## 简单语言概率模型中文分词

1. 开发环境：
   1. Java 9.0.1
   2. Intellij IDEA 2017.3
   3. Maven 3.3.9 （Intellij 已经集成的）

为保证正确运行，请尽量使用相同的开发环境，Java 8 应该也可以，但是7 及以前不行，因为用到了 Java 8 及以上的新特性。

注意：涉及到中文文本的处理，请使用 UTF-8 编码！

1. 基本思想：

将分词转化为一个标注问题，对每一个“字”标记一个标签，用 b 表示词首，c 表示词的第二个字，d 表示词的第三个字，e 表示词的第四到第n个字，所以是一个4-Gram 语言模型。在训练语料上学习一个模型，通过 Viterbi 算法，找出概率最大的标注序列，从而得到分词结果。

具体参考：<http://spaces.ac.cn/archives/3956>

1. 项目结构：

语料来自 <http://sighan.cs.uchicago.edu/bakeoff2005/>

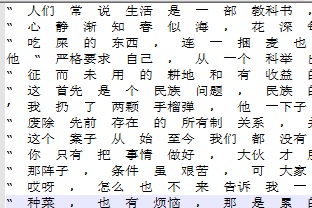
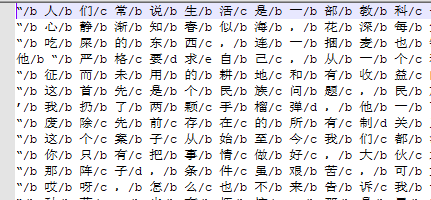
1. Corpus 文件夹，是训练、测试语料的目录
2. Corpusraw 文件夹，是原始语料
3. Dictionary 文件夹，是可以用户自定义的字典
4. Lib 文件夹，用到的第三方包，使用Maven 可以自动下载
5. Model 文件夹，语言模型的参数存储
6. Src 文件夹，项目代码
7. Target 文件夹：Maven 生成的项目class 文件
8. Pox.xml，Maven 的配置文件
9. Results.txt，得到的分词结果
10. Segment.iml，Intellij的配置文件
11. 代码文件说明

本项目是 Maven 项目，只用到了一个第三方资源包（Intellij 的 GUI 设计包），直接导入 Maven 项目，会自动下载配置，需要等待一段时间。

文件的代码（Segment\src\main\java\com\segment\目录下）主要包括：

* 1. CorpusPreProcess.java

数据集预处理，将训练语料进行分词标注

 -> 

* 1. PLM.java

算法的核心实现，结合相关的注释看即可

* 1. SegUI.form, SegUI.java

GUI 设计，通过 Intellij IDEA 内置的 GUI Designer 辅助设计的

* 1. WordSegment.java

程序的入口

1. 运行结果说明

根据官方发布的评测标准，准确率达到80%，而本项目中使用的测评标准实现比较粗暴，所以显示只有50%+。

这也说明了概率语言模型不是十分适合分词这项任务的。如果有一个比较好的用户词典，使用最大双向匹配的算法通常是快速且准确的做法。

附官方说明地址：

http://www.52nlp.cn/%E4%B8%AD%E6%96%87%E5%88%86%E8%AF%8D%E5%85%A5%E9%97%A8%E4%B9%8B%E8%B5%84%E6%BA%90#more-2885