APRENDIZADO DE MÁQUINA - KAGGLE - NLP

INTRODUÇÃO

Para aula de hoje, o nosso obetivo será participar de uma competição pela plataforma Kaggle! O Kaggle é uma plataforma online onde datasets são disponibilizados e competições são criadas para engajar a comunidade de ciência de dados a resolver diferentes problemas do mundo real.

Para trabalhar com bibliotecas como Pandas e ScikitLearn, é recomendado o uso de Jupyter Notebooks. Estes ambientes são interativos e permitem escrever e executar código Python em porções. Eles podem ser acessados nas diretamente nas máquinas dos laboratórios da PCURS ou através de portais online como o Google Colab e o próprio Kaggle.

OBJETIVO

O objetivo da atividade é desenvolver um modelo que consiga classificar se um determinado comentário em sobre uma revista (review) é negativo (bad) ou positivo (good). Geralmente quando temos dados textuais ligados a um sentimento, tratamos esse problema como sendo uma tarefa específica de classificação chamada de Análise de Sentimentos. Para essa finalidade, utilize da biblioteca ScikitLearn para instanciar os modelos vistos em aula (KNN, Naïve Bayes e Árvores de Decisão) e da biblioteca Pandas para ler os dados como uma tabela.

- Análise de Sentimentos

O cnojunto de dados disponibilizado possui três arquivos, um contendo os dados de treinamento (anotados), um como o conjunto de teste (sem as anotações) e um arquivo de exemplo para sobmissão. Os testes e modelagem devem ser realizados utilizando o conjunto de treinamento, o conjunto de teste serve apenas para gerar o arquivo de submissão.

Ao gerar o arquivo, você deve fazer a submissão do mesmo no link da competição. Os resultados serão avaliados automaticamente e um valor de escore será atribiudo.

A atividade pode ser realizada individualmente ou em grupos de até 3 pessoas.

Utilize os notebooks de exemplo disponibilizados no Moodle (um para cada model visto em aula) para saber quais bibliotecas podem ser utilizadas para analisar, visualizar e trabalhar com os dados. O formato de notebook é recomendado: localmente ou pelo Google Colab e Kaggle, que fornecem plataformas.

PONTOS EXTRA

Aqueles que conseguirem atingir mais de 87% de acurácia no segundo desafio receberão 0.5 pontos na primeira prova da disciplina. Os pontos a mais só irão valer se o aluno/grupo postar o notebook utilizado para gerar o resultado no fórum disponibilizado no Moodle e seja possível reproduzir os resultados atingidos.

Apenas um aluno do grupo precisa fazer a submissão para contabilizar a presença e o acréscimo na nota da prova.