

Disciplina: Inteligência Artificial
Professora: Cristiane Neri Nobre
Data de entrega: 21/06
Valor: 5 pontos EXTRAS

Questão 01 (2,5 pontos extras)

Para fazer esta questão, sugiro que estude o material sobre **Agrupamento** que está no CANVAS. Assista também os vídeos disponibilizados sobre este assunto. Está junto com os slides.

Além disso, acesse o notebook “**Kmeans.ipynb**” e a base de dados **Iris**, disponibilizados no CANVAS.

Com base nisso, faça:

1. Realize todos os pré-processamentos necessários: identificação de outlier, normalização e etc
2. Encontre os agrupamentos, discuta a qualidade destes agrupamentos (usando Silhouette e Elbow) e caracterize os agrupamentos obtidos
3. Explique como se obtém estas duas métricas, ou seja, explique as equações matemáticas.
4. Investigue, explique e implemente, pelo menos, mais 1 métrica de avaliação dos agrupamentos, diferentes das 2 anteriores
5. Uma vez que a base é classificada (setosa, virgínica e versicolor), mostre **visualmente** que instâncias foram agrupadas **incorretamente** pelo kmeans. Discuta os resultados.
6. Faça um pequeno relatório, bem simples, explicando todas as etapas de pré-processamento realizadas e explicando todos os resultados obtidos.

Coloque os links para os códigos produzidos ao final de cada questão

Questão 02 (2,5 pontos extras)

Para fazer esta questão, sugiro que estude o material sobre “**Mineração de texto**” que está no CANVAS. Considere também os arquivos **ReutersGrain-train.csv** e **ReutersGrain-test.csv** que estão no CANVAS.

Além disso, investigue os links abaixo:

1. <https://www.kaggle.com/code/leandrodoze/sentiment-analysis-in-portuguese/notebook>
2. <https://drive.google.com/drive/folders/1eBdTcmBA5Z7ALFEtfBEyBNAYNbYsJjfa>
3. https://www.tensorflow.org/text/guide/word_embeddings

E, finalmente, veja aqui também uma pequena função com parte do processamento de um texto.

```
def preprocess(text):  
  
    #Tokenização  
    tokens = word_tokenize(text.lower())  
  
    #Remoção de stopwords  
    stop_words = set(stopwords.words('english'))  
    tokens = [token for token in tokens if token not in stop_words]  
  
    #Stemming  
    tokens = [stemmer.stem(token) for token in tokens]  
  
    #Reune os tokens em uma string  
    processed_text = ' '.join(tokens)  
  
    return processed_text
```

O que você precisa fazer:

- 1) Carregar os dois arquivos que estão no CANVAS: **ReutersGrain-train.csv** e **ReutersGrain-test.csv**
E verificar o quanto o modelo acerta na classificação dos textos que falam ou não de grãos.

Faça o pré-processamento da base de dados. Adicione mais etapas de pré-processamento ao texto além dos citados acima (por exemplo: eliminação de marcadores, números e etc).

E utilize pelo menos **2 algoritmos** de aprendizado. Mostre os resultados e **disponibilize o código** implementado.