代码相似度评估研究

背景需求：

在平时工作中，需要评估两段代码的相似度。

思路：

本研究提出了一种简易的评估模型，使用 pytorch 实现对两段代码进行分析，并给出了评估值计算表达式。

操作步骤：

对输入的数据进行清洗，并字典化，结合 pytorch 余弦相似度计算方法，评估出代码的相似度。

评估指标：

$$S = \max(f(x), g(x))$$

其中，f(x)为字典评估值，g(x)为余弦相似度评估值，取两者的最大值为最终评估结果。

若结果 s 为 1，则完全相同，否则，比 1 小，且值越小，不相同度越高。

收获：

通过查阅相关资料，熟悉了部分 pytorch 接口的使用方法。

不足之处：

暂时只能处理英文代码；

由于时间有限，没有合适的数据集，暂未使用训练生成模型的方法评估。