1.简介：

按照如图的栏目进行分别爬取

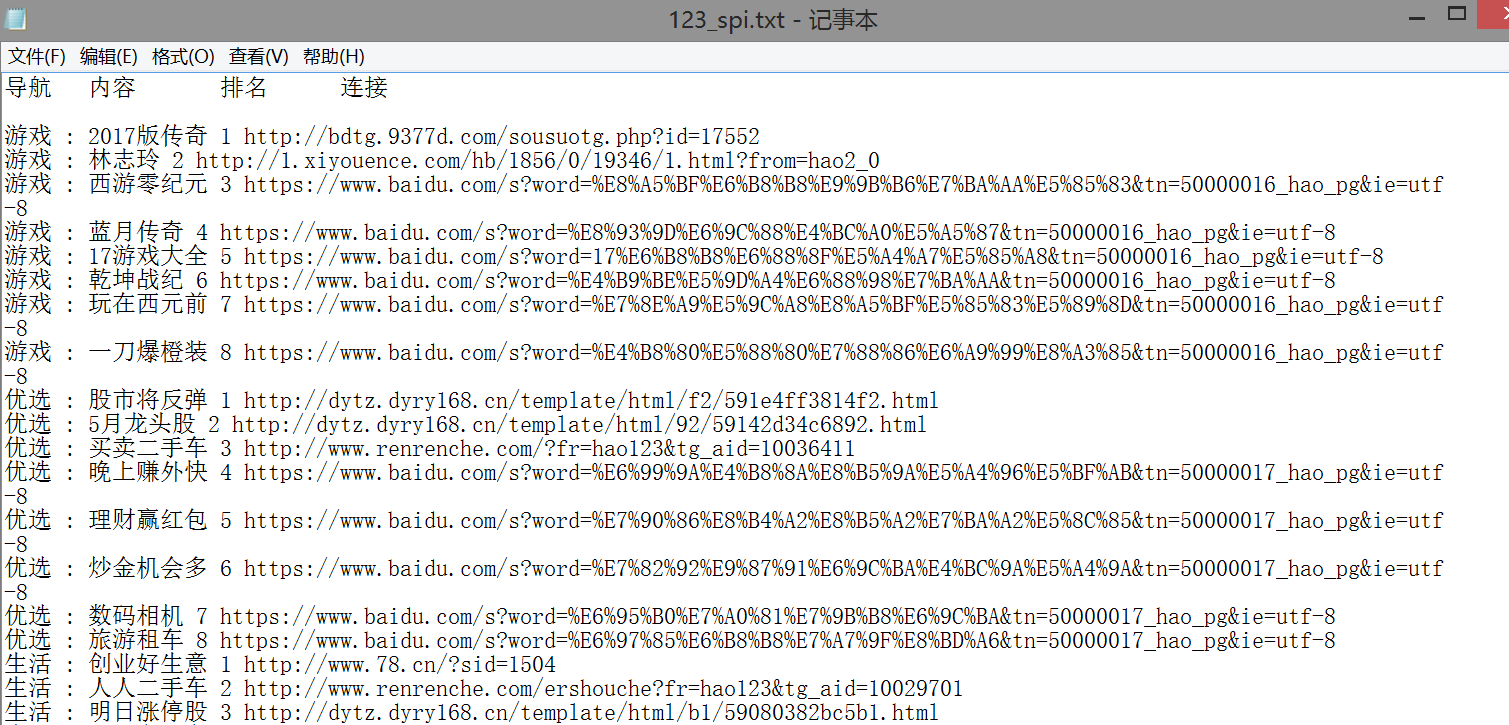
比如游戏的后边就是他的具体内容

每个内容都有自己的排名

比如传奇就是第一个 林志玲是第二个 

爬取结果：

先存入txt 而且每个字段之间是用空格隔开的虽然看着不方便 但是易于分析



1. 编程思想：

（1）首先我们来看一个导航他的代码

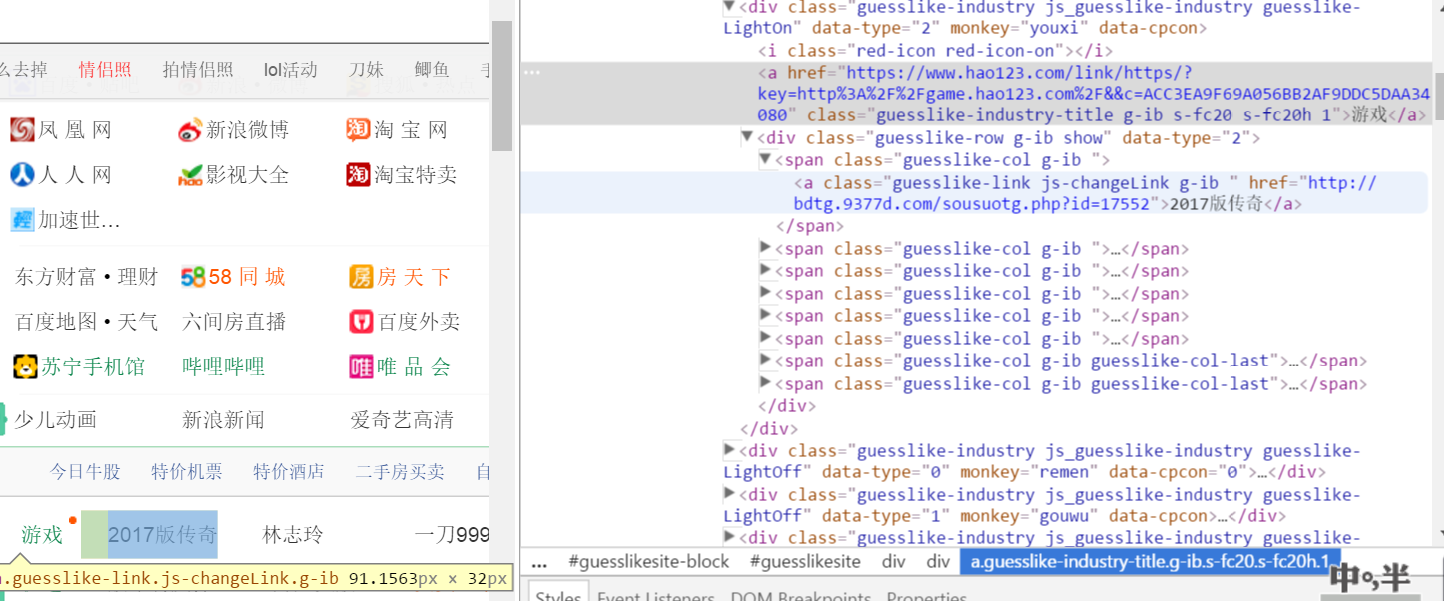
可以发现这个导航是一个a标签，

他的字段游戏可能不是唯一的，但是class + string（文字内容）是可以唯一确定这个标签的

因此使用beautifulsoup模块，使用find\_all函数首先确定该标签，

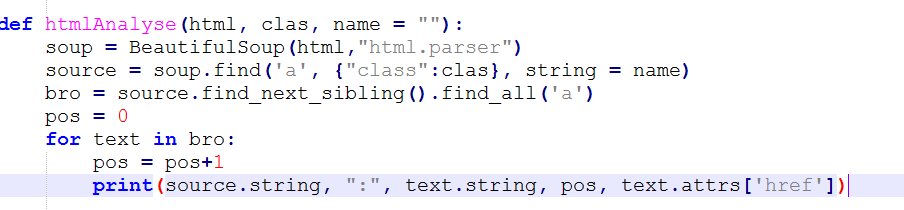


然后看看后边的内容和该标签的关系：

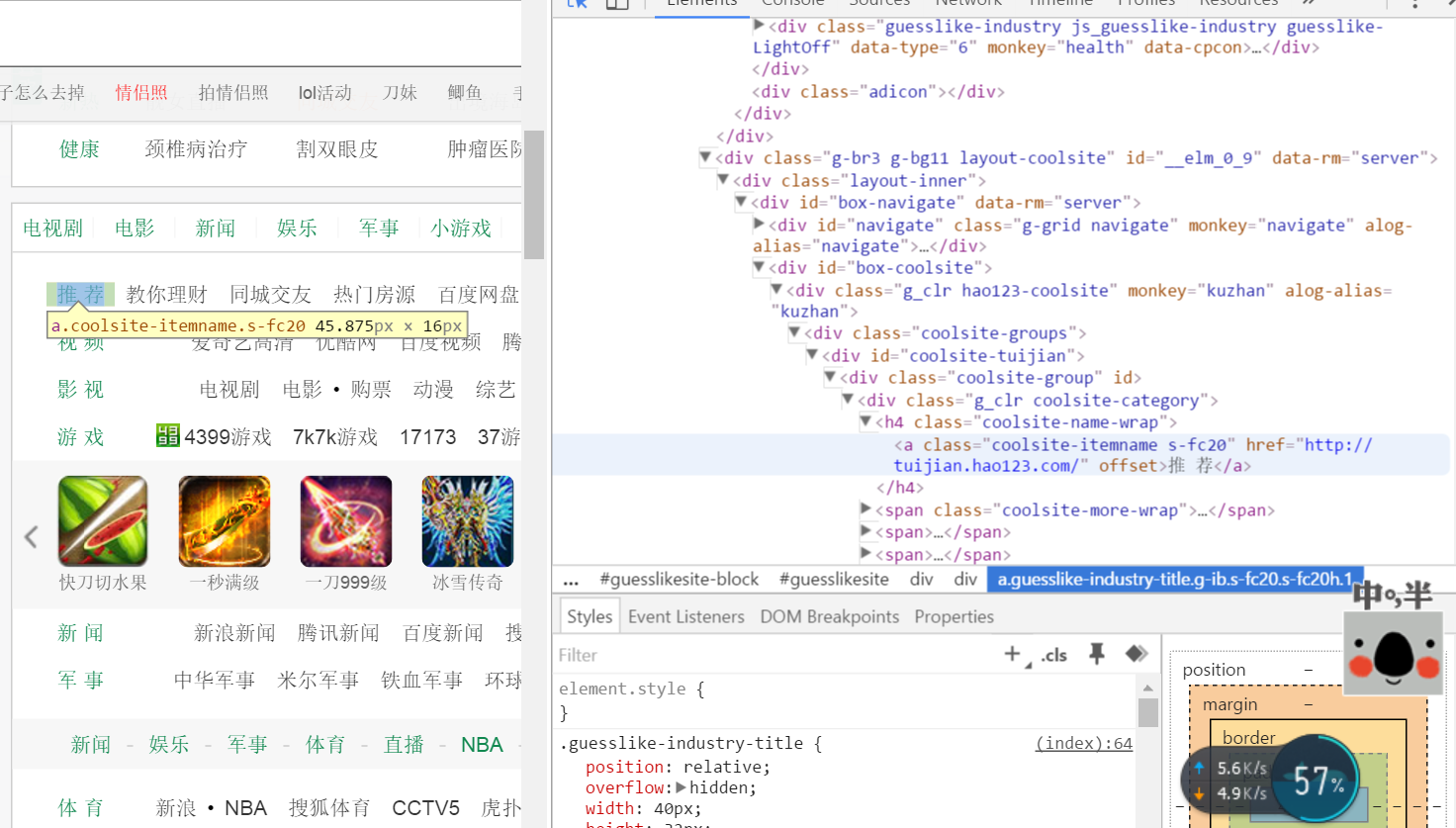


2017版传奇是上一个游戏标签的弟弟的儿子的儿子，之后的每一个内容都是该儿子的兄弟节点，因此直接遍历就可以

代码如下：



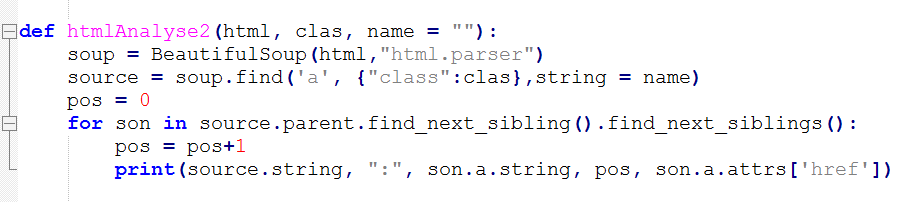
同理继续分析



从这里开始结构是和上边的结构不一样的

从这里开始的结构和下边的完全一样的，因此需要编辑第二个函数，在第一个基础上改一改

代码如下：



接下来是遇到的难点：

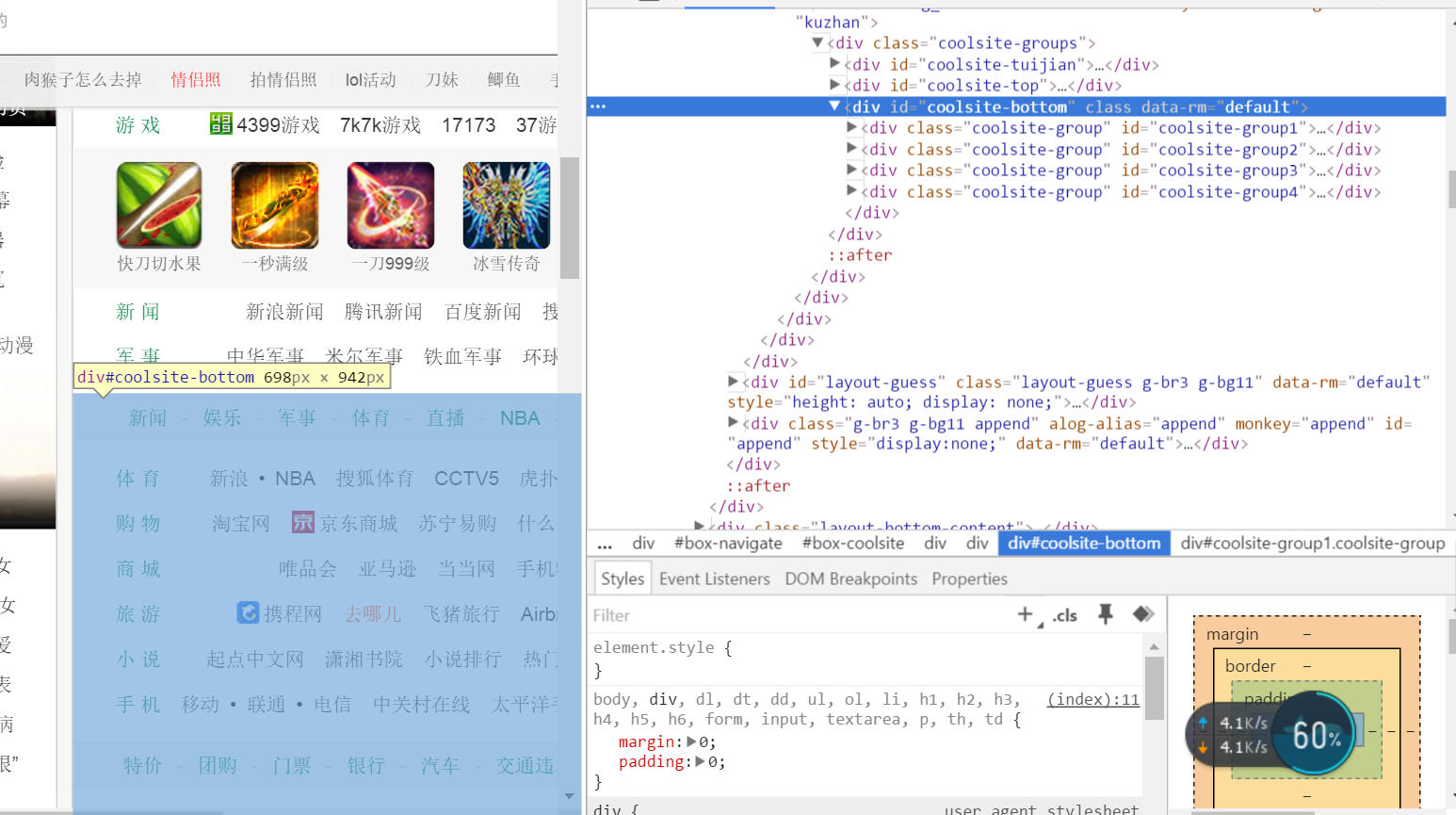
我们都知道hao123 是一个懒加载的网页 就是说如果你浏览器浏览不滑动鼠标的话，他下边并不会加载出来。

如图这是爬取的123网页没网的时候的样子



因此爬取军事以下的部分是非常难的。

因为原先的div是空的，就是这里



现在是有网的状态里边才会加载代码，如果是没网或者是爬虫的话里边是空的标签

那么我们可以看一下爬取的源代码 下边的部分去哪了



如图我们搜索关键词 发现这一部分代码还在html中 但是不在原先的div中

再一个code标签里

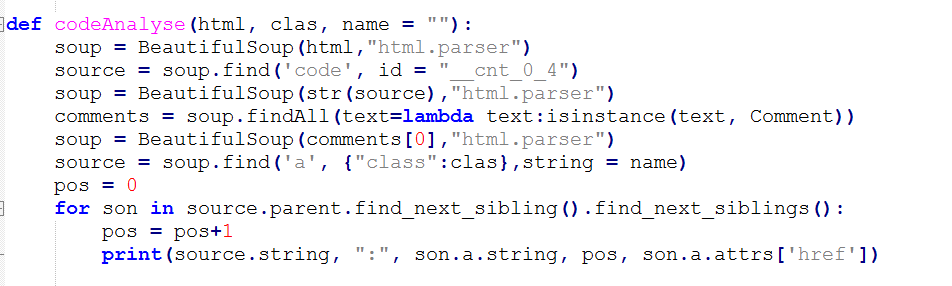
我想应该是浏览器在滑动鼠标的时候，js生成时间，然后代码讲这些加入到div中。那么解决方法有两种，第一种 在code中爬取，第二种模仿浏览器行为，进行爬取。老师告诉我第二种后边牵扯太多。因此采取第一种。

首先找到code 标签 通过唯一的ID

这里引入一个新的模块，因为代码中有注释，然后将注释中的内容提取出来



然后就和上边的第二种分析方式一样的，他们的结构相同



之后将每一个关键词，都存入链表中，进行循环就完成了。

因为是用class + string 进行唯一标识的爬取

