



EACH

ACH2118 Introdução ao Processamento de Língua Natural

# Exercício prático 2

Ivandr  Paraboni  
EACH - USP

# Tarefa

- Usando os conjuntos de dados *ep2\_train.zip* contendo requisições de usuários da plataforma e-sic, desenvolver classificadores individuais para estimar a profissão (government, private, academic) dos usuários da plataforma eSIC.
- Em qualquer linguagem, usando qualquer técnica ou algoritmo.
- Objetivo é praticar o desenvolvimento de um modelo de classificação e textos simples em e esclarecer os critérios de avaliação

# Diferenças em relação ao EP1

- Um problema de três classes
- Dados reais
  - Desbalanceamento
  - Maior volume (mas você não precisa usar tudo)
- Foco na avaliação / análise dos resultados
  - Interpretação do modelo usando eli5 pode exigir classificador adicional como suporte

# Entregas

- Parte 1 (treino):
  - Relatório (modelo diferente do EP1, ver enunciado da tarefa) indicando todos os métodos testados e o resultado apenas do método escolhido como o melhor, mais a análise de características importantes (que pode exigir a criação de um modelo simples como reglog de tf-idf só para esse fim).
  - Código de todos os métodos relatados.
- Parte 2 (teste):
  - Arquivo CSV contendo apenas uma coluna com os rótulos de predição. Sem textos.
  - Slides da apresentação em formato PDF.

# Cr terios de avalia  o

- Acur cia de teste (70%).
- M todo inovador / bem elaborado (20%) – modelos Tf-Idf de regress o log stica, SVM, NB e afins n o pontuam neste quesito.
- Relat rio (5%), slides+apresenta  o (5%).

# Normalização e média final

- Trabalhos são ordenados por resultado global, e a nota máxima é ajustada para dez, e as demais na mesma proporção.

# Cronograma

- **Entrega 1 (treino) até 10/11**

Imediatamente após este prazo, serão divulgados os dados de teste.

- **Entrega 2 (teste) até 17/11**

- **Apresentação dos trabalhos em 19/11 e 26/11.**

Obrigado