# 知识表示学习课程　第一次作业要求：

**通过指定方法或模型在指定数据集上完成一个命名实体识别任务，提交代码和报告。**

**指定方法：**1、隐马尔可夫模型；2、长短期记忆网络（LSTM）

**指定数据集：**Conll03 数据集（文件为 json 格式），非 BIO 格式，需要自行转格式，会在压缩包中给出。

**提交时间限制：**从 9 月 27 号算起，给予三周时间完成，即截至 10 月 18 号（中午 12 点整）。

**提交方式：**将程序源码、报告、以及执行结果文件（不包括模型）打包发送到邮箱：cwt_0139@ruc.edu.cn，提交时邮件主题为“知识表示学习第一次作业”关键字，压缩包文件以姓名+学号的格式命名。

# 评分规则：

## 1、代码（12 分）

### 1.1 可执行（6 分）

要求说明：以 Python（3.5 版本及以上）作为编程语言，完成代码设计和运行调试，并保证程序的可执行和可重复。满分 6 分，两种方法各 3 分。

备注（程序设计思路）：

1、 数据预处理阶段：
   包括划分训练、测试、验证集；分词；数据格式转换（BIO 格式和实体位置+标签格式）、构建字典（词汇字典、标签字典）
2、 模型构建与训练：
   选择合适的模型，并将数据分批次（如果需要的话）输入模型，设定损失函数，进行训练。最后保存模型。（可采用神经网络 Pytorch 框架中的 LSTM 模型。HMM 的训练部分可以采用监督式的极大似然估计， Baum-welch 算法（EM 步迭代）作为可选项）
3、 结果测试
   从文件中读取模型，设计验证函数，得出结果。

### 1.2 程序结果验证（4 分）

要求说明：对一段文本，识别出其中的实体类别和实体边界，对一个实体来说，两者都

正确才算识别正确。计算出查准率（Precision）、查全率（Recall）、以及 F1 分数作为评价指标。两种方法的实现分别计算出 P、R、F1 的评价指标，计算给出结果，并在报告中列出。

备注：

1、评价指标计算（这部分的评分算在 2.4 结果说明中）。对 F1 分数有 weighted-F1（加权 F1）和 Balanced F1（平衡 F1），本项目中取 Balanced F1。即：

$$P = \frac{a}{a+c}$$

$$R = \frac{a}{a+b}$$

$$F1 = \frac{2 * P * R}{P+R}$$

| | Correct | Not Correct |
|---|---|---|
| Selected | a | b |
| Not Selected | c | d |

2、结果文件输出（4 分）。

会给出两个示例文件，依照给定文件格式输出，文件命名参考“模型名称+文件格式“的形式，两个模型，两种格式的输出文件，共 4 个文件，各一分。

（1） BIO 格式

（2） 实体类型+实体位置格式

### 1.3 关键注释和代码整洁（2 分）

要求说明：一个良好的程序需要在关键位置有注释，且变量定义符合规范，满分 2 分。

备注（程序设计要求）：

1、 代码对齐（python 自动要求对齐）

2、 避免重复代码（使用函数调用）

3、 条理清晰（变量定义规范，适当使用注释）

## 2、报告（8 分）

要求说明：一个完整的报告应包括：介绍（Introduction）、相关工作（Related Work）、实验设计（Experiments）、结果和分析（Results and Analysis）4 个部分。满分 8 分，每部分 2 分。

### 2.1 项目简介（2 分）

介绍主要说明本项目的目的，背景，和项目结果概述（300~800 字）；

### 2.2 相关工作（2 分）

相关工作主要以总结性语言阐述解决 NER 问题的相关方法（不少于 10 种，可以查阅相关论文，包括基于规则（Rules-based）、基于概率(Classifier-based)、基于神经网络(Neural-based)等等）；（300~800 字）；

## 2.3 实验说明（2 分）

实验设计包括设计思路和参数设置；关键参数可以以表格的形式列出（如必要的话）。同时，还要对数据集进行分析，如分析训练、测试样本数目；标签分布情况等。

## 2.4 结果分析（2 分）

结果和分析需要列表格说明结果，并给出自己的分析（如有的话），或可以选若干样本进行说明。需要辅以表格说明，示例如下。

表 1 实验数据结果

| Model | precision | recall | F1 score |
|---|---|---|---|
| model_001 | - | - | - |
| model_002 | - | - | - |

注：如果少一种方法实现，将总计扣 1.1（-3）、1.2（-2）、2（-1）共 6 分。