Bootcamp: Analista de Dados

Trabalho Prático

Módulo 3: Técnicas para Análise de Dados

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Praticar os conceitos de análise de textos vistos nas aulas teóricas e práticas.
2. Usar o KNIME como uma plataforma para montar o workflow de análise de dados.

Enunciado

O presente trabalho tem como objetivo fazer a análise de sentimentos de reviews de usuários aos filmes listados no IMDB. O objetivo dessa análise é elaborar um modelo que seja eficiente para classificar o sentimento de usuários em positivo ou negativo.

Para isso, o engenheiro de dados forneceu a você uma base de dados com os reviews dos usuários e a classificação de cada um.

Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Fazer download da base de dados neste link.
2. Abrir o arquivo com o nó File Reader.
3. Clique na aba encoding e escolha a opção UTF-8. Clique em OK e carregue os arquivos.
4. Fazer a estratificação dos dados usando o nó Partitioning com as seguintes opções:
    a. Absolute: 1000
    b. Stratified Sampling: Sentiment
    c. Random Seed: 1
5. Crie os documentos usando o nó Strings to Document e a opção OpenNLP Simple Tokenizer como tokenizador. Os documentos devem ser criados com a coluna text_pt e a opção “use categories from column” deve estar marcada selecionando a coluna “sentiment”.
6. Repita as etapas de processamento e análise de texto vistas nas aulas práticas para responder às perguntas desse desafio, mas com alguns pontos de ATENÇÃO:
    a. Mantenha a coluna ID no seu dataset. Ela será importante para a resolução de algumas questões.
    b. Após a execução do nó “Number Filter” o dataset deve estar assim:

Row ID	id	Document
Row109	110	""
Row182	183	""
Row185	186	""
Row215	216	""
Row336	337	""

    c. Para executar as predições do ZeroR e do LibSVM (3.7) é preciso remover a coluna document do dataset antes.
    d. O trabalho foi realizado na versão 4.6.3 do KNIME.

Qualquer dúvida extra, o tutor está disponível nos fóruns. Bom desafio e divirta-se!