## baidu searching result form

based on the div class:

<div class="c-container">**【content】**</div>

# python程序：解析m.baidu.com文件

import re

'''

<div class="c-container">

...<h3 class="c-title c-gap-top-small">

标题

</h3>

<p class="。。。

内容

</p>

</div>

'''

import sys

file=sys.argv[1]

if file:

with open(file,"rb") as f:

s=b''

for line in f:

s+=line

su8=s.decode('utf8')

regexItem=r'<div\s\*class="c-container">.\*?(?=<div\sclass="c-container">)'

regexTitle=r'<h3 class="c-title[^>]\*>.\*?(?=</h3)</h3>'

regexContent=r'<p class="[^>]\*>.\*?(?=</p>)</p>'

items=re.findall(regexItem,su8,re.MULTILINE)

titles=[]

contents=[]

for item in items:

titles.extend(re.findall(regexTitle,item,re.MULTILINE))

contents.extend(re.findall(regexContent,item,re.MULTILINE))

i=1

for title,content in zip(titles,contents):

# print(i," : ",title.replace('\n','\\n'),"==>",content.replace('\n','\\n'))

i+=1

title=re.sub(r'<[^>]\*>',' ',title)

title=title.strip()

content=re.sub(r'<[^>]\*>',' ',content)

content=content.strip()

print("==============================")

print(title,":\n\n\t\t",content)

# 一般性的接口搜索

项目内涵名：fultonsearch

使用temp目录:~/temp/fultonsearch

获取网页：使用httplib2

分析网页：使用re

可变部分：每个网站的网页/每个请求的网页格式都是不同的，但是组成基本相同。单项构成

m.baidu.com/s : item{title,content}

结果：需要完整的网址和内容

常用的检索目标:（均取第一项，并统一使用www作为访问主机前缀）

百度百科

知乎

豆瓣

stackoverflow

百度网盘

csdn博客

新浪博客

开源中国 http://www.oschina.net