切分单词时，出现大小写变化的，也进行切分。http://fun.youth.cn/yl24xs/gxq/tp/201605/W020170107578325771520.jpg

1. 密集文本正文中没有图片，但是密集文本正文中，下方的推荐中有图片，错误提取推荐中的图片。针对这种情况，应该考虑下。

http://399.epaper.rmzxb.com.cn/c/2017-03-24/1439357.shtml

1. Pic url名字中包含关键词的情况，可以考虑。下面名字是ico的被召回了。

<http://acg.gamersky.com/news/201703/885030.shtml>

<http://artist.baozang.com/15666/works19032>（包含banner）

1. 名字长度，比如长度为1的，宽度有值且比较小的，没有利用起来，可以把这个利用起来。（**看一下名字长度的权重**）

<http://admin.huaian.gov.cn/web/center/ztbd/zxzt/jhhh/2012/07/30/2668525.html>

1. display属性为none时，被召回了。<http://api.yfcnn.com/comment-4660322.html>

如果页面评论过多，正文过少，可能导致有问题。

<http://api.yfcnn.com/comment-4660322.html>

1. 通过判断img外面的a标签，url是否是当前页，或者url和当前page url的子路径是否相同来过滤（已添加）
2. 关键词中增加login/logo/face/tool/center/qq（已添加）
3. <http://bjrb.bjd.com.cn/html/2017-03/27/content_120499.htm>的<http://bjrb.bjd.com.cn/images/2017-03/27/01/Page_b.jpg>错误标注

<http://blog.sina.com.cn/s/blog_78f9fd9b0102eaxw.html>下的图片标注有问题

<http://collection.sina.cn/yejie/2017-03-24/detail-ifycstww1041104.d.html?vt=4&HTTPS=1>标注有问题

<http://comic.ishangman.com/read/35762.html> 标注有问题

<http://fj.china.com.cn/p/317531.html> 标注有问题

<http://gupiao.southmoney.com/zhuli/201703/264041.html>（结构相关，内容不相关）

1. 整个block的a标签密度，作为这个block里img的特征。注意考虑a标签数，不能单纯考虑比例。A标签数整个body的标签数的比例。（已完成，目前是单纯考虑比例）
2. 整个block的li标签密度，作为这个block里img的特征（已完成，目前是单纯考虑比例）
3. Img是否在li里面，同时距离li的距离。或者是li的第几层孩子结点。（这个距离有点弱，暂时不考虑）
4. 如果有宽和高的，则判断下宽高和计算下宽高比例（已添加）
5. 对于pic url的path或者名字，命中下关键词，比如ico/logo等信息（已添加）
6. 增加一个特征，pic url是否被a标签包含，同时，a的跳转网址和当前页面的相似度。越低越好。跳转是否同一个站点/跳转是否同一个path/跳转是否同一个page url（正在计算这个特征，其他特征暂时没有计算）（已完成）
7. 对同一个block内，path长度相同的pic url进行平滑处理
8. html title和alt/title的相似性（已添加）
9. 尝试对class/id进行ngram
10. 对img的pic url中特点，比如ico等信息进行挖掘。（已添加）
11. 看下现有PA，怎么进行pic url的路径过滤的。
12. 看下对错误召回的非pic url的错误情况。
13. 看下目前不能正确分类的页面情况。
14. 扩充下数据
15. 对于有width和height的pic url考虑进去。（已添加）

**看一下程序为什么pic url列会出现page url的值。**

<http://a.66game.cn/app/33880.html>这个页面是导致特征抽取非常慢的典型

// 判断page url和linked url的相似性时，将根域名这种形式排除掉，比如：<http://pic.news.sohu.com/group-790050.shtml>

**strict的page url linked url相似性有问题，javascript://**

**解析宽度和高度时，最好将style中的width和height解析出来**