Universidade Federal do Paraná
Departamento de Informática
INFO 7004 – Aprendizagem de Máquina

Especificação do Trabalho

Considere a base IMDB disponível na página da disciplina, a qual é composta de duas classes e contém 100000 registros.

1) Divida a base de dados da seguinte forma:
    a. 1 a 25000: Treinamento
    b. 25001 a 50000: Validação
    c. 50001 a 75000: Teste1
    d. 75001 a 100000: Teste 2
2) Extraia uma representação da sua escolha (Bag of Words, Word Embedding, etc).
3) Implemente o classificador kNN como discutido em sala de aula. O classificador deve ser executado em linha de comando e deve receber como parâmetros os arquivos de treinamento, teste e o valor de k (por exemplo, knn train test 3). A saída deve ser a taxa de reconhecimento e a matriz de confusão.
4) Avalie diferentes valores de k e métricas de distância na base de validação. Verifique o desempenho das suas escolhas nas bases de teste.
5) Escreva uma relatório reportando seus experimentos.