知识表示学习课程　第二次作业要求：

1、 任务一：通过任意 neural-based 的方法在给定数据集上完成关系抽取
2、 任务二：通过远程监督（bag-level）方法完成对给定无标注语料的关系抽取，提交代码和报告。

提交时间限制：从 10 月 18 号算起，给予三周时间完成，即截至 11 月 8 号（中午 12 点整）。

提交方式：将 1.程序源码、2.报告、3.一个文件（任务二中的识别结，results.txt）（不包括模型、数据、执行过程中的文件!!!）打包发送到邮箱：cwt_0139@ruc.edu.cn，提交时邮件主题为"知识表示学习第二次作业"关键字，压缩包文件以姓名+学号的格式命名。

评分规则：

1、任务一：有监督关系抽取（6 分）

任务描述：
给定一个有监督关系抽取数据集 semeval，已经按训练集、测试集、验证集划分完成。**要求：**1.使用训练集（semeval_train.txt）和验证集（semeval_val.txt）的数据完成训练，2.并在测试集（semeval_test.txt）上进行测试，其中关系类别标签见文件（semeval_rel2id.json）

1.1 数据处理（2 分）
涉及分词、字典构建等
1.2 模型设计及训练（2 分）
涉及模型构建、实体表示嵌入或其它实体位置标明的方法、模型训练等
1.3 测试和验证（2 分）
可采用的指标为正确率和 F1 分数
（微平均和宏平均均可，不做要求，报告中说明用了哪一种即可）

2、任务二：基于远程监督的关系抽取（8 分）

任务描述：
给定一批无任何标注的文本语料（unlabelled_data.txt）（语料格式为若干句子的集合）和一个已知知识库（kb.txt），**要求：**1. 通过远程监督方法训练一个 bag-level 的关系抽取模型。2. 训练完毕后，用于识别未知头实体（h）和尾实体（h）对之间的关系，其中未知实体对（h,t）需要自行从文件（unseen_kb.txt）中提取。3. 将识别的关系存入文件（results.txt），并和文件（unseen_kb.txt）比较，判断识别的效果。
其中，为方便同学处理，已经将涉及到的实体（entities）和关系（relation）分别放入文件（entities.txt）和（relation.txt）中。

2.1 实体匹配，数据生成（2 分）

通过直接匹配实体词的方式，构建远程监督数据

2.2 数据处理（2 分）

对生成的数据进行初步处理，分词，构建词典，划分 bag 等

2.3 模型设计及训练（2 分）

数据转为可以输入模型的 tensor，模型构建、模型训练等

2.4 推理和测试（2 分）

通过模型对未知实体对（h,t）进行推理，得出最终结果。

将识别结果和“unseen_kb.txt”中的三元组进行比对，计算 F1 分数。

3、报告与提交规范（6 分）

任务描述：完成实验设计思路的简要说明，并以正确的格式提交到作业邮箱。其中设计思路两个任务分开写，各 2 分，整个报告尽量不超过两页。

3.1 设计思路说明（任务一）（2 分）

设计思路、关键部分说明、结果

3.2 设计思路说明（任务二）（2 分）

设计思路、关键部分说明、结果

3.3 代码规范及提交规范（2 分）

代码部分：整洁，合理注释等。

提交规范：1、压缩包命名包含：姓名学号；2、邮件主题正确；3、不要附带额外数据或模型文件。（如果同时违反两点及以上，扣 1 分）