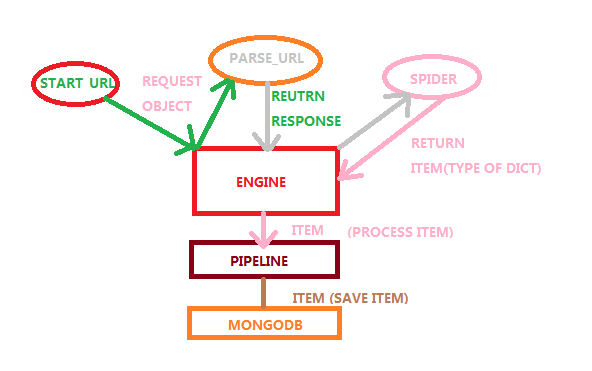
# Baidu\_Spider\_Log

## 爬取数据流程（PROCESS）



流程图

### 构造url地址

本次爬取的网站是百度贴吧，因此需要研究分析出网页的入口，确定哪个为起始的url地址。

通过分析，可以得到起始的url地址为：

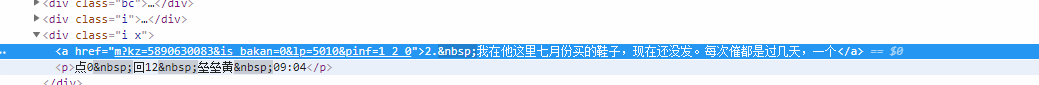
<http://tieba.baidu.com/mo/q----,sz@320_240-1-3---2/m?kw=cat&lp=5011&lm=&pn=0>



你会发现什么鬼，贴吧是这个样子的吗，哈哈，这个是贴吧极速版，没有广告，没有中间杂质，完完全全是纯粹的贴吧。而且利用手机版可以容易获取数据

### 访问这个网页

访问这个网页后，按下F12,开发工具，分析获取的数据在哪个地方，如图所示，titlezai a标签中的text中，链接在href属性中。



分析过后就可以构造一个dict对象，把数据保存进去即可。

### 访问每个获取到的每个内容的链接

得到内容链接之后，就可以访问内容链接了，得到的是如图所示的页面



我要获取的是其中的图片， 因此只要分析图片在哪个标签中，通过xpath就可以获取内容了。

### 获取下一页地址，并访问

我们爬取贴吧中的每个吧，肯定是要爬个底朝天，因此要获取下一页地址，才能够完整爬取整个贴吧。

打开开发工具，分析下一页，通过xpath获取下一页地址即可



然后访问它，并再次获取下一页地址，

当运行程序时，你会觉得奇怪，为什么访问了一页的贴吧，程序就停止了呢。为什么呢？？？？？

哈哈哈

这个问题刚开始我也是比较诧异的，琢磨了半天，还是觉得我的程序逻辑没有问题，响应也是在response中的，为什么呢？

原来在进入第二页的页面后，获取下一页的地址时，居然获取不到url地址，返回的是None,这也就导致了只爬取一个页面就停止程序了。

### 解决获取下一页问题

既然有问题，那就去解决，不要去想这个是不是百度在搞得鬼，反爬虫？？？，哈哈不要这么想，只要有问题就去想办法，然后去解决就够了，管它是谁搞得鬼，毕竟我门是爬取别人的网站，看透不说透嘛！！！

既然xpath获取不到下一页，那么在起始页的时候获取总页数，只要把这个总也数用一个字典保存起来，传递到下一个处理中，那么就解决了这个问题，当前页面只要从起始页面加一即可，而且也要保存进这个字典中，我的处理是这样的，如图所示：



这样就解决了获取下一页的问题