ACH2118 Introdução ao Processamento de Língua Natural

Semestre 2023-2 - Exercício prático 2: classificação de faixa etária a partir de requisições na plataforma eSIC

1. O objetivo do trabalho é implementar um classificador de faixa etária (a1,a2,a3,a4) a partir de textos de solicitações publicadas na plataforma eSIC. com base no conjunto de dados disponível na atividade.
2. O EP consiste de uma tarefa de classificação de 4 classes representando a faixa etária do auto do texto fornecido.
3. O trabalho é desenvolvido em duas partes: a primeira entrega consiste em desenvolver o classificador e reportar resultados médios de medida de **acurácia** usando validação cruzada de 10 partições sobre o conjunto de dados de treinamento completo, a ser avaliado de acordo com os critérios discutidos a seguir. Nesta primeira parte, programadores Python devem usar

   result = cross_val_score(clf, x_train, y_train, scoring='accuracy'', cv=10).mean()

4. Os dados de treinamento são fornecidos em um arquivo ep2_pln-train.xlsx que pode ser aberto em Excel, Google Sheets ou similar, ou diretamente em Python. Não é fornecido vocabulário adicional, e portanto os futuros dados de teste incluirão palavras desconhecidas em tempo de treinamento.
5. A segunda parte da avaliação consiste em gerar as predições de cada classificador para o conjunto de teste que será fornecido após a primeira entrega.
6. O que entregar:
   - **Entrega 1** (treinamento): uma pasta ZIP contendo (a) o código completo da implementação; e (b) o relatório descritivo dos classificadores considerados, deixando explícito qual deles é o classificador final que está sendo proposto, juntamente com os resultados de validação cruzada *apenas do classificador final* sobre os dados de treinamento. Ou seja, apresente apenas UM classificador final e apenas UM valor final de acurácia, e não valores para múltiplos classificadores ou *folds*.

     Observe que este relatório é diferente do utilizado no EP1, e contém alguns itens adicionais.
   - **Entrega 2** (teste): uma pasta ZIP contendo (a) slides para apresentação (15 min) e (b) arquivo XLSX com as predições sobre os dados de teste, conforme modelo a ser divulgado, contendo uma coluna de rótulos exatamente na mesma ordem dos dados do conjunto de teste, sem nenhuma linha a mais ou a menos (pois neste caso não haveria como avaliar).

A entrega deve ser feita ANTES do prazo estipulado, certificando-se de que o arquivo realmente foi carregado com sucesso, que é a versão correta, e que não está corrompido.

7. Avaliação

- **Critérios de avaliação**: acurácia de teste (60%), método inovador / bem elaborado (20%), item "análise" do relatório (10%), demais itens do relatório (5%), apresentação em aula (5%).
- **Penalidade por baixo desempenho**: EPs cuja acurácia média seja inferior à média de um baseline de classe majoritária recebem nota 3,0 e não são avaliados. Assim, por exemplo, se a classe majoritária obtém 60% de acurácia no teste, a acurácia do modelo não pode ser inferior a 60%, já que isso tornaria o aprendizado de máquina sem razão de ser.
- **Penalidade por formato inválido**: 5% de desconto na nota final por entrega de arquivo de predições em formato diferente do esperado.
- **Ranking da turma**. As entregas válidas e completas (treino e teste) terão suas notas ajustadas pelo melhor resultado (que receberá nota dez). É desejável portanto que você obtenha não apenas um bom resultado, mas melhor do que o de seus colegas. ☺ Entregas incompletas não terão a nota ajustada.

O EP deve ser desenvolvido pelos mesmos indivíduos ou duplas do EP1.