**日志解析项目介绍**

## 项目内容

用Python完成对《chinese\_english\_logs.csv》和《english\_logs.csv》中的content列的日志内容的解析，得到EventId、EventTemplate（日志模板）、Occurrences（模板出现的次数）和ParameterList（日志中的参数）的值，具体如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 原始数据表 | 解析方法 | 解析结果表 | 日志行数 | 得到EventId的个数 |
| chinese\_english\_logs.csv | 统计法 or Drain3 | chinese\_english\_logs\_parse\_by\_statistic.csv | 6163行 | 29个 |
| chinese\_english\_logs.csv | Drain3 | chinese\_english\_logs\_parse\_by\_drain3.csv | 6163行 | 40个 |
| english\_logs.csv | Drain3 | english\_logs\_parse\_by\_drain3.csv | 36220行 | 69个 |

## 方案介绍

### 1.1、基于统计的方法

1.1.1、方法介绍

《chinese\_english\_logs.csv》中content列的内容，含有中文的日志为6155条，占比为6155/6163= 99.8%。观察数据的特点，可以用统计方法得到EventId、EventTemplate、Occurrences和ParameterList，具体流程如下：

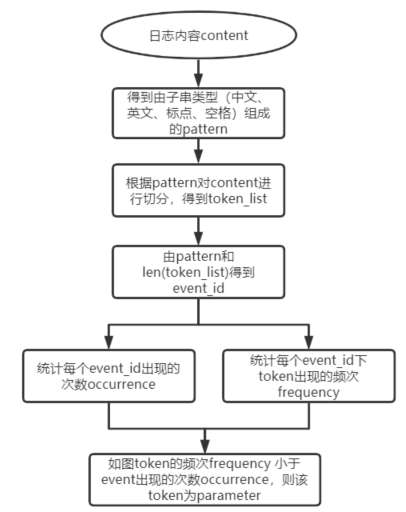


图1 统计方法的流程图

根据pattern对content进行切分的流程图如下：

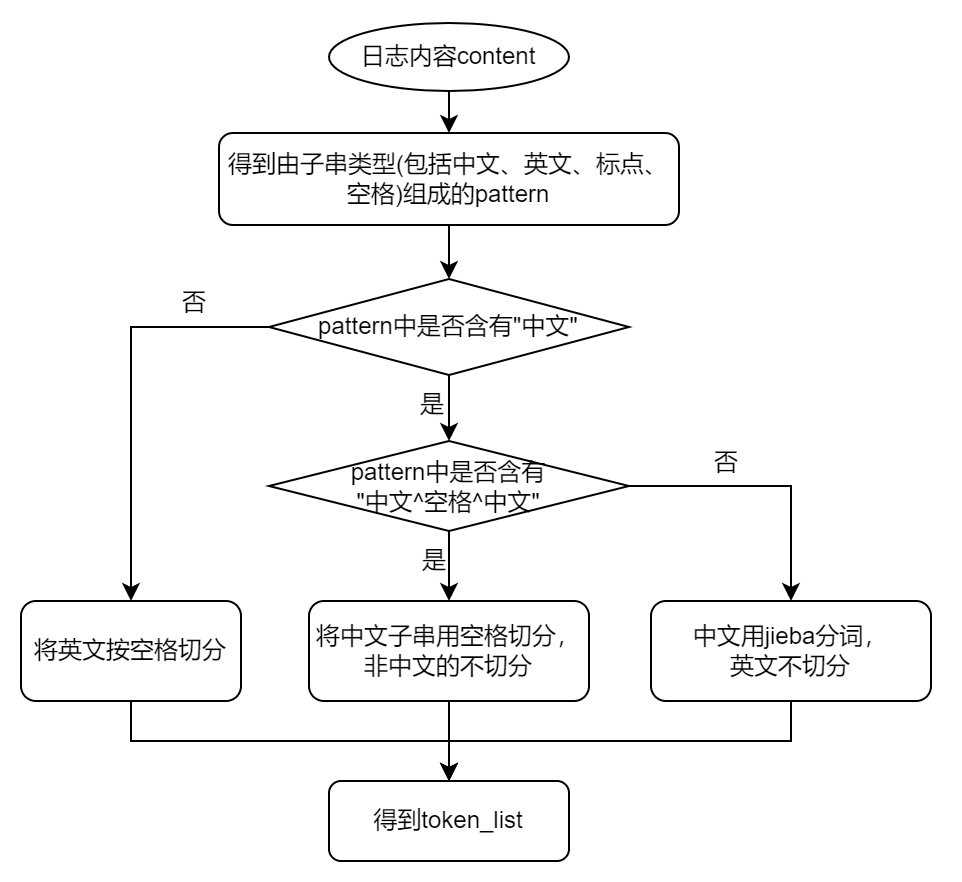


图2 根据pattern对content进行切分的流程图

#### 1.1.2、举例说明

以《chinese\_english\_logs\_parse\_by\_statistic.csv》中的第4行为例，输入的content为“Google 更新服务 (gupdate) 服务处于 正在运行 状态。”，则获取到的pattern为*“英^空格^中^空格^英^空格^中^空格^中^空格^中^标点”*，切分得到的token\_list为*['Google', ' ', '更新服务', ' ', '(gupdate)', ' ', '服务处于', ' ', '正在运行', ' ', '状态', '。']*，根据pattern和len(token\_list)，得到event\_id = 0，EventTemplate为*“Google 更新服务 (gupdate) 服务处于 \* 状态。”，*得到的parameter\_list为*['正在运行']*。

#### 1.1.3、方法缺点

如果日志content中不包含中文，且两个content中含有的英文单词个数相同，则它们的pattern和len(token\_list)均相同，但是它们很有可能不是对应同一个event\_id。

例如：“DSNY001I -PB3A SUBSYSTEM STARTING”与“DSNR002I -PB4A RESTART COMPLETED”的pattern均为“英”，且len(token\_list) = 4，但是它们不是同一event\_id。

### 1.2、基于Drain3的方法

由于基于统计的方法，对处理不含中文的content存在问题，故尝试开源的Drain3方法，该方法的源码为<https://github.com/logpai/Drain3.git>。原始的Drain3方法不适合处理含有中文的日志，经过改造后，能处理中文、英文和中英文混杂三种情况。