离线切词环境使用文档

概述

组内切词工具，工具主要由两部分组成

1. 词典构建，对输入的词表文件生成切词资源
2. 切词环境，将切词资源拷贝到切词环境中，完成真正的切词操作

词典构建

生成切词词典
工具目录 build_dict
输入：词表文件
输出：切词词典资源
运行脚本 ./shell/get_worddict.sh 输入文件
脚本会生成一个文件夹，其中是切词工具需要的资源，
举例，如果输入文件是 vocab，则生成资源文件夹名为 worddict.vocab

切词环境

完成切词过程
工作目录：offline_word_splitter
配置： conf/scw.conf
资源： 从 build_dict 中生成的切词资源，将文件夹拷贝到当前环境。
并将配置中 Language_dictpath 配置设置为对应的切词资源。
运行方式：
因为./bin/wordseg_for_yuyin 只能针 utf8 格式的文本进行切词，因此需要先将文本转换编码格式成 utf8，具体运行方式
iconv -f gbk -t utf8 输入文件| ./bin/ offline_word_splitter conf/scw.conf | iconv -f utf8 -t gbk > 输出文件

标准环境

root@10.150.130.43:/data/lele/offline_word_splitter
其中包括
build_dict 词典构建
offline_word_splitter 切词

词表构建环境中词表只有两个词
中国
人民

切词环境中，利用生成的资源对测试文本进行切词，测试文本中只有一行
中国人民

切词之后的结果
中国 人民