自然语言处理与文本挖掘 作业一
中文分词

2020.3.5

本次实验希望大家理解和应用常见的分词方法

数据集说明：

我们使用 SIGHAN Second International Chinese Word Segmentation Bakeoff 评测数据来进行模型的训练和测试，官方提供的数据为 icwb2-data.zip （评测数据主页为 http://sighan.cs.uchicago.edu/bakeoff2005/）。

我们只使用其中的”pku”部分数据，即使用压缩包中的 training/pku_training.utf8 作为训练集，testing/pku_test.utf8 作为测试集（gold/pku_training_words.utf8 为训练集的词典，可供参考）。

你的模型应将测试集的分词结果输出到一个文本文件中，格式与训练集相同。然后参考 “测试方法” 文档中的说明来测试你的分词结果。

作业提交说明：

需要提交的内容有：报告文档（包括结果数据，分析等，见题目具体要求），程序源代码及其运行方法（可以复现报告中的结果数据）。

编程所使用的语言不限。

1. 基于字典匹配的分词
   实现一种基于字典匹配的分词方法，在数据集上评价其性能。观察输出的分词结果，分析并举例说明字典匹配方法的效果如何，有哪些问题？如果你有兴趣，可以实现不同的匹配方式并比较其特点。

2. 基于序列标注和 CRF 的分词
   将分词视为序列标注问题，实现一个基于 CRF 的分词模型，并在数据集上进行训练和测试。请给出训练集和测试集上的性能指标。你的模型使用了哪些特征？调整使用的特征的种类，比较不同特征对模型性能的影响。
   注：CRF 部分可以使用开源工具包，如 CRF++（https://taku910.github.io/crfpp/）。

3. （选做）基于神经网络的分词
   实现一种基于神经网络的分词模型，例如使用前馈神经网络、RNN(LSTM)等，可检索并参考其他文献中的模型，并在报告中给出准确的引用信息。给出你的模型设计、训练和测试的流程，以及测试集上的性能指标。观察分词结果，比较神经网络方法与前两种方法的特点。你还能想到哪些提升神经网络模型性能的方法？
   注：神经网络部分可使用开源工具实现，如 Tensorflow，Torch，pyTorch 等。