

## Practice exercises 2

You will explore different aspects related to data: quality, loading, and licenses. This document is an overview of the practice exercises. Detailed instructions and explanations are provided in the accompanying Jupiter notebook.

### Investigating data-induced bias in the models

We explore the bias within and across text encoders (SentenceBert, CLIP, GloVe) in the context of job gender bias. The bias in these models is caused by the training data on which the encoders were trained. You are provided with a list of 100 job titles (obtained as output of the ChatGPT 3.5 Prompt: "Provide a list of 100 jobs"). You will learn how to embed the job titles and words "man" and "woman", and compute their similarity in the embedding space.

**2.1** Compute the difference between the cosine similarity of the word "man" and the job title, and "woman" and the job titles.

**2.2** Compute the variance and the mean of the delta values per model.

### Familiarizing with the dataset and image augmentation

**2.3** Learn to load an image dataset, create a dataset loader and image batches.

**2.4** Visualize the images and the class distribution.

**2.5** Use augmentation techniques to improve the diversity of the data for training.

### Getting to know data license

**2.6** Familiarize yourself with data licensing: check the initial and current images' license, and count images changed the license from *Attribution* to *All Rights Reserved*.