实验报告 1

一、实验目的

1. 了解基本的 html 语法，学习提供的 Parser 库 BeautifulSoup

二、实验内容

1. 学习万维网基本要素 HTML 的基本结构
2. 使用 BeautifulSoup 来抽取网页上的文本、连接、图片等内容。

三、实验环境

1. Firefox + Firebug 插件或 Chrome
2. Python 2.7 + easy_install + BeautifulSoup

四、实验步骤

首先更据提供的连接下载 python2.7 以及功能强大的 BeautifulSoup 插件，配置环境变量，开始实验。

这是我们首次接触此类与网络相关的课程，更据 word 和 ppt 上的例子，用百度的网页做样例，进行实验。从中了解 html 这一超文本标记语言的相关内容。了解了 tag，属性等基本的结构与语言结构，也知道了连接，图表，图片是如何表示的。

接着我们深入研究，通过 BeautifulSoup 处理网页，尝试解析 html 树，学习了用来匹配的正则表达式，以及一些正则表达式的书写格式。之后，我又尝试了许多别的网站，BS 的处理网页的功能真的不容小觑。

五、问题及其解决

课件的最后留有三道练习题，这正好也是对之前学习成果的一次检验。

第一题比较简单，搜索‘a’的标签名，把 href 的内容提取出来，写入 text 就大功告成了。然而，我遇到了不少的问题。比如会有形如//www.baidu.com/more/的不完整的地址出现。还有无关的 javascript:;出现了。

我搜索了相关的资料，前者的问题是这不是直接的地址，而是间接的,import urlparse 然后 urlparse.urljoin()利用这个函数即可迎刃而解。后一个问题我并没有找到有效的方法，就是添加了一个判断把含有 javascript 的不是地址的剔除，我仍会寻找更优的解决方案。

第二题找 img，比较容易。

第三题问题不少。首先，直接从那个百科网站上抓是不行的，得伪装成浏览器，才能顺利扒下来。毕竟是比较大型的网页，我们所需要的 tag，连接什么的都藏的很深，当时调用开发者工具用浏览器找的时候就花了不少时间，

由于发现百科的 tag 都有 qiushi_tag_这样的标识，通过正则表达式很快就能找到，接下来就是字符串的处理之类的，难度不是很大。

六、实验总结

电工导的课程实用性都非常强，上学期的 app 相关的入门，这学期的和网络，搜索这些要素息息相关，对平时我们大量的寻找资料等都有很大的帮助。但我感觉这门课程也不是那么容易的，应为我并没有什么基础。虽然我们大一上的时候学过 python，但是那是学的太浅，而且印象不够深刻，导致学了一学期 C++以后 python 的书写格式都忘的差不多了，还

是得花一些时间去熟练熟练，毕竟 python 是最近非常流行的语言之一。能通过计算机，来减少我们的人类的工作量，方便我们的生活，想必这也是计算机领域的魅力所在吧。

F1403023　5140309534　韩坤言