目前为止，能够抓取到包含文本最多的节点，不管是什么主页还是文章页面，主页的标题和简介组成的多个DIV，和文章的主标题和正文DIV

### 提取文章标题和正文

找到文本最多的父节点

将子节点保存到数组中

再检查每一个子节点，如果只有一个子节点中包含文本，说明还要再深入，深入之前要判断能否深入，不能深入的条件就是 p h strong等非容器节点，直到有至少两个节点包含文字，如果大于两个节点，反正只有第一个设置成标题，后面的都归于文本，细致的分类后面再说

拿到的两个节点，主观上如果只有两个是以第一个为标题，如果有很多个还要判断哪一个中

包含h1,h2等标题节点的设置为键

包含大量p或者大量文字的设置成值

最后创建字典保存

### 提取首页标题和简介等信息

找到子节点最多的节点，可能是导航栏或者我们要找的DIV，

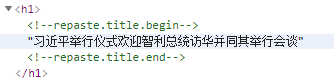
如果里面字数很少那就是导航栏

现在找到了包含标题和简介的父节点，放到列表

检查每一个子节点下面的子节点，和上面的原理一样，就是找到一个即只包含h字段的父节点又是其它字段的兄弟节点

标题模式

百度新闻文章标题使用的h1



百度新闻首页在找到大量文本的DIV后，可以通过查看兄弟节点确定标题



知乎首页标题也是h2，兄弟节点div就是简介，但是没有p节点，文字和节点之间也还有几个节点





CSDN首页同样结构，虽然中间有很多层，但是标题和简介还有一些属性总有几个兄弟节点

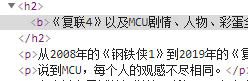


也就是说对于首页来说，只我们先找到包含h1h2这样的标题字段，就说明有可能是标题，他的父节点的兄弟节点有可能是简介，也可能继续往上，我们可以h2的标题节点一直往上，直到突然文本多起来，说明这个节点下包含了标题和简介，确定这个节点下的n个子节点，排除到包含标题的节点，其它节点便是这个标题对应的信息

## 正文模式

百度新闻一篇文章中含有大量p

知乎文章h2标题，还是有很多p，都在一个div里面



百度知道就有点麻烦，有点和找标题有点类似，现在能找到标题和正文的位置关

