题目4

基于FFNN的方案

- 输入和输出表示：
  - 输入：将影评文本的每个词表示为数值向量，可以使用独热模型，也可以使用已有的word2vec等词嵌入模型，从而映射为一个向量表示，方便输入网络。
  - 输出：电影的星级评定，可以视为一个5分类问题，输出概率最大的类别。
- 网络结构：
  - [图1]
  -
  -
  -
  -
  - 输入层：由于FFNN是传统的神经网络结构，只能处理固定长度的输入向量，因此我们不能一下处理所有词。可以借助n-gram构造很多个定长的输入，比如使用滑动窗口来读取所有连续的n个词作为输入，拼接每个词的向量得到输入向量，同时保留了上下文信息。
  - 隐藏层：可以包含多个隐藏层，每个隐藏层使用激活函数（如ReLU）引入非线性变换。
  - 输出层：对于每次的输入，使用 softmax 函数输出分类的概率结果，并对一个文本内的所有输入的分类概率求积并归一化，从而得到文本的分类结果。
- 训练过程：
  - 数据集建立：可以在前一阶段要求部分用户写完影评文本后手动给出评分星级，构成数据集，再划分来构建训练集和验证集。
  - 数据预处理：对影评文本进行分词，并使用选定的模型来生成输入向量。
  - 训练：使用训练集对 FFNN 进行训练，采用与正确标签的交叉熵损失函数作为评价指标，对于从输入层到隐藏层的权重等参数，使用梯度下降算法或其变种来最小化损失函数。最后比较验证集上的准确率，调整超参数，直到满足一定的正确率要求。
- 推理过程：
  - 对于一个新的影评文本，首先进行分词，然后使用训练好的模型来预测其星级评定。FFNN的逻辑可以简单的理解为，筛选出一些关键词或词组（n-gram）所形成的特征维度，比如“好看”、“不好看”、“喜欢”、“不喜欢”等，然后根据这些维度上测试文本的向量大小即关键词的出现情况来判断影评的星级。