**视角抽取算法程序说明**

本部分包含了视角抽取的总体过程，所有过程都基于网络爬虫形成的词典：

**一、基于多模匹配，利用ViewExtract.java()抽取出两部分**

一部分是单个视角，此类视角由于精确匹配，字面上一定是准确的，该文件输出

为view\_part1.txt

连续的多个视角，此类视角由于连接在一起，极有可能是同一个视角，但也有可能不是同一个视角，因此由基于最长公共子串匹配的python部分程序进一步划分, 该部分文件分为combined\_views.txt

**二、基于最长公共子串的模式匹配**

利用split\_combined\_views.py ,处理 **1** 当中**b**部分产生的combined\_views.txt，得出view\_part2.txt

将view\_part1.txt和view\_part2.txt组合成最终的generate\_view.txt

将结果进行尾部处理排序输出进入后期的规则处理并形成最终的generate\_view.txt

**三、运行步骤**

1. 用ideaIntelliJ打开，选择maven方式,修改split\_combined\_views.py中的base\_path 为该文件当前所在的绝对路径，编译运行/src/test/java/org/viewextract/ViewExtract.java，得到generate\_view.txt
2. 在linuxterminal中运行：

cd 到 ViewExtract.java 所在目录， 运行/bin/bash view\_review.sh, 进行排序去重等操作