# 中文分词程序说明

#### 小组成员：

杨欣松，熊炜伟，张冠群

#### 任务定义：

Chinese word segmentation: 10 points

This task provides PKU data as training set and test set (e.g., you can use 80% data for model training and other 20% for testing ), and you are free to use data learned or model trained from any resources.

Evaluation Metrics:

Precision = (Number of words correctly segmented)/(Number of words segmented) \* 100%

Recall = (Number of words correctly segmented)/(Number of words in the reference) \* 100%

F measure = 2\*P\*R / (P+R)

#### 输入输出：

输入：

1998-01-105-带音.txt

输出：

preLearn.txt //学习前预处理

dictionary.txt //词典

answer.txt //分词原文

preSeg.txt //分词前预处理

result.txt //分词结果

analyse.txt //结果分析

#### 方法描述：

**最大匹配（贪婪算法）：**

先对文本进行预处理，取1998-01-105-带音.txt前80%生成字典，后面20%作为测试文本。

扫描文本时先读入默认最大项长度（本程序设为5）的字符串，如果不匹配则向前回退一位，直到分词成功，把指针移动到分词位置，如此循环直到文本结束。

分析结果时统计每对”/”在文本中的相对位置，比较结果和正确答案相同的个数，即正确分词的个数。

#### 结果分析：

corret = 206682

segNum = 235275

num = 223107

precision = 0.87847

recall = 0.926381

F = 0.901789

长度超过默认最大项（本程序为5）的词，词典中没有的词，和最大匹配不正确的词不能正确分出。

#### 源码运行环境：

VS2013 C++

segment.cpp

Segment.exe