1、<http://zz-nn725.blog.163.com/blog/static/16222221201153102744566/>，这个页面中下面的图片，怎么识别名称，是不是为0？

<http://blog.163.com/bitMaxImageGen.do?url=http://zz-nn725.blog.163.com/blog/static/16222221201153102744566&w=100&h=100&act=qbbkfxtj_20150420_05>

如果图片周围文本都很少，有什么好的解决办法没？

2、可以考虑下alt或者title和html title的相似性。或者周围文本和html title的相似性。同时加一个是否存在alt或者title，统计下，长度也可以。

3、<http://zx.meilele.com/ideabooks/item-746852/>，可以考虑下，img元素的real src是否和全文img的real src一致。如果都来自src，则一致。如果个别来自非src，则以这个基准，将其他的标注为不一致，即为0。如果存在多个，取并集。只要和其中任何一个一致都可以。（这个特征不太好，除非统计下，含有这样特征的图片的比例有多少，如果比较多时，可以考虑，否则效果不好。）

4、<http://zx.cjn.cn/qiche/2017-04-20/2019839.html>，这个有点神奇，周围文字很丰富，为什么还提取错误了？可能是因为类似国家/地区，这种li非常短，导致位置比较靠后，可以考虑不计算这类的li。可以尝试将这个删除掉，然后重新分类下。

5、目前对于列表性质的图片，效果不好，还有就是图片周围文本过少，基本没有文本的情况。（**还因为目前没有对该类数据进行标注**）

6、图片的path路径属性信息需要增加

7、图片距离时间的特征需要增加。

8、<http://blog.sina.com.cn/s/blog_49a0573e0100051g.html>，这个有问题，文字多，没有抽取成功。**这个是不是页面解析有问题，考虑只统计tr，不算td/th，去除li过多的情况，合并其为一个ul/ol**

9、可以统计下连续图片的密度，以text问分割间隔（注意去除&nbsp;），让后看看连续的图片密度。**这个是有影响的**

可以考虑统计下图片的分布密度，就是将网页看成图片结点和文本结点的一维向量，然后统计下密度。

10、看一下这个页面去掉空格，和不去掉空格的分类准确率：

<http://blog.sina.com.cn/s/blog_8c5ce1a00101hsqt.html>

11、如果图片和图片之间的文本过少，可以考虑使用前面的文本，不过应该限制tag个数。（**这个暂时不考虑**）

12、尝试将<!—取消作为分割标签（**这个取消后，效果不好**）

13、提供了text block的上一个text block距离是指text block之间含有的tag数量，包含opentag和closetag

14、图片文件的后缀名，可以作为一个弱特征，因为.ico，较少可能是正文图片，除了专业ico图片。

15、图片周围文本长度的变化，就是用方差表示，文本长度的变化的方差

16、对于连续注释的文本，可以考虑合并为一个。注意中间不能有任何tag

17、对于path相同的（只考虑tag name、class、id）pic，在进行一次平滑处理。

开一个窗口，计算下text的个数，作为特征。