# 自然语言处理作业 1

## 一、作业内容

本次作业内容为利用相关工具执行自然语言处理特定任务并进行评价。个人独立完成作业。

- 主要内容：选择中文分词、词性标注或命名实体识别中的两个任务，利用相关工具在给定的数据集上进行测试并汇报结果：
  - 任务选择两个。
  - 相关工具可以在 spaCy, Stanford CoreNLP, NLTK 中三选一。
  - 提供的数据均为 UTF-8 编码。对于中文分词，数据格式为纯文本(.txt)，每行包括一个句子，词与词之间用空格分开。对于词性标注和命名实体识别，数据格式为制表符分隔的值(.tsv)，每行包括一个词和其标签，空行表示句子结束。
  - 中文分词、命名实体识别采用 Precision, Recall, F-Score 作为评价指标；词性标注采用 Accuracy 进行评价指标。如数据中的标签集与工具的标签集不同，请先确定标签间的对应关系然后进行评价。

- 扩展内容（可选任务，非必需）：分析测试结果中的错误情况和可能的原因，可进行考察的方面包括文本风格、文本长度、歧义现象和同义现象等。鼓励对问题出现的规律作一定的总结。

## 二、作业形式

作业形式为实验报告。建议统一用中文书写，对于特殊术语，可考虑在中文后用括号表明对应英文；报告中需要说明所使用的工具及其版本。作业提交时还需要包括实验代码与测试结果（按给定数据的格式）。

## 三、作业提交

- 截止日期：11 月 26 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）。必须包含实验报告、源码及测试结果。
- 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->教学内容 -> 作业 1）。

## 四、作业评分

- 主要对报告内容的详实程度和规范程度进行评分，如包括扩展内容则还包括分析的合理性。其它评分点还包括作业工作量与实验代码风格等。
- 鼓励在截止日期之前留一定余量提交。
- 作业可多次提交，成绩以最后一次提交的为准。