中文分词说明文档

姓名：刘泽君

班级：2015211306

学号：2015211304

内容目录

[一、任务定义 3](#__RefHeading___Toc316_728037814)

[二、输入输出 3](#__RefHeading___Toc318_728037814)

[三、方法描述 3](#__RefHeading___Toc320_728037814)

[四、结果分析（性能评价） 3](#__RefHeading___Toc322_728037814)

[五、源码运行环境 3](#__RefHeading___Toc324_728037814)

# 一、任务定义

中文分词：

此任务提供PKU数据集作为训练集和测试集（你可以用80%的数据来训练模型，剩下20%来测试模型），你可以使用任何从其他资源处得到的训练过的模型或是学习过的数据。

评估标准：

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)

Chinese word segmentation: 20 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)

# 二、输入输出

* 输入：

训练数据：文件夹中的“pku\_training.utf8”

测试数据：文件夹中的“pku\_test.utf8”

来自：icwb2-data：pku 数据集网站：<http://sighan.cs.uchicago.edu/bakeoff2005/>

* 输出：

分词结果：“seg\_result.txt” 编码格式：utf-8，windows下的记事本打开会乱码

# 三、方法描述

* 数据集：

此分词系统运用 http://sighan.cs.uchicago.edu/bakeoff2005/ 提供的 icwb2-data 中PKU的语料库作为训练集，测试集使用其中PKU的test测试集。

* 源码word\_segment.py说明：

Dictionary类将“ pku\_training.utf8”中的词语提取出来，建立词典，以便在此词典的基础上进行分词。

MM类使用了最大前向匹配分词的方法。此分词器将词典中最大单词长度设置成了11个汉字（utf8编码33个字节）。其中，用word来寻找匹配的词：如果字典中没有word，则将word右侧的长度减一；如果匹配到了word，则将分词结果输出到seg\_result.txt 文件，并将单词加入到test词典中。

# 四、结果分析（性能评价）

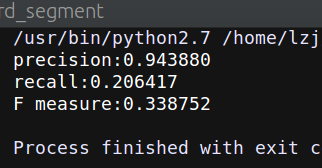
* 评价标准：

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)

* 分词器运行结果：



# 五、源码运行环境

编程语言：python 2.7

运行环境：ubuntu 16.04

编码格式：utf-8