爬虫项目第二阶段报告

第11组 组员：陈靖野、舒心、盛意林

我们使用webmagic的爬虫框架，爬取weibo.cn页面上的数据，以及weibo.cn上的链接页面。可以爬取的数据包括用户名、发布时间、发布设备、微博内容等。得到的数据可以进行分析，比如统计微博用户是使用什么设备发微博的。由于其中可能有动态加载的内容，需要使用selenium的技术。

小组分工：组员一起讨论，查阅相关资料，舒心负责搭建webmagic的框架，陈靖野和舒心主要负责编写爬虫程序，预计花一周左右的时间。然后一起做数据分析，预计在一周内完成。最后组员一起写文档。

难点：动态加载的内容的获取。

webmagic+selenium的Demo：

package us.codecraft.webmagic.samples;

import us.codecraft.webmagic.Page;

import us.codecraft.webmagic.Site;

import us.codecraft.webmagic.Spider;

import us.codecraft.webmagic.downloader.selenium.SeleniumDownloader;

import us.codecraft.webmagic.pipeline.FilePipeline;

import us.codecraft.webmagic.processor.PageProcessor;

/\*\*

\* 花瓣网抽取器。<br>

\* 使用Selenium做页面动态渲染。<br>

\* @author code4crafter@gmail.com <br>

\* Date: 13-7-26 <br>

\* Time: 下午4:08 <br>

\*/

public class HuabanProcessor implements PageProcessor {

private Site site;

@Override

public void process(Page page) {

page.addTargetRequests(page.getHtml().links().regex("http://huaban\\.com/.\*").all());

if (page.getUrl().toString().contains("pins")) {

page.putField("img", page.getHtml().xpath("//div[@id='pin\_img']/a/img/@src").toString());

} else {

page.getResultItems().setSkip(true);

}

}

@Override

public Site getSite() {

if (null == site) {

site = Site.me().setDomain("huaban.com").setSleepTime(0);

}

return site;

}

public static void main(String[] args) {

Spider.create(new HuabanProcessor()).thread(5)

.addPipeline(new FilePipeline("/data/webmagic/test/"))

.setDownloader(new SeleniumDownloader("/Users/yihua/Downloads/chromedriver"))

.addUrl("http://huaban.com/")

.runAsync();

}

}