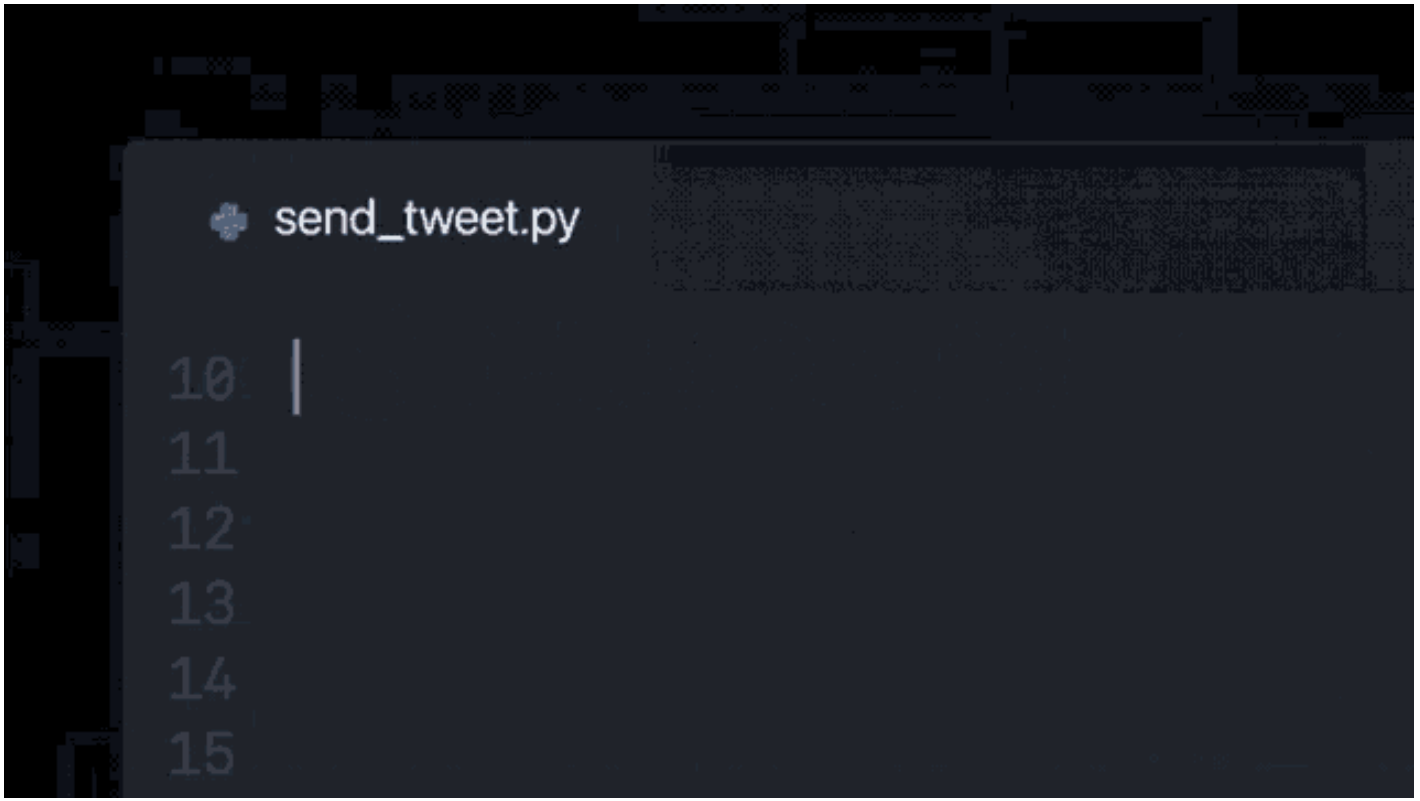
# Machine Learning

## Lab2

**Fall 2023**

**Instructor: Xiaodong Gu**

# 基于GPT2的程序生成



# 基于GPT2的程序生成

- 任务:

  参考课堂讲解的 GPT2 finetuning 以及示范代码实现程序自动生成系统并进行调参实验。最后对生成的程序样例进行展示。

- 数据:

  https://github.com/wangcongcong123/auto_coding/tree/master/dataset 也可采用其他数据集如CodeNet (https://github.com/IBM/Project_CodeNet)或自行收集数据

- 参考代码：

  https://github.com/wangcongcong123/auto_coding

有问题请联系助教林雅岚 linyalan@sjtu.edu.cn

# AI编程

- **提交：**SID_NAME.zip
  - 代码及运行说明
  - 实验报告。包括但不仅限于系统设计、训练过程（如loss曲线）、调参实验及结果（不同参数下的perplexity等指标）、样例展示等。**训练结果指标仅作为一项参考，不是主要的评价标准！**

    系统设计

    模型设计

    训练方法

    实验结果

    训练过程 （如loss曲线）

    调参实验及结果 （如模型在不同超参数下的精确度）

    生成代码展示

- 截止日期：2023年11月27日

# AI编程

注意事项：

1. 需要配置开发环境, 如：PyCharm+Anaconda, python=3.7, torch=1.10.1
2. 如果使用 gpt2 训练较慢, 可以适当减小 seq_len、使用 distilgpt2 (https://huggingface.co/distilgpt2)
3. 读取示例数据的代码已经给出,可以参考此代码构建自己的 Dataset 类
4. DataLoader 中可以设置如 num_workers=4 提高计算效率
5. 在架构较新的GPU上 (18 年后发布的NIVDIA GPU), 可以使用混合精度训练提高效率
6. 为了减轻工作量，可以参考开源代码（报告中注明来源），但要有自己的发挥。

# AI编程

- 评分：综合评价功能、质量和工作量

**功能：**

代码无法运行

能完成功能、鼓励举一反三、尝试新方法

**质量：**

生成内容无意义、报告质量低

生成可读程序、报告完整思路清晰

**工作量：**

直接提交示例代码或完全照搬开源代码

显示出对代码有理解、重构、或改进