MBA em Ciência de Dados

Técnicas Avançadas de Captura e Tratamento de Dados

Avaliação Final

Luis Gustavo Nonato e **Moacir Antonelli Ponti**

Cemeai - ICMC/USP São Carlos

A avaliação vale 10 pontos. As questões de 1 a 4, caso respondidas da forma correta, já totalizam 10 pontos.

ATENÇÃO: Quando terminar de exame, você deve fazer um "upload" do notebook no _moodle_.

Questão 1 (2.5 pontos)

Considere o arquivo modcovid.pdf (disponível para download no moodle). Escreva um código para extrair o texto (ASCII) do arquivo PDF e escreva o texto extraído em um arquivo chamado `modcovid.txt` .

In [ ]:

Questão 2 (2.5 pontos)

Leia o arquivo `modcovid.txt` e realize as seguinte operações:

1. Extraia todas palavras contidas no arquivo e armazene em uma lista de palavras (utilize o método word_tokenize do pacote nltk.
2. Remova da lista de palavras todos os "palavras" que não sejam formadas exclusivamente de caracteres do alfabeto.
3. Quantas palavras com apenas 1 caractere sobraram na lista?

In [ ]:

Questão 3 (2.5 pontos)

Antes de começar, carregue o arquivo `artists.csv` e armazene em um pandas DataFrame.

In [1]:
```
import numpy as np
import pandas as pd

df = pd.read_csv("artists_mba.csv")
```

a) (0.5 pontos)

Crie um novo atributo no dataframe, chamado `birth` , pegando os 4 primeiros caracteres do atributo `years` e convertendo para inteiro.

Posteriormente, exiba o tipo do novo atributo e a estatística descritiva do novo atributo linhas usando a função `describe()`

In [ ]:
```

```

b) (1 ponto)

Execute uma função que identifique outliers com base no intervalo interquartil. Mostre as linhas referentes a outliers detectados por esse método no atributo `paintings` para valores para além de mais ou menos $2 \times IQR$.

In [ ]:
```

```

c) (1 ponto)

Crie um novo atributo numérico, codificando em inteiros o atributo `nationality` . Posteriormente, compute a correlação de Pearson entre esse novo atributo e o atributo `paintings`

In [ ]:
```

```

Questão 4 (2.5 pontos)

Dada uma imagem `painting.jpg` de uma pintura da qual não sabemos o artista, gostaríamos de fazer uma busca numa base de dados e recuperar obras similares. Para isso utilizaremos uma composição de descritores:

1. Histograma de cores (R, G, B) considerando 4 bins por canal de cor (total 12 características)
2. Descritor LBP utilizando raio 2.5 e 16 pontos (total 18 características).

Concatene esses dois descritores e use-o como descritor de cor e textura da imagem. Faça uma busca no diretório `paintings` , retornando as 5 imagens mais similares de acordo com esse descritor e a distância Euclidiana. Exiba a imagem de consulta e também as 5 imagens retornadas, com seus nomes e valor da distância obtido.

In [3]:
```
# inclua os pacotes necessários e as funções necessárias
```

In [4]:
```
# inclua o código para carregar as imagens, gerar os vetore
s de características e obter as distâncias
```

In [5]:
```
# inclua o código para obter as 5 imagens mais próximas com
base nas distâncias computadas e exibi-las
```