1. 爬虫要求

对2000多个网站（网站地址存储在txt文件里）按照广度优先策略，爬取其前100个页面的文本以及网页源码，每个网站爬到100个页面就停止。

重复的url不要爬取，每个网站只要爬站内页面，不要爬取出栈链接。

1. 环境要求

python 3.6, scrapy 1.3.3 , 最好用分布式爬取，在windows环境下运行。

1. 数据要求

存储到mysql中的websites表中，数据格式：

siteurl char(100),# 网站网址，作为爬虫的starturl

pageurl char(200),# 网站的某一个网页

content text,#网页的内容，即去除html tag之后的所有文本

html text,#网页源码

primary key (siteurl,pageurl)

关于爬虫数据抽取的代码已写好：

**def** parse\_item(self, response):

#add some code here

……

#获取网页源码

html=response.text  
soup=BeautifulSoup(html,**'lxml'**)

#清除script标签下的文本

**for** script **in** soup.find\_all(**'script'**):  
 script.clear()

**#清除style标签下的文本**

**for** style **in** soup.find\_all(**'style'**):  
 style.clear()

**#去除所有标签，留下文本**

content=re.sub(**'<.\*?>'**,**''**,formal)  
content=**''**.join(content.split())

！！！！最终提供程序代码和mysql中的数据