algoritmo.

Pede-se para escrever um software que, dada uma determinada oração, deverá ser capaz de rotular comentários ofensivos escritos em língua portuguesa brasileira (mineração/ de opinião).

Instruções

Há dois arquivos disponíveis no MOODLE:

- a) *treinamento.7z*
- b) *teste.7z*

O primeiro arquivo deverá ser utilizado, evidentemente, para treinar o algoritmo (utilizando validação cruzada). O segundo, deverá ser utilizado para avaliar o desempenho do algoritmo (experimento).

Entregáveis

A entrega deverá ser efetuada pelo MOODLE em um arquivo compactado contendo os seguintes arquivos:

- a) O código fonte (**somente os arquivos .java**).
- b) Um arquivo de texto contendo o nome dos acadêmicos.
- c) Um arquivo .pdf relatório técnico.

Para a entrega, a equipe deverá elaborar um **relatório técnico**, seguindo a metodologia científica da UDESC, contendo as seguintes seções, devidamente explicadas:

- a) Resumo
- b) Introdução
- c) Explicação do *design* e arquitetura da solução proposta



DISCIPLINA: Programação e Inteligência Computacional

Professor: Felipe Appio

d) Metodologia utilizada:

- a. Quantos *folds* foram utilizados para o treinamento e avaliação
 - i. Deverá ser utilizada ***k-fold cross-validation***.
- b. Qual métrica de desempenho foi utilizado para avaliar o modelo
 - i. Deverá ser devidamente justificada o porquê optou-se em utilizá-la
 - ii. Supondo que $k=10$, para cada *fold* haverá uma métrica. Pede-se que seja elaborada uma tabela contendo cada uma das métricas coletadas para cada *fold*.
 - iii. Construir a tabela verdade demonstrando as métricas coletadas **pelo *fold* com maior desempenho**. Assim, supondo $k=10$ e que o *fold* com a melhor métrica (seja ela qual for) seja o *fold* número 4, então deverá ser exibido uma matriz de confusão com os verdadeiros e falsos positivos/negativos.
- e) Todos os parâmetros do algoritmo (mesmo que em valores *default*) deverão ser devidamente explicados.
- f) Experimento utilizado para avaliar o classificador.
 - a. Os dados utilizados para o experimento estão no arquivo *teste.7z*

A entrega deverá ser efetuada até às **22h20 do dia 25/05** pelo **MOODLE** na atividade: “Atividade P2”.