1. 判断是否进入阅读模式。

判断方法：通过sequenceFilter.java中的addNodetypeForNode（）方法识别页面中是否存在title（标题）和content（正文），如果存在，则进入阅读模式。

Title存在识别方法：存在<h1> 标签

<body><h1>a</h1></body> 里内容= <title></title>的内容

大部分情况下是h1标签。有的不是h1标签。有的是span标签等。

优化：对于<body><h1>a</h1></body> 里内容= <title></title>的内容

并且h1不存于form表单里，有些网站<body><h1>a</h1></body> 里内容= <title></title>的内容但这些网站不应进入阅读模式。

1. 取得正文内容：

根据标签：通过weightNode算的该网页所有node中的最大值。其中设置各个标签的权重值。然后遍历根据最大值，取出需要的那一块。即：这一块就是需要的正文内容。

3.阅读模式处理

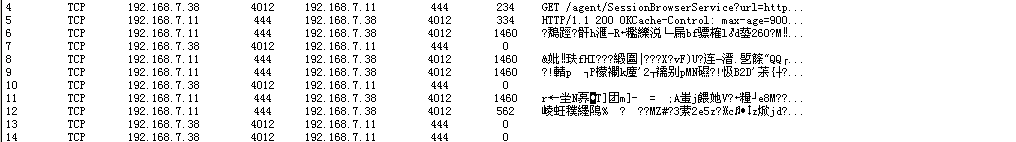
（1） 对权重值最大的那一块进行处理。

抽取掉Content中除<p>,<b>,<small>,<big>,<i>,<tt>,<br>,<center>,

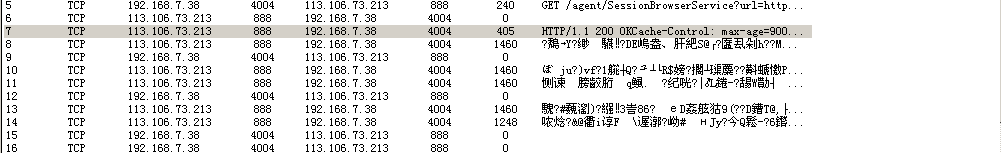
<em>,<strong>,<dfn>,<code>,<samp>,<div>,<span>以外的所以标签。

(2)优化：再次取抽取掉content中内容后的最大节点，重点对于有些网站<div>标签里的内容不属于正文内容应去除，节省流量。

以页面http://novel.hongxiu.com/a/286613/3518203.shtml为例，测试优化前后流量对比，无法保证所有页面，仅供参考；  
优化前：  
   
优化后：  
   
  
通过对比发现优化前服务器返回数据总量为7493字节，优化之后的服务器返回的数据总量为6729字节，对比优化前有所减少，此例子仅供参考。



本地抓包数据



深圳抓包数据

特殊情况处理：

1．某些内容是经过特殊处理的。例如：div里面的内容，并不是需要的，需要进行特殊处理。类似xpath

2.自动加载的时候，有的下一页是图片，有的情况是最后一页了，还有下一页这个链接等这些情况进行特殊处理。