

1. 由于数据集过大，故未放到作业压缩包中

分词与词性标注使用的数据集为 corpusZh，地址：<https://github.com/liwenzhu/corpusZh>；  
命名实体识别使用的数据集为人民日报语料库。

2. 训练集测试集的划分

对分词与词性标注任务，由于数据集过大，故取前 10000 个句子为训练集，10000~12000 的句子做测试集。

3. 使用说明

分词与词性标注直接运行即可，控制台输出为分词和词性标注的准确率、查全率、f1、效率。

命名实体识别直接运行即可，控制台输出为实体识别的准确率、查全率、f1。