

Trabalho prático 2 - Execução dos códigos

Julianny Favinha Donda
156059
julianny.favinha@gmail.com

José Henrique Ferreira Pinto
155976
joseh.fp@gmail.com

É necessário que as bibliotecas **numpy**, **matplotlib** e **sklearn** estejam instaladas.

O Moodle aceita apenas 10MB como entrega. Por isso, é necessário fazer o download da pasta

<https://www.dropbox.com/s/ahkim9u103v0q9i/health-dataset.zip>

É necessário que a pasta **health-dataset** esteja no mesmo diretório que os programas listados a seguir. Os programas a seguir utilizam apenas os arquivos bags.csv e health.txt.

Kmeans.py

Este arquivo contém o uso do algoritmo Kmeans da biblioteca Sci-Kit Learn. Para executá-lo, digite no terminal

```
$ python3 Kmeans.py
```

A saída desse programa é o gráfico da inércia pelo número de clusters e, na saída padrão para cada número de clusters, o coeficiente de silhueta, a inércia e o tempo decorrido. Também gera dois arquivos texto, de nomes DictionaryOfClusters e MedoidsOfClusters, sendo que o primeiro contém, para cada cluster, as sentenças que pertencem ao mesmo e o segundo arquivo contém, para cada cluster, a sentença medóide e as duas sentenças mais próximas do medóide.

DBSCAN.py

Este arquivo contém o uso do algoritmo DBSCAN da biblioteca Sci-Kit Learn. Para executá-lo, digite no terminal

```
$ python3 DBSCAN.py
```

A saída desse programa é, na saída padrão, o epsilon e min_samples utilizado, o número de clusters formado, o número de observações que não foi considerada *outlier*, o coeficiente de silhueta e o tempo decorrido. Também gera um arquivo texto de nome DictionaryOfClusters que contém, para cada cluster, as sentenças que pertencem ao mesmo.