

以下是本门课程第二项编程作业的具体内容及要求。

## 作业二：文本分类

### 1.1 目标

- 实现一种基于 RNN 的模型（可以是 vanilla RNN、LSTM、GRU 等）以及一种基于 CNN 的模型进行文本分类任务，具体为对英文篇章级文本进行分类。评估指标是 Accuracy。由自己实现相关的评估指标。
- 要求个人独立完成作业二，推荐使用已有的神经网络实现的库函数或者自动求导的框架，包括但不限于 Pytorch, Tensorflow 等，鼓励在框架的神经网络组件基础上自己实现传统或者创新的神经网络架构。不能抄袭现有的开源实现和别人的代码，实验报告不能抄袭。
- 要求对比基于 RNN 的模型和 CNN 模型的效果，鼓励具有创新性的神经网络模型设计。

### 1.2 数据

- 训练集和验证集为 train.csv 和 dev.csv，两者格式相同。
- 测试数据为 test.csv，要求按测试数据的格式，将占位的第一列标签替换为预测标签，占位标签不代表真实标签，真实标签不公开，作业上交后评测分数将由脚本自动评测，所以请务必按照测试数据的格式来处理作业格式。

### 1.3 提交

- 截止日期：1 月 10 日 23:59 分
- 提交内容：作业需要压缩为压缩包提交，作业文件名命名方式为：学号-姓名（如 1200011111-张三）。必须包含 code 文件夹（内含所有源码），test.csv（测试集输出，其结果应为所实现的所有模型中验证集上效果最好的模型的预测结果）和 report.pdf（实验报告）。实验报告包括实验方法，实验设置和步骤以及取得的验证集实验结果。请务必按照提交格式进行提交，以免自动评测脚本出错。
- 提交方式：作业的提交方式为上传至教学网（教学网->自然语言处理导论->教学内容->编程作业二）。

### 1.4 评分

- 对报告的书写内容，格式是否规范，实现的原理以及工作量和代码风格进行评分，使用中英文均可，英文报告会考虑酌情加分。
- 对测试集效果进行排名评分，如果测试集格式错误导致脚本无法识别，需要手动修改格式或者重新提交才可以评测，会扣除一些分数。
- 鼓励具有创新性的神经网络设计。
- 鼓励在截止日期之前留一定余量开始提交，不建议压截止日期提交。
- 作业可多次提交，成绩以最后一次提交的效果和时间为准。