作业一：中文分词

1.1 目标

- 实现结构化感知器进行中文分词来完成中文自动分词任务，即把连续的中文文本切分成词序列。评估指标是 Precision，Recall，F-score。会提供自动评测的 perl 脚本文件。可以使用提供的脚本文件，也可自己实现相关的评估指标，以提供的脚本分数为准。
- 要求个人独立完成作业一，不能使用已有的机器学习库函数的实现，包括但不限于 sklearn，Pytorch，Tensorflow 等，但是可以使用线性代数计算或者科学计算的库函数进行科学计算，例如 numpy 等。不能抄袭现有的开源实现和别人的代码及实验报告。希望大家自觉遵守学术诚信，我们会对大家提交的代码进行内部查重，以及与网上主流代码、历年学生代码比较进行外部查重，来发现各种抄袭行为。

1.2 数据

- 训练集和验证集为 train.txt 和 dev.txt，均为 utf-8 编码，每行包含一个分词以后的句子，词与词之间用空格分开。
- 测试集为 test.txt，utf-8 编码，每行包含一个未分词的句子。
- 提交的答案格式应当参考 answer.txt(不代表真实测试数据答案)，采用 utf-8 编码。score 为评测使用的 perl 脚本，真实答案不公开，评测分数为脚本自动评测，所以请务必按照示例答案文件的格式提交答案。

1.3 提交

- 截止日期：11 月 27 日 23:59 分
- 提交内容：需要提交的作业内容以压缩包形式提交，作业文件名命名方式为：学号-姓名（如 1200011111-张三）。作业文件必须包含 code 文件夹(内含所有源码)，result.txt(测试集输出)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。
- 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->编程作业一）。

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。会根据大家提交时间和先后顺序酌情给分。
- 作业可多次提交，成绩以最后一次提交的效果和时间为准。