

## Analyzing the Large-Scale CelebFaces Attributes Image Dataset

This tutorial concerns with handling an image dataset. We will showcase how to load & explore the data, analyze the images, predict image attributes and build a face recognition pipeline.

### Exercise 1: Data Loading and Exploration

- Download the CelebFaces Attributes Dataset (CelebA) from <https://openbigdata.org/>. It suffices if you download the aligned zip file instead of the raw images.
- Identify the number of images and attributes in the dataset.
- Explore the distribution of binary attributes in the dataset. Identify the most common attributes and visualize their frequency in a bar plot.
- Check for any missing values in the dataset and determine the appropriate approach for handling them.
- Compute summary statistics for the landmark locations in the dataset.

### Exercise 2: Image Analysis

This exercise aims to analyze the images in the CelebA dataset using neural embeddings and dimensionality reduction.

- Create a list containing 1000 random image file paths symbolizing an image subset to do for the next steps.
- Use a pre-trained deep learning model (e.g., VGG16 or ResNet) to extract features from the images.
- Visualize the clusters using PCA and t-SNE to reduce dimensionality and plot the clusters in a 2D space.
- Implement a clustering algorithm (e.g., k-means) on the extracted features to group similar images together.

### Exercise 3: Attribute Prediction

In this exercise, we try to classify the CelebA attributes using our previously obtained feature embeddings.

- Split the dataset into training and testing sets.
- Select a suitable machine learning algorithm (e.g., logistic regression, random forest) and train it to predict binary attributes based on the image features.
- Evaluate the performance of the trained model using appropriate metrics such as accuracy, precision, recall, and F1-score.
- Experiment with different algorithms and hyperparameters to improve the model's performance.

## Exercise 4: Face Recognition

Another task that can be performed using the CelebA dataset is face detection (Is there a face?) and recognition (It's that face!).

- a) Build a basic face detection pipeline using OpenCV and a Haar cascade.

The next exercises focus on creating a face recognition pipeline using eigenfaces.

- b) Identify the top 10 most common celebrities in the dataset and get 30 file paths from each.
- c) Prepare the data for training: 1. Remap the IDs to integer labels 0-9. 2. Do a train/testsplit, where the test set contains exactly 5 images from each of the top 10 celebrities. 3. Load the files as grayscale images to memory as a numpy array.
- d) Create a face recognizer with OpenCV and the eigenfaces algorithm. Hint: You'll need the opencv-contrib-python package.
- e) Create a face recognizer with scikit-learn and a SVM backbone.