一． 作业简介

要求所有同学两两组队（不得多于两人），在分词/词性标注/命名实体识别三项任务中选择两项独立完成。其中**分词任务必选**，**词性标注和命名实体识别两项中任选一项**（多选有加分）。在给出的 trainset/devset/testset 上进行代码设计和调优，最终提交**测试集上的结果和实验报告**，并在实验报告中分析多次实验的结果及其原因，并对最终模型的错误类型进行分析。

二． 数据简介

1. trainset 文件夹中包含训练集数据，devset 文件夹中包含验证集数据，testset1 中包含测试集数据。
2. 文件夹中以_cws.txt 后缀的文件中包含的是分词任务的训练/验证/测试语料，其中词与词之间由空格分开；以_pos.txt 后缀的文件中包含的是词性标注的训练/验证/测试语料，词与词之间由空格分开，每个词与其对应词性由'/'分割开；以_ner.txt 后缀的文件中包含的是命名实体识别的训练/验证/测试语料，词与词之间由空格分开，每个 NER 开始位置词拥有'['，结束位置词拥有']'，方括号后的英文表示 NER 类型。
3. ner_evaluate.py/pos_evaluate.py 文件中包含 ner/pos 的官方端到端测试代码，用来进行测试和模型评价。分词任务测试代码较为简单，不建议同学使用端到端测试（可以使用简单代码直接判断每个汉字的分词标记，按 precise/recall/f1 进行计算）。

三． 任务引导

以上三个任务均需要学生动手完成数据预处理+模型设计+模型调优+实验报告撰写。由于以上三个任务都属于序列标注任务，所以数据预处理和模型设计部分有诸多相似之处。

参考模型：

1.机器学习模型：HMM，CRF，SVM 等分类模型（参考 python 包：sklearn）

2.深度学习模型：RNN，LSTM，LSTM+CRF 等（参考 python 包：tensorflow，pytorch 等）

参考代码格式，尽量分成以下四个文件，分别按以下功能设计：

1.dataset.py(预处理数据，输入是原始数据，输出是供模型使用的格式)

2.model.py(模型调用)

3.evaluate.py(评价结果，并给出最终 precision/recall/F1)

4.run.py(调用以上三个文件中函数或类，完成整个训练+测试+预测过程)

参考实验报告格式：

1.实验目的（阐明任务）

2.实验原理（描述模型）

3.实验内容（描述实验步骤）

4.实验结果（描述实验结果并分析原因）

5.误差分析（对 bad case 进行分析）

6.思考题

7.实验分工及感想

四． 评分标准

这次实验中评分标准分为基础分数+加分项+扣分项，其中基础分数按 100 分计，扣分和加分标准按照对应项计算，得到本次作业结果后与第一次作业结果最终加权求和，以得到本次课程成绩。

1.基础分数：

1.1.实验完成度（从实验报告中模型修正的次数【至少三次参数调整或模型架构变化】和误差分析【从错误案例角度分析至少一次】），占 60%

1.2.模型理解程度（对于自己使用的模型需要有详细的介绍和参数理解，并能对应模型的优缺点调参），占 20%

1.3.实验最终成绩（验证集的最终结果），占 20%

2.扣分标准：

2.1.抄袭论文或虚报实验结果，一经发现，直接按课程成绩 59 分论处

2.2.使用网络上他人代码或分词/pos/ner 工具，一经发现，直接按课程成绩 59 分论处

2.3.代码整洁度（有清晰的模块分割和接口设计），不足者扣 5-10 分

2.4.个人工作量明显不足者，扣除工作量不足部分占总工作量之比，区间在 20%-100%

2.5.论文格式混乱，提交数据有误，酌情扣除 5-10 分

2.6.态度不端正，及作业实在不堪入目者，直接按课程成绩 59 分论处

3.加分标准：

3.1.完成全部三个任务，整体成绩提升 5%-10%

3.2.最终结果在全班成绩中达 top-3，整体成绩提升 5%-10%

3.3.实验报告完成度高，结果分析有理有据，深刻理解模型，整体成绩提升 5%-10%

3.4.使用模型创新型强或有其他优异 trick，整体成绩提升 5%-20%

3.5.思考题

五． 思考题

根据这次作业提出几个开放性问题，供学生思考；大家可以把这部分体现在实验报告或代码中，每道题酌情加 0-10 分

1. NER 部分存在同一个词被多个 NER 嵌套的情况，但是序列标注的 NER 模型对于一个词往往只能有一个 NER 标注，如何解决该问题？
   e.g:[[肺动脉]bod 狭窄]sym
2. 分词任务与 pos/ner 任务其实是紧密相关的，如果在分词阶段出现错误，该部分误差就会传递下去，如何解决该问题？
3. 训练数据量十分有限，在不使用人工标注的情况下，如何扩充数据量，并进一步优化模型效果？

六． 时间节点

6.1 号之前必须完成所有任务，并提交测试集分数和实验报告。

推荐时间安排：5.18 之前完成数据预处理模块设计，并完成模型调研；5.25 之前完成模型设计，并进行多次实验调优；5.30 之前完成实验报告撰写和模型细节修改。

七． 其他

1. 鼓励同一组同学之间适当分工，无论任何专业，都应该参与代码设计过程，并对整体模型框架足够熟悉，这样才能起到锻炼目的，甚至也可以成为简历的光辉一笔（尤其是未来希望从事 NLP 方向的同学）。
2. 严禁使用开源工具，代码抄袭和实验报告抄袭，扣分项有足够严厉的惩罚，本课程不保证所有同学都能通过。
3. 无须过分重视模型效果，本课程旨在启发学生的动手能力，只要认真完成作业，在代码和实验报告中体现态度，即使最终结果不佳，也完全不用担心成绩问题。
4. 鼓励创新，对于能够使用创新模型或框架的学生，会有大幅成绩奖励。

5. 最终成绩优秀或者模型新颖的同学，欢迎在最后一节课上进行汇报分享，也会有对应的成绩奖励。
6. 有任何技术问题欢迎在群里 at 助教，当然更鼓励同学们使用网络搜索解决。