

Задание 2

Практическая работа

Тема:

Формирование данных для машинного обучения в формате NumPy на основе датасета S3DIS

Цель работы:

Научиться:

- Загружать облака точек из датасета S3DIS;
 - Формировать единый массив данных NumPy (таблицу), объединяющий признаки и метки классов;
 - Выполнять базовую предобработку: нормализацию координат и признаков;
 - Подготавливать данные для подачи в модели машинного обучения (например PyTorch).
-

Задача:

Сформировать массив данных для каждого помещения всех областей, где каждая строка соответствует одной точке, а столбцы представляют:

X	Y	Z	R	G	B	label
---	---	---	---	---	---	-------

Реализовать выше изложенное в виде отдельной функции.

Исходные данные:

- Путь к каталогу помещения
 - Выходной файл содержит: X Y Z R G B label
-

Инструкции к выполнению:

1. Загрузка и разбор данных:

- Считать данные с помощью `numpy.loadtxt()` или `np.genfromtxt()`;
- Выделить все нужные признаки;
- Преобразовать цвета R, G, B в диапазон [0, 1].

2. Формирование таблицы:

- Объединить все признаки и метку в одну таблицу с помощью `np.hstack()`;

3. Предобработка:

- Выполнить нормализацию координат X, Y, Z (например, центрирование и масштабирование);
- (по желанию) Нормализовать интенсивность.

4. Сохранение данных:

- Сохранить таблицу `dataset` в формате `.npy`:

```
np.save('semantic3d_dataset.npy', dataset)
```

- Предусмотреть сохранение в форматах `.txt` и `.h5`

5. Реализовать функцию визуализации распределения меток по файлу

- Построить простую визуализацию распределения меток (в виде гистограммы).

Что нужно сдать:

1. Готовый Python-скрипт или Jupyter Notebook;
2. Файл `s3dis_dataset.npy`;
3. Скриншот/вывод содержимого первых 5 строк массива;
4. Ответы на контрольные вопросы.

Вот подборка **контрольных вопросов** по теме **S3DIS** и подготовке данных в формате **NumPy** для машинного обучения. Эти вопросы можно использовать для самопроверки, устного опроса или включить в итоговый тест по теме.

Контрольные вопросы по теме S3DIS и работе с данными

Общие вопросы по датасету S3DIS:

- 1. Что такое S3DIS и для чего он используется?**
- 2. Какие типы данных содержит S3DIS?**
- 3. Сколько пространств (Areas) включает датасет?**
- 4. Что представляют собой строки в .txt файлах S3DIS? Опишите формат.**
- 5. Какие задачи машинного обучения можно решать с помощью S3DIS?**
- 6. Какие классы (объекты) включены в разметку S3DIS? Приведите примеры.**

Вопросы по подготовке данных:

- 7. Какие признаки можно извлекать из данных S3DIS для обучения модели?**
- 8. Что такое нормализация координат, и зачем она нужна при подготовке 3D данных?**
- 9. Почему важно сохранять данные в формате NumPy (.npy , .npz) при работе с ML-моделями?**
- 10. Как разделить данные на входные признаки и метки классов?**
- 11. Чем отличается np.save() от np.savez() и np.savez_compressed() ?**
- 12. Какие потенциальные проблемы могут возникнуть при чтении и обработке S3DIS?**