使用经过人工分词后的北京大学《人民日报》标注语料库

（含train/valid/test三个文件夹）

完成基于**马尔科夫模型**及**smoothing策略的句子概率计算**。

对上述语料库做适当处理，如：在每一句开头增加一个句子起始符<BOS>, 在其结尾处增加一个句子结束符<EOS>。 （可以认为这两个特殊字符是“虚拟”地加上去的）

基于词的**bigram**（一阶马尔科夫模型）计算如下句子的概率（参考lesson4part2课件）：

* 扶贫 开发 工作 取得 很 大 成绩 （句子1）
* 扶贫 开发 工作 得到 很 大 成绩 （句子2）
* 你再任给2个其他句子（要求：其中至少一个词在train中没有出现过）

要求用如下几种smoothing方法：

1. 根据在train + valid + test整个语料库中的统计结果，使用adding-的smoothing策略计算上述句子的概率。要求分别测试=0.1及1.0共2组参数。
2. 以train为基础，将valid以及 valid + test分别当做Held out语料库，使用Held out estimation, 计算上述句子的概率。
3. 分别在train + valid以及 train + valid + test上，使用Good-Turing estimation，计算上述句子的概率。

对上述各种配置下得到的结果进行必要的观察、比较和分析。