**爬取方案:通用爬虫的方式**

**url/翻页分析:**

[http://.\*?.blog.hexun.com/.\*?\_d.html](http://.*?.blog.hexun.com/.*?_d.html)

http://%s.blog.hexun.com/%s\_d.html %(username,article\_id)

<http://lfkls.blog.hexun.com/118353670_d.html>

http://whbsuwc.blog.hexun.com/118354180\_d.html

**数据提取:…**

标题

//span[@class=’ ArticleTitleText’]/a/text()

<span class="ArticleTitleText">

<a style="cursor: pointer;text-decoration: none;" href="[http://lfkls.blog.hexun.com/118353670\_d.html](view-source:http://lfkls.blog.hexun.com/118353670_d.html)">春雪喜降天宁寺</a>

链接

Request.url

article\_id

blog\_id

<script type="text/javascript" src="[http://cache-sidebar.blog.hexun.com/inc/ARecommend.aspx?blogid=3575809&articleids=118353670&t=2](view-source:http://cache-sidebar.blog.hexun.com/inc/ARecommend.aspx?blogid=3575809&articleids=118353670&t=2)"></script>

点击量

评论数

http://click.tool.hexun.com/click.aspx?articleid=118353670&blogid=3575809

无法直接访问,需加上refer和header

document.getElementById("articleClickCount").innerHTML = 720;

document.getElementById("articleCommentCount").innerHTML = 24;