网页去噪：

学长们的网页去噪实际上指的是去除html中html特有的诸如 <title>之类的语句，原理时是通过直接获取innerText，也就是从每一个形如<xx>内容</xx>的句子中获取内容。然而距离我们想象的去噪还是相去甚远。。诸如下列举了一个百度知道的例子（节选）

**before**

<meta name="description" content="如何检查单向链表有无环 cnblog1, 最简单的方法, 用一个指针遍历链表, 每遇到一个节点就把他的内存地址(java中可以用object.hashcode())做为key放在一个hashtable中. 这样当hashtable中出现重复key的时候说明此链表"/>

<link rel="canonical" href="http://zhidao.baidu.com/question/1736572532645081587.html"/>

<meta name="robots" content="noindex,follow" />

<script>

if (location.hash && location.hash.indexOf('bdres2exe=') > -1) {

document.write('<script src="//m.baidu.com/static/as/res2exe/js/res2exe\_1.0.3.min.js?v=091818"></' + 'script>');

}

</script>

<script type="text/javascript">

!function(){var n={},t={};n.context=function(n,e){var i=arguments.length;if(i>1)t[n]=e;else if(1==i){if("object"!=typeof n)return t[n];for(var o in n)n.hasOwnProperty(o)&&(t[o]=n[o])}},"F"in window||(window.F=n)}();;

**after**

change', arguments.callee); fn(); } }); } }; domReady(function(){ new HTTPSWarningLog(); for(var i=1; i<6; i++) { !function(i){ setTimeout(function(){ new HTTPSWarningLog(); }, i\*i\*i\*1000); }(i); } }); }(document, window); if (location.hash && location.hash.indexOf('bdres2exe=') > -1) { document.write('<script src="//m.baidu.com/static/as/res2exe/js/res2exe\_1.0.3.min.js?v=091818"></' + 'script>'); } !function(){var n={},t={};n.context=function(n,e){var i=arguments.length;if(i>1)t[n]=e;else if(1==i){if("object"!=typeof n)return t[n];for(var o in n)n.hasOwnProperty(o)&&(t[o]=n[o])}},"F"in window (window.F=n)}();; F.context('user', {"isLogin":"0","stoken":"","name":"","imId":"","id":"","wealth":"","gr