# 2月25日第一次作业

**AI Studio用户名：赵杭天**

## 作业1-1：

（1）下载飞桨本地并安装成功,将截图发给班主任：



（2）学习使用PaddleNLP下面的LAC模型或Jieba分词
LAC模型，安装并测试成功：



（3）对人民日报语料完成切词，并通过统计每个词出现的概率，计算信息熵
使用Jieba分词：





删除因.MD格式而出现的符号，并使用/作为分隔符以实现并行处理，最终结果：

**文本文件数：447，不同词的数量：21758，总词数：172731，信息熵：7.467906**

## ****作业1-2****

（1）思考一下，假设输入一个词表里面含有N个词，输入一个长度为M的句子，那么最大前向匹配的计算复杂度是多少？

 最坏的情况：词表所有的词都没有出现在句子里，句子被判定为当前未处理句子长度（记为S1\_len）个单字，图解如下：



所以，总的比较次数=$N\frac{\left(M-1+2\right)×(M-1)}{2}=\frac{N(M^{2}-1)}{2}$，复杂度为$O(N×M^{2})$。

（2）给定一个句子，如何计算里面有多少种分词候选，你能给出代码实现吗？



（3）除了最大前向匹配和N-gram算法，你还知道其他分词算法吗，请给出一段小描述。

【基于深度学习的端到端的分词方法】

最近，基于深度神经网络的序列标注算法在词性标注、命名实体识别问题上取得了优秀的进展。词性标注、命名实体识别都属于序列标注问题，这些端到端的方法可以迁移到分词问题上，免去CRF的特征模板配置问题。但与所有深度学习的方法一样，它需要较大的训练语料才能体现优势。



BiLSTM-CRF的网络结构如上图所示，输入层是一个embedding层，经过双向LSTM网络编码，输出层是一个CRF层。下图是BiLSTM-CRF各层的物理含义，可以看见经过双向LSTM网络输出的实际上是当前位置对于各词性的得分，CRF层的意义是对词性得分加上前一位置的词性概率转移的约束，其好处是引入一些语法规则的先验信息。

从数学公式的角度上看：



其中，A是词性的转移矩阵，P是BiLSTM网络的判别得分。



因此，训练过程就是最大化正确词性序列的条件概率P(y|X)，似的工作还有LSTM-CNNs-CRF：

