# 1. 任务定义

使用最大匹配算法的中文分词方法处理测试集，并进行性能评估

# 2. 输入输出

**输入:**

- 字典:dict_810K.txt
- 测试集:pku_test.utf8
- 答案集:pku_test_gold.utf8

**输出:**

- 分词后的结果:out.utf8

# 3. 方法描述

## 正向最大匹配算法

根据设置的max_len从正向进行匹配，由max_len长度从字典中找到匹配的词则返回结果，否则删去最后一个字，词长减去1继续匹配，长度为1时直接返回。

## 逆向最大匹配算法

根据设置的max_len从逆向进行匹配，由max_len长度从字典中找到匹配的词则返回结果，否则删去第一个字，词长减去1继续匹配，长度为1时直接返回。

# 4. 结果分析（性能评价）

对SIGHAN Bakeoff 2005比赛中自带的pku_test数据进行逆向最大匹配算法分词，运行时间5327秒

用自带的score脚本，经过测试，结果如下

- SUMMARY:
- TOTAL INSERTIONS: 4355
- TOTAL DELETIONS: 13378
- TOTAL SUBSTITUTIONS: 21849
- TOTAL NCHANGE: 39582
- TOTAL TRUE WORD COUNT: 83697
- TOTAL TEST WORD COUNT: 74674
- TOTAL TRUE WORDS RECALL: 0.579
- TOTAL TEST WORDS PRECISION: 0.649
- F MEASURE: 0.612
- OOV Rate: 0.485
- OOV Recall Rate: 0.521
- IV Recall Rate: 0.633

# 5. 源码运行环境

unix环境,python3,安装库sklearn,re

**源码结构:**

- test.py 测试代码
- segmentation.py 分词模块

# 性能评价脚本使用

The script 'score' is used to generate compare two segmentations. The script takes three arguments:

1. The training set word list
2. The gold standard segmentation
3. The segmented test file

You must not mix character encodings when invoking the scoring script. For example:

% perl scripts/score gold/cityu_training_words.utf8 gold/cityu_test_gold.utf8 test_segmentation.utf8 > score.ut8