

# 中文分词分析报告

学号-2014211280-谢非

2016 年 11 月 18 日

## 目录

<b>1 实验目的</b>	<b>2</b>
<b>2 实验要求</b>	<b>2</b>
<b>3 实验环境</b>	<b>2</b>
<b>4 实验原理介绍</b>	<b>2</b>
4.1 词典的动态导入 . . . . .	2
4.2 前向匹配和后向匹配 . . . . .	3
4.2.1 匹配模式选择 . . . . .	3
4.2.2 实现原理 . . . . .	3
<b>5 运行示例</b>	<b>3</b>
<b>6 性能评价</b>	<b>3</b>
<b>7 实验总结</b>	<b>4</b>

## 1 实验目的

1. 熟悉中文分词的具体步骤和细节
2. 应用最大前向和后向匹配实现中文分词
3. 了解对于分词程序的评价

## 2 实验要求

1. 实现一个可以前向或后向匹配的分词器
2. 分词器具有灵活的配置选项和输出

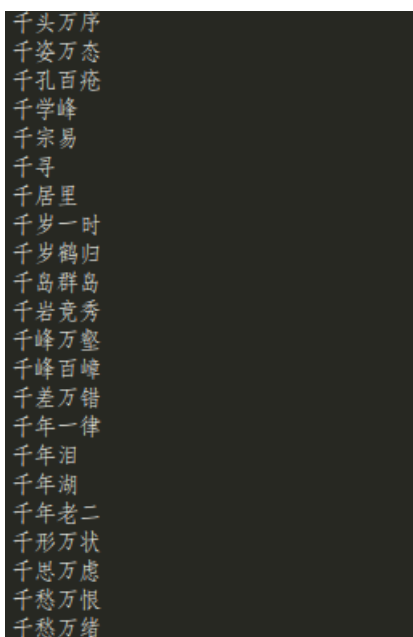
## 3 实验环境

实验实现的 python 分词器在 python3.5 环境可运行, 没有外部依赖包实验的文件输出可选择是在 window 下还是类 unix 下.

## 4 实验原理介绍

### 4.1 词典的动态导入

实验实现的程序可以动态导入词典, 自带的词典 dict.txt 是从网上找到的多个词典叠加处理后得到的. 自定义的词典格式简单, 只需要把要添加的词语每行一个, 存入词典文件中. 命令行执行 fenci.py 时加参数 -d \*\*.txt(词典文件) 可使用自己的词典. 如果也想用自带的词典, 可以将参数改成 -c \*\*.txt(词典文件), 此时会应用自带词典和用户添加的词典.



千头万序  
千姿万态  
千孔百疮  
千学峰  
千宗易  
千寻  
千居里  
千岁一时  
千岁鹤归  
千岛群岛  
千岩竞秀  
千峰万壑  
千峰百嶂  
千差万错  
千年一律  
千年泪  
千年湖  
千年老二  
千形万状  
千恩万虑  
千愁万恨  
千愁万绪

图 1: 词典文件如图

## 4.2 前向匹配和后向匹配

### 4.2.1 匹配模式选择

实验程序可以添加参数配置 `-m 1` 选则匹配模式是前向匹配还是后向匹配。

```
# thief @ thiefuniverse in ~/PycharmProjects/Fenci [2:25:30]
$ ./fenci.py -l "这件事和尚不清晰的现状是有原因的。"
这件/事/和尚/不/清晰/的/现状/是/有/原因/的/./ /

# thief @ thiefuniverse in ~/PycharmProjects/Fenci [2:53:52]
$ ./fenci.py -l "这件事和尚不清晰的现状是有原因的。" -m 2
这件/事/和尚/不/清晰/的/现状/是/有/原因/的/./ /
```

### 4.2.2 实现原理

加载词典文件之后，分词器会具备一个 dict 对象，键表示词语长度，值表示一个词语 list(包含该长度的所有词语)。对一个句子分词时，如果是前向匹配，则从词典中长度最大的往下匹配，如果找到，则把匹配词弹出，存入结果中；如果到长度为 1 时还没有找到，则弹出句子中的一个字存入结果中，然后继续匹配。

## 5 运行示例

示例如下：

```
# thief @ thiefuniverse in ~/PycharmProjects/Fenci [4:21:30]
$ ./fenci.py -l "我的电脑很好用"
我/的/电脑/很/好/用/

# thief @ thiefuniverse in ~/PycharmProjects/Fenci [4:21:43]
$ ./fenci.py -l "我们中国现在的发展也是蛮快的。"
我们/中国/现在/的/发展/也/是/蛮/快/的/./ /

# thief @ thiefuniverse in ~/PycharmProjects/Fenci [4:22:23]
$ ./fenci.py -l "比起斐济，我更想去巴黎或者哥伦比亚度假。"
比起/斐济/，/我/更/想去/巴黎/或者/哥伦比亚/度假/./ /

# thief @ thiefuniverse in ~/PycharmProjects/Fenci [4:26:16]
$ ./fenci.py -l "历数这些寻求合作的企业，既有江淮、奇瑞这样中国汽车工业异军突起的新军，也有东风公司徐平这样的车坛宿将，更有国内外金融投资巨头。"
历数/这些/寻求/合作/的/企业/，/既有/江淮/、/奇瑞/这样/中国/汽车/工业/异军/突起/的/新军/，/也/有/东风/公司/徐平/这样/的/车坛/宿将/，/更有/国内/外/金融/投资/巨头/。

# thief @ thiefuniverse in ~/PycharmProjects/Fenci [4:26:21]
$ ./fenci.py -l "比如我们的发动机项目技术含量很高，国内外都看好，一旦开展合作，将会出现产能不够的问题，加上我们推进国际化战略，肯定需要更多的战略投资者。"
比如/我们/的/发动机/项目/技术/含量/很/高/，/-国内/外/都/看好/，/-一旦/开展/合作/，/-将/会/出现/产能/不够/的/问题/，/-加上/我们/推进/国际化/战略/，/-肯定/需要/更多/的/战略/投资者/。"
```

## 6 性能评价

测试时从网上找了一个带有答案的测试文件（包含 3654 个句子，将其分成了三份进行测试，一份 1144 句，一份 1000 句，一份 1510 句），实际分词后对比得到分词正确率为 84.24%(P 值)，分词正确占词引用数比值为 95.23%(R 值)，F 为 89.39%。

<sup>1</sup> 1 为前向匹配，2 为后向匹配。

## 7 实验总结

本次实验过程中我对分词的具体实现有了更加深刻的认识，同时也对其他不同的分词方法产生了浓厚的兴趣，期待之后可以用更加复杂而精确性更高的分词方法。