以下的描述中， xxx，yyy，zzz都用于表示我们需要处理的值。 … 表示任何我们不需要关心的内容。

1. 名词解释：

* 无关的标签：本软件不需要关心的标签，在任何位置或任何层次加添这些标签都不应该对本软件的输出产生任何影响。
* TagsPattern： html 的标签可以无限次嵌套，为了方便描述，我们用TagsPattern来表示某种固定的标签组合。

TagsValuePattern 中的外层标签和内层标签之间， 可能有也可能没有其它我们不需要关心的标签，为了准确的描述，我们在此约定我们的书写规范如下：

* 1. xxx 和 yyy 之间没有任何其它的标签：

<xxx>

<yyy>zzz</yyy>

</xxx>

* 1. xxx 和 yyy 之间有0个或多个和yyy处于同一级别的无关的标签， 这些无关的标签可以有也可以没有子标签。但是，yyy一定没有某个祖先标签是xxx的子孙标签：

<xxx>

…

<yyy>zzz</yyy>

…

</xxx>

* 1. xxx 和 yyy 之间有0个或多个和yyy处于同一级别或处于不同级别的无关的标签：

<xxx>

… …

<yyy>zzz</yyy>

… …

</xxx>

* TagValue-A: 当前标签的值可能被多个子标签分割为多个段，TagValue-A表示被分割的各段的字符串的和,不包括子标签的值。如：

<span>xxx<sub>zzz</sub>yyy </span>, 标签span的TagValue-A的值为xxxyyy

* TagValue-B: 当前标签的值可能被多个子标签分割为多个段，TagValue-B表示被分割的各段的字符串的和加上子标签的TagValue-A。 组合字符串时，以字符串出现的先后顺序为准， 如：

<span>xxx<sub>zzz</sub>yyy </span>, 标签span的TagValue-B为xxxzzzyyy

1. 需要处理其值的TagsPatte列表

任何时候，出现在需要处理TagsPatter列表之外的任何TagsPatter的值都需要被忽略。我们需要处理其值的TagsPatte列表如下：

* 名称：

<div class="webtop-g">

…

<h2 class="h">xxx</h2>

…

</div>

此TagsPatter可能在同一个html文件中出现在多次，我们只将当前文档的第一个匹配的TagsPatter的TagValue-B视为名称。

* 词性:

<div class="webtop-g">

…

<span class = "pos">xxx</span>

…

</div>

此TagsPatter可能在同一个html文件中出现在多次，我们只将当前文档的第一个匹配的TagsPatter的TagValue-B视为词性。

* 音标：

<div class="pron-gs ei-g">

…

<span class="pron-g" … geo=”n\_am”>

</span>

</div>

在div区间内，有两个 <span class="pron-g">区间， 我们只取出<span class="name">NAmE</span>的父标签（<span class="phon">）的TagValue-B作为音标。因为<span class="name">是需要忽略值的标签，所以” NAmE”需要从TagValue-B中去掉， 如本例中音标的值为：/yyy/

* 释义

一个释义包括当前释义以及当前释义相关的例句。每一个TagsPatter <li class="sn-g"> </li> 或者<span class="sn-g"> </span>的TagValue-B都是一个释义。

* 任何时候，以下TagsPatte的值都应该被忽略（虽然这些标签的值需要被忽略， 但是我们仍然可能需要读取这些标签，用来判断当前TagsPattern是否是我们需要关注的TagsPattern）：
  1. 所有包含css格式 {display: none} 的标签的值。
  2. <span class="collapse">， Example: “look” 的 “Synonyms”：

<span class="collapse" title="Synonyms"> … … </span>

* 1. <span class="idm-gs">， Example: “look” 的 “Idioms”：

<span class="idm-gs" id="look\_1\_\_426"> … … </span>

* 1. <div class="sound audio\_play\_button pron-us icon-audio">, Example: “look”的音标部分。
  2. [

("span", [("class", "pron-g”)])

("span", [("class", "prefix”)])

]

* 1. [

("span", [("class", "sym\_first")])

]

1. css格式控制

|  |  |  |
| --- | --- | --- |
| tag | css style | pattern |
| .webtop-g | ---- |  |
| h2 | ---- |  |
| .pos | ---- |  |
| .pron-g | ---- |  |
| .pron | ---- |  |
| .name | display: none | [  ("span", [("class", "name")]),  ], |
| .pron-gs .wrap | display: none | [  ("span", [("class", " pron-gs")]),  ("span", [("class", "wrap")]),  ], |
| .cf | display: inline |  |
| .exp | display: inline |  |
| .sn-gs | display: block |  |
| .sn-g | display: block |  |
| .sn-g:befor | {  content: counter(num) " ";  counter-increment: num;  } |  |
| .num | display: none |  |
| .gram-g | ---- |  |
| .gram | ---- |  |
| .wrap | ---- |  |
| .def | display: inline |  |
| .x-gs | display: block |  |
| .x-g | display: block |  |
| .x-g:befor | content: url(…); print(“: ”) |  |
| .rx-g | ---- |  |
| .x | ---- |  |
| .xr-gs | display: block |  |
| .xr-gs:befor | content: url(…); print(“-> ”) |  |
| .prefix | ---- |  |
| .oxford3000 | ---- |  |

* display: block 格式的处理办法：根据css对display: block的定义，我们需要在每次进入和退出带此格式的tag时都输出一个换行符号。 但是为了避免在输出的txt中看到多余的空行，我们在最终打印输出结果时， 我们将只有一个换行符号的空行忽略掉。

1. html 字符实体

|  |  |
| --- | --- |
| string | css style |
| &nbsp; | 空格 |
|  |  |
|  |  |

1. 容错性需求

* 测试发现, 源html文件有误, 几乎每个文件都有一行<meta name='viewport' content='width=device-width, initial-scale=1'>, 这个”meta” tag 就没有匹配的tag结束标志. 对于这种情况, 我们忽略掉此tag的开始标志, 程序继续处理其他的tag.

1. 可维护需求

可在短时间内增加/修改/删除 需要被处理或被忽略的TagsPatte。