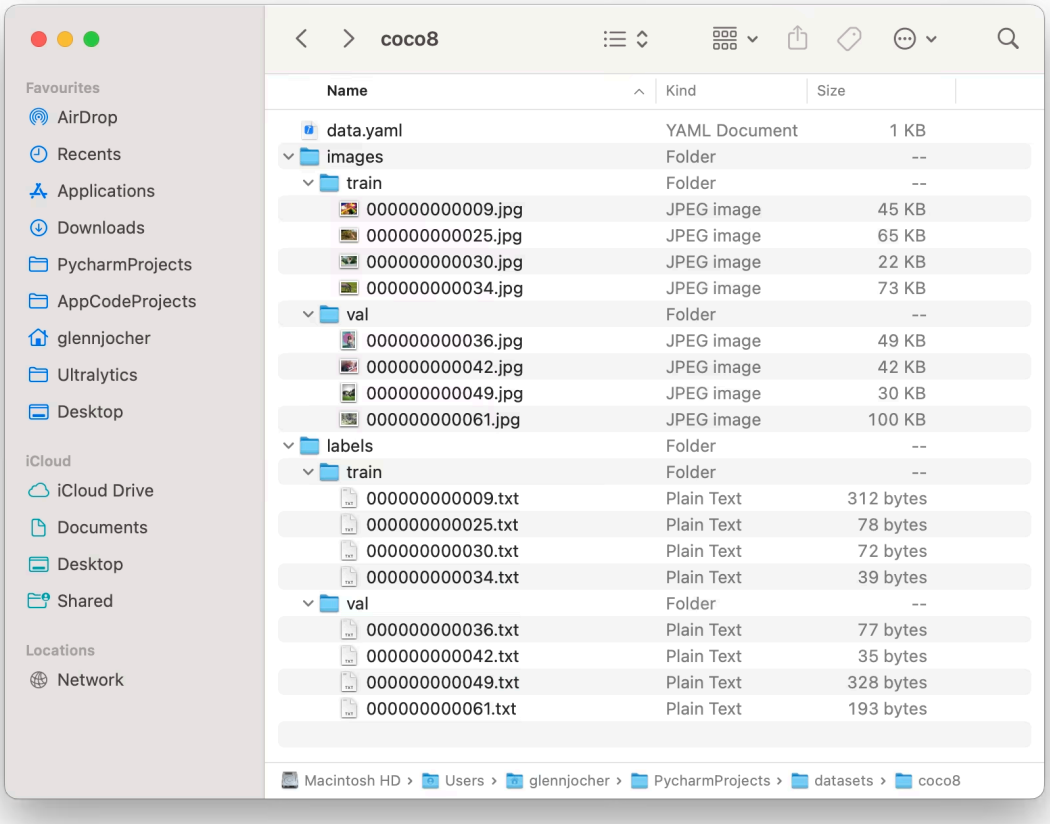
2024-10-22

Практика 3. Сбор изображений с веб-сайта

1. Написать функцию, которая принимает на вход URL (строку), а на выходе у нее — ZIP-архив со всеми изображениями, которые можно получить по этому URL (упаковать через модуль `zipfile` стандартной библиотеки).
2. Написать функцию, которая принимает на вход `ZIP-архив с изображениями`, `путь к директории в файловой системе` и `пропорции датасета`, после чего создает по указанному пути структуру директорий для набора данных для тренировки YOLOv8 (см. пример в Object Detection Datasets Overview - Ultralytics YOLO Docs) и размещает в этих директориях изображения из ZIP-архива в правильных пропорциях (например, `train` / `val` — 80%/20%); каждый вызов функции распределяет изображения случайно



3. Предусмотреть, что изображения могут быть в разных форматах (особенно обратите внимание на `.webp`) и не всегда лежать только в тэге `img` (например, в случае `picture` нужно загрузить именно `source` https://developer.mozilla.org/en-US/docs/Web/HTML/Element/picture)

4. Предусмотреть при сборе задание минимального и максимального разрешения изображения (чтобы не собирать совсем уж маленькие или большие картинки)
5. Ограничиться только сайтами, где контент передается со стороны сервера (т.е. где не нужны инструменты вроде Selenium). Хороший пример — сгрузить все картинки со страницы поисковой выдачи DuckDuckGo или Яндекса

Примеры ссылок, откуда можно подтянуть картинки

- https://drom.ru
- https://unsplash.com/
- (e:ltr cn>=452) or (e:ltc cn>=411) · Scryfall Magic The Gathering Search
- https://sipi.usc.edu/database/
- https://www.jstor.org/images#classifications
- Louvre site des collections
- https://www.npg.org.uk/collections/about/photographs-collection/photographic-holdings-collections/
- https://www.npg.org.uk/collections/about/photographs-collection

```
import glob
import zipfile

with zipfile.ZipFile('images.zip', 'w') as zpf:
    for img_file in glob.glob('*.jpg'):
        zpf.write(img_file)
```