1. **如何利用这些文本特征：**

**利用<!--和script以及style作为简单分块标记。在分块的基础上统计图片周围的文本的分布信息。**

备选：

计算img周围文本属性时，可以考虑利用父结点具有id作为结束结点来统计属性，或者利用最多往上遍历1-5个父结点就作为终结结点。**可以统计下标注页面中，目标页面距离body页面多远，以及距离其他正文img相差几个父结点。**

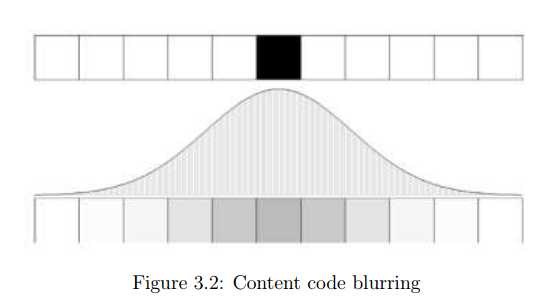
（2）正文图片在dom中的高度，一般应该算是相同，可以根据dom高度不同以及其他特点对图片进行聚类，然后在聚类的基础上构造特征。

（3）String Length Smooth技术：构造字符串的一维长度向量，然后利用平滑技术进行处理，最后利用方差进行平滑，并根据阈值判断是否是正文。对于遇到非文本元素，则直接赋值为0。

（4）对于论坛页面可以先进行识别，然后利用结点树相似性来定位论坛的正文区域。就是毕业设计采用的技术。

（5）Document Slope Curve，通过累加计算相应tag数量的变化，对于变化平滑的，可能是正文区域。可以将图片划分为正文区域和非征文区域，这个作为一个特征，将相应的概率值作为特征或者0/1值，即是/否作为特征

（6）Content Code Blurring，是通过将code和content进行0/1编码，然后利用函数进行调整。选择某些区域作为正文区域，这个也可以作为一个特征。



（7）统计每行文字和tag的比例。而不是以结点为单位。相对于平均值，利用标准方差作为阈值，同时在过滤前，利用平滑器做平滑处理。不过这种思路一般适用于文本区分度较搞的网页。来源：**Text Extraction from the Web via Text to Tag Ratio**

（8）可以统计下正文img和非正文img中，包含哪些常规和非常规属性，作为分类参考。

（9）通过Density Sum来计算可能是潜在正文，来源：DOM Based Content Extraction via Text Density。

（10）计算图片和潜在标题或者重要文本的距离，比如相隔多少个TextUnit或者picture，不能简单计算code 距离，因为如果列表性质的，后面图片表示时，就不正确。最好是父结点的距离比较好。

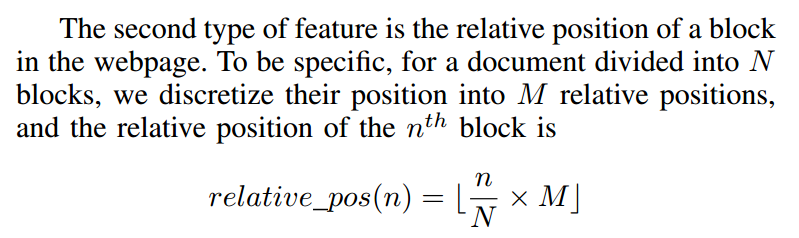
（11）首先通过多页面线下挖掘一批非正文的图片，比如logo/icon/广告等信息。主要是统计出现频次较高的图片。

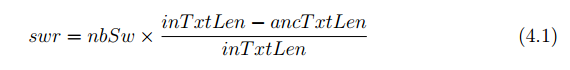
（12）归一化处理时，可以采用和最大值相除，或者和body中的数据相除。

（13）有些文本，太短，就包含一个字，如果单纯依靠比率，则可能有失偏颇，可以参考决策树中的方案来解决此问题。

（14）对本次特征进行归一化或者正则化时，可以针对页面整个统计量或者针对该页面该统计量的最大值。亦或是采用方差之类的。

1. 对于文本较少的网页，可以考虑将网页分为长文网页和短文网页
2. 通过正文抽取定位到的短文本，如果确实是正文，则该文本距离正文图片的距离应该不远。
3. 对于推荐和正文图片的区分可以添加：周围文本长度变化情况。因为推荐的布局一般较规整。
4. 对于html5，有些标签有特殊含义
5. 统计一下正文中属性的特殊分布。
6. 图片tag在html源码中的位置也很重要。因为太靠后一般不会是正文。

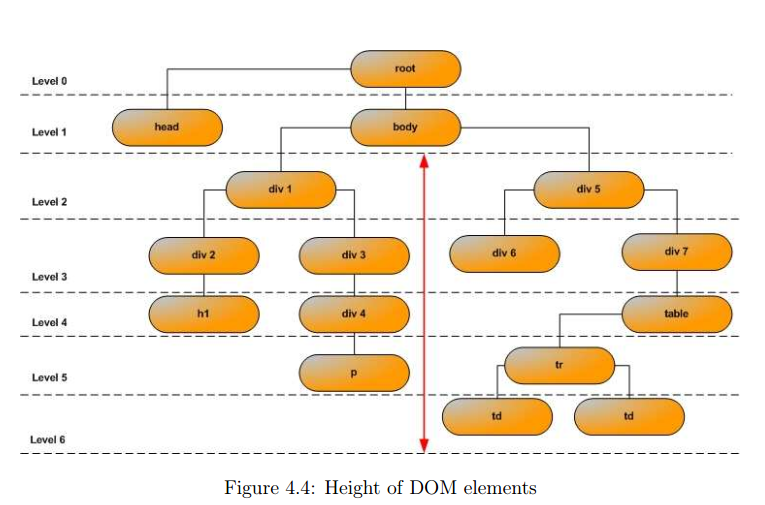


1. 非正文文字周围的图片也可能不是正文。
2. Stop word ratio：the ratio of full stops to the overall number of words。其他文档中对stop words ratio的定义：**stopWordRatio** The stopWordRatio is the number of stop words that contained in all of the text nodes of a particular DOM node. Based on our observation, typically, there are many stop words occur in the main content. However, there is a possibility that the stop words occur in the noisy contents such as recent comments section of blogs which rich of anchored text. Therefore, we introduce a penalty for stop words. Given the inner text length (*inTxtLen*), anchored text lengh (*ancTxtLen*), and number of stop words (*nbSw*), the stopWordRatio (*swr*) is calculated using the formula in Equation (4.1).  
   
3. Header around：header应该是指H1, H2, H3, H4, H5, H6等，该特征和h\*个数类似。其他文档：**headerAround**

This feature indicates whether there are any header elements near a particular DOM node. In order to do this, we check the parent node, sibling nodes, and children nodes of the DOM node whether any header element exist. In this case, the header element is denoted by h1, h2, h3, h4, and h5 tags. Based on our observation, most of the main contents such as news articles, blog posts have title around them. The title usually represented by header tags.

1. Dom height，其他文档中dom height的描述：**domHeight**

The domHeight is the maximum depth that can be reached a particular DOM node to a certain leaf node. The example of DOM tree with its associated depth level is shown in Figure 4.4, in the flgure, the DOM height of element body is flve. The explanation as follows, from body element, we can reach four leaves, namely h1 (at Level 4), p (Level 5), div 6 (Level 3), and two td elements (Level 6). Thus, the maximum depth is the path from body to td elements which give flve units height. In order to do segment classiflcation, we should select a certain DOM node which has good granularity level. The DOM height could be used as a feature to determine whether a DOM node is a good segment or not.



1. Link to text ratio，其他文档中：**linkToTextRatio**The linkToTextRatio in a DOM node is the ratio between the length of the anchored  
   text and the total length of the text in the DOM node. A DOM node which has high  
   value of link to text ratio most likely would be navigational elements and not main  
   contents.

同时这个属性在Link Quota Filter中也有使用。通过hyperlinked content和none hyperlinked content的比例来进行识别nav和adc部分。

1. 平均单词长度：Average word length（in our definition words are white-space delimited character sequences which at least contain one letter or digit）
2. 句子的平均长度，句子分割时，只是感觉简单的启发式规则average sentence length (the sentence boundaries are identified by a simple pattern-based heuristic checking for the presence of full stops, question or exclamation marks as well as semicolons) and **the absolute number of words**
3. 时间符号的比例：**the number of date/time-related tokens**
4. （这个是弱属性，暂时不提取）导航/菜单符号的比例：**the number of vertical bars “|”(these characters can sometimes be found in navigational boilerplate text).**
5. A标签的数量/全文A标签的数量：Moreover, we also compute the link density (called anchor percentage in [15]), as the number of to-kens within an A tag divided by the total number of tokens in the block;

其他论文提到的：linkNum

The linkNum is number of anchor element (hyperlink) inside a DOM node. If a DOM node contains quite a lot of links most likely it could be a navigational elements and not a main content.

1. （这个属性不可靠）CSS标签和属性使用的越多，越可能不是正文：The more CSS is used, the less important the semantics of an HTML tag becomes ——it is perfectly legal to only use DIV tags and describe the “semantics" of a particular division using style-sheet classes.

The presence of a particular headline tag (H1, H2, H3, H4, H5, H6), a paragraph tag (P), a division tag (DIV) and the anchor text tag (A) as an HTML element that encloses a particular text block.

18 、 **innerTxtLength，和the absolute number of words（第13条）**The innerTxtLength is the length of all of the text string in all of the text nodes in a  
DOM node. The length of the inner text in a DOM node can be an indication that  
it contains main content.  
**19 、（这个不需要，只采用innerTextLength就行）innerHtmlLength**The innerHtmlLength is the length of all HTML string between start tag and end  
tag of an object. The length of the inner HTML in a certain DOM node can be  
used as heuristic to determine whether a DOM node is a good segment in segment  
classiflcation or main content in content classiflcation.

20、**imgNum**The imgNum is the number of image elements inside a DOM node. Typically, headers  
and footers of web pages have images without text elements on it. Meanwhile, the  
main content sometimes has image in it as an illustration. Therefore, number of  
images can be used as a feature to determine whether a DOM node is a main content.

**21、interactionNum（包含input/select元素的个数，这些元素一般不会放到正文中）**The interactionNum is the number of input, select elements inside a DOM node.  
HTML elements such as input, select usually used to provide interaction between  
the user and the application. These kinds of interaction components are rarely part  
of the main content. Therefore, it could be useful as a feature to distinguish whether  
an element is a main content or noisy content.

**22、formNum**The formNum is the number of form element inside a DOM node.  
**23、optionNum（包含option元素的个数，类似select/input元素的个数分布）**The optionNum is the number of option element inside a DOM node. The motivation  
to choose optionNum as a feature also similar to interaction components and forms

**24、tableNum（包含table/tr/td/th元素的个数，用来识别是否位于正文区域）**The tableNum is the number of table, tr, td elements inside a DOM node. Before  
the extensive use of div tag and Cascading Stylesheet (CSS) in the modern web  
documents, tables are often used to define the layout structure of the document and  
as a placeholder for main contents. However, sometimes we still can find ad hoc web  
pages that still use tables for layout purposes. **Therefore we use this as a feature  
to determine whether a DOM node is a good segment and whether it contains main  
contents.**

**25、paraNum（包含p元素的个数）**The paraNum is the number of p element inside a DOM node. In web documents  
such as blogs, news websites, the main content usually is placed inside the paragraph  
element.  
26、**divNum**

The divNum is the number of div elements inside a DOM node.In the modern web documents, div tags are frequently used to define the layout structure of the document. **The number of div tags inside a DOM node can be used as a hint whether it is a good segment and main content.**

27、**tagName（这个感觉类似divNum，可以统计下li/ul的个数）不过在正文定位中是直接作为一个特征的，估计是利用one-hot进行特征构造。**

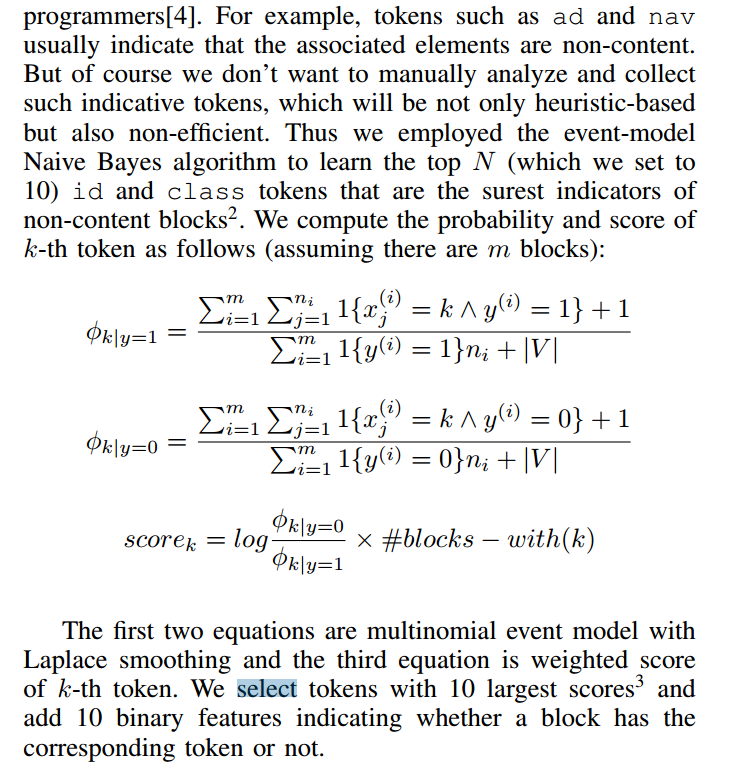
The tagName is the name of the HTML tag of the DOM node. Difierent HTML tags  
usually have difierent purposes for document presentation. For example, li and ul  
tag usually used to make a linked list of hyperlinks or navigational elements while  
div and table used to deflne the document layout structure and content. Therefore,  
HTML tag can be used as a feature to identify whether a DOM node is a good  
segment and whether it contains the main content.

28、img元素中包含的非常规属性的个数，比如data-src等这些属性。

29、统计包含的javascript数，一般是统计上下文，或者距离

30、统计包含css标签数。一般是统计上下文，或者距离

31、提取img到body的id列表/class列表，判断该列表中包含main content/none-content（nav，ico）意思的关键词数量。



A sample list of such tokens from our are: menu, module, nav, widget, sidebar, footer, right, dropdown, cat, nagivation.

将获取的属性值作为单独的列，转换成one-hot编码。

32、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

33、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。

34、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

35、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

