BeautifulSoup 和 lxml 选型文档

BeautifulSoup

优点：

1.易用性: BeautifulSoup 用起来比较简单，API 非常人性化

缺点：

1.因为 sgmllib 的问题，导致解析出错

```
from BeautifulSoup import BeautifulSoup
html = u'<a onclick="if(x>10) alert(x);" href="javascript:void(0)">hello</a>'
print BeautifulSoup(html).find('a').attrs
```

2.BeautifulSoup 是基于 DOM 的，会载入整个文档，解析整个 DOM 树，因此时间和内存开销都会大很多

lxml

优点：

1.速度比 BeautifulSoup 要快 10 倍(lxml 只会局部遍历，另外 lxml 是用 c 写的，而 BeautifulSoup 是用 python 写的，因此性能方面自然会差很多)

缺点：

1.lxml 也有它自己的问题，那就是多线程方面貌似有重入性问题，如果需要解析大量网页，那只能启动多个进程来试试了。(未测试，不确定是否是真的这样)

2.中文文档较少

BeautifulSoup4

优点：

1.可以选择解析器，如指定 lxml 为解析器

2. 易用性: BeautifulSoup 用起来比较简单，API 非常人性化，支持 css 选择器

3.有中文文档

缺点：

1.即使用了 lxml 解析器的 BS4 时间和 lxml 相差 6 倍

总结：

在 BeautifulSoup4 和 lxml 之间选择：

1.对于速度的需求： lxml 更好

10000 个最基本的网页解析

lxml： 1.0945 s

BS4 ：6.6950 s

2.对于准确性的需求 ：通过

“10) alert(x);"

href="javascript:void(0)">hello”

“<div class=我的 CSS 类>hello ”

3.**对于快速开发的需求：BeautifulSoup 更好**

4.根据文档结构树,判断网页是否更新 ：未测试

原来以为速度会相差很少，但是即使是用了 lxml 解析器的 BS4 速度还是达到了 6 倍,同时，因为需求比较确定，性能要求较高，所以最终应该选择 lxml。

参考文档：

http://stackoverflow.com/questions/1922032/parsing-html-in-python-lxml-or-beautifulsoup-which-of-these-is-better-for-wha
http://my.oschina.net/apoptosis/blog/118647?p=1

https://www.zhihu.com/question/26494302

https://www.crummy.com/software/BeautifulSoup/bs4/doc.zh/ 等