**package** com.lshs.mytest;  
  
**import** com.lshs.entry.Urlstable;  
**import** com.lshs.service.UrlsTableService;  
**import** org.jsoup.Connection;  
**import** org.jsoup.Jsoup;  
**import** org.jsoup.nodes.Document;  
**import** org.jsoup.nodes.Element;  
**import** org.jsoup.select.Elements;  
**import** org.junit.Test;  
**import** org.junit.runner.RunWith;  
**import** org.springframework.beans.factory.annotation.Autowired;  
**import** org.springframework.test.context.ContextConfiguration;  
**import** org.springframework.test.context.junit4.SpringJUnit4ClassRunner;  
  
**import** java.io.\*;  
**import** java.net.HttpURLConnection;  
**import** java.net.MalformedURLException;  
**import** java.net.URL;  
**import** java.net.URLConnection;  
**import** java.text.SimpleDateFormat;  
**import** java.util.ArrayList;  
**import** java.util.List;  
  
*/\*\*  
 \** ***@Description:*** *网页爬取工具类 爬取彼岸图网的4K图片总共是1037页数据的图片 http://pic.netbian.com/  
 \** ***@author:*** *LuShao  
 \** ***@create:*** *2018-08-16 21:10  
 \*\*/*@RunWith(SpringJUnit4ClassRunner.**class**)  
@ContextConfiguration(locations = {**"classpath:applicationContext.xml"**})  
**public class** MyTest {  
*// https://movie.douban.com/tv/#!type=tv&tag=%E7%83%AD%E9%97%A8&sort=recommend&page\_limit=20&page\_start=0  
// https://movie.douban.com/tv/#!type=tv&tag=%E7%83%AD%E9%97%A8&sort=recommend&page\_limit=20&page\_start=20* @Autowired  
 **private** UrlsTableService **service**;  
  
 */\*\*  
 \* 得到列表页的 list里面的a href  
 \** ***@param url*** *\** ***@return*** *\** ***@throws*** *IOException  
 \*/* **public static** List<String> getManyAdress(String url) **throws** IOException {  
 **long** sss = System.*currentTimeMillis*();  
 List<String> list=**new** ArrayList<>();  
 Document document =**null**;  
 **try** {  
 document = Jsoup.*connect*(url).get();  
 } **catch** (Exception e) {  
*// e.printStackTrace();* **return null**;  
 }  
*// 根据className得到元素内容得到这个链接的页面内容* Elements elementsByClass = document.getElementsByClass(**"list"**);  
*// Elements lis = document.getElementsByTag("li");* **for** (Element byClass : elementsByClass) {  
 Elements bs = byClass.getElementsByTag(**"b"**);  
 Elements as = bs.get(0).getElementsByTag(**"a"**);  
 **for** (Element a : as) {  
 String href1 = a.attr(**"href"**);  
 **if**(href1.contains(**".htm"**)){  
 String href = **"http://www.netbian.com"** + href1;  
 list.add(href);  
 }  
 }  
 }  
 **return** list;  
 }  
  
 */\*\*  
 \* 存第一级首页链接  
 \** ***@return*** *\*/* **public static** List<String> urlList(){  
 List<String> list=**new** ArrayList<>();  
 list.add(**"http://www.netbian.com/2560x1440/"**);  
 list.add(**"http://www.netbian.com/1920x1080/"**);  
 list.add(**"http://www.netbian.com/1680x1050/"**);  
 list.add(**"http://www.netbian.com/1600x900/"**);  
 list.add(**"http://www.netbian.com/1440x900/"**);  
 list.add(**"http://www.netbian.com/1366x768/"**);  
 list.add(**"http://www.netbian.com/1280x1024/"**);  
 list.add(**"http://www.netbian.com/1280x800/"**);  
 list.add(**"http://www.netbian.com/1280x720/"**);  
 list.add(**"http://www.netbian.com/1024x768/"**);  
 **return** list;  
 }  
  
 */\*\*  
 \* 得到单个图片的 href  
 \** ***@return*** *\*/* **public static** String getSingleUrl(String url) **throws** IOException {  
 Document document = Jsoup.*connect*(url).get();  
 List<String> alist=**new** ArrayList<>();  
 Elements elementsByClass = document.getElementsByClass(**"pic"**);  
 Elements a1 = elementsByClass.get(0).getElementsByTag(**"a"**);  
 String href = a1.get(0).attr(**"href"**);  
  
 String s = **"http://www.netbian.com"** + href;  
 **return** s;  
 }  
  
 */\*\*  
 \* 得到最后一张大图的URL  
 \** ***@param url*** *\** ***@return*** *\** ***@throws*** *IOException  
 \*/* **public static** String getOnePic(String url) **throws** IOException {  
 List<String> endList=**new** ArrayList<>();  
 Document document =**null**;  
 **try** {  
 document = Jsoup.*connect*(url).get();  
 }**catch** (Exception e){  
 **return null**;  
 }  
 Elements table = document.getElementsByTag(**"table"**);  
 Element endimg = document.getElementById(**"endimg"**);  
 Elements a = endimg.getElementsByTag(**"a"**);  
 String href1 = a.get(0).attr(**"href"**);  
 **return** href1;  
 }  
  
 **private static** SimpleDateFormat *sdf*=**new** SimpleDateFormat(**"yyyy-mm-ddhhmmss"**);  
  
 @Test  
 **public void** demo1() **throws** IOException {  
*// List<String> urllist = urlList();* **long** time1=System.*currentTimeMillis*();  
 String surl=**"http://www.netbian.com/1366x768/"**;  
 String spx = surl.replace(**"http://www.netbian.com/"**, **""**).replaceAll(**"/"**, **""**);  
 String fdir=**"f:/webfile/imgs"**+ File.***separator***+spx;*//F:\webfile\imgs\2560x1440\* File dir=**new** File(fdir);  
 **if** (!dir.exists()) dir.mkdirs();  
 String txt=fdir+File.***separator***+spx;*//F:\webfile\imgs\2560x1440\2560x1440* List<String> alists=**new** ArrayList<>();  
 **for** (**int** i=1;i<=200;i++){  
 String url=**""**;  
 **if**(i==1) url=surl;  
 **else** url=surl+**"index\_"**+i+**".htm"**;  
 List<String> aList = *getManyAdress*(url);  
 **if**(aList!=**null**) alists.addAll(aList);  
 }  
  
  
 String s1=txt+**"\_1.txt"**;*//F:\webfile\imgs\2560x1440\2560x1440\_1.txt* PrintWriter pw1=**new** PrintWriter(**new** FileWriter(s1),**true**);  
 List<String> singles=**new** ArrayList<>();  
 String s2=txt+**"\_2.txt"**;*//F:\webfile\imgs\2560x1440\2560x1440\_2.txt* PrintWriter pw2=**new** PrintWriter(**new** FileWriter(s2),**true**);  
 **for** (String a1 : alists) {  
 pw1.println(a1);  
 pw1.flush();  
 String singleUrl = *getSingleUrl*(a1);  
 System.***out***.println(**"中型图~~~"**+singleUrl);  
 pw2.println(singleUrl);  
 singles.add(singleUrl);  
 pw2.flush();  
 }  
 pw1.close();  
 pw2.close();  
  
 **long** time2=System.*currentTimeMillis*();  
 System.***out***.println(**"共爬取"**+alists.size()+**"条记录~~~"**);  
 **long** sss=(time2-time1)/1000;  
 System.***out***.println(**"总共用时（秒）"**+sss);  
  
 }  
  
  
 */\*\*  
 \* 大图的地址存入数据库  
 \** ***@throws*** *IOException  
 \*/* @Test  
 **public void** demo2() **throws** IOException {  
 List<String> urls = getUrls();  
*// String txturl="F:\\webfile\\imgs\\2560x1440\\2560x1440\_3.txt";* **for** (String url : urls) {  
 String onePic = *getOnePic*(url);  
 **if**(onePic!=**null**&&(onePic.contains(**".jpg"**))){  
 System.***out***.println(**"大型图~~~"**+onePic);  
 Urlstable table=**new** Urlstable();  
 table.setTxtPath(**"1366"**);  
 table.setUrlPath(onePic);  
 table.setScreenpx(**"1366x768"**);  
 **service**.add(table);  
 }  
 }  
 }  
  
 */\*\*  
 \* 根据文件整理url  
 \** ***@return*** *\** ***@throws*** *IOException  
 \*/* **public** List<String> getUrls() **throws** IOException {  
*// String txturl="F:\\webfile\\imgs\\2560x1440\\2560x1440\_2.txt";* String txturl=**"F:\\webfile\\imgs\\1366x768\\1366x768\_2.txt"**;  
 BufferedReader br=**new** BufferedReader(**new** FileReader(txturl));  
 String line=**null**;  
 List<String> list=**new** ArrayList<>();  
 **while** ((line=br.readLine())!=**null**){  
 list.add(line);  
 }  
 **return** list;  
 }  
  
 */\*\*  
 \* 从数据库取出url并下载图片  
 \** ***@throws*** *IOException  
 \*/* @Test  
 **public void** demo3() **throws** IOException {  
 Urlstable table=**new** Urlstable();  
 table.setScreenpx(**"1366x768"**);  
 List<Urlstable> select = **service**.select(table);  
 **long** secound=System.*currentTimeMillis*();  
  
 **for** (**int** i=0;i<select.size();i++){  
 **if**(i>=100){  
 **if**(i%100==0) secound=System.*currentTimeMillis*();  
 }  
 String dir=**"f:/webfile/download/imgs/1366x768/"**+secound;  
 File fdir=**new** File(dir);  
 **if** (!fdir.exists()) fdir.mkdirs();  
 Urlstable urlstable = select.get(i);  
 String urlPath = urlstable.getUrlPath();  
 **int** index = urlPath.lastIndexOf(**"/"**);  
 String img = urlPath.substring(index + 1);  
 String fimg = fdir + File.***separator*** + img;  
  
 System.***out***.println(**"开始下载~~~"**+urlPath+**" 下载到~"**+fimg);  
 URL url=**new** URL(urlPath);  
 URLConnection connection = url.openConnection();  
 InputStream is = connection.getInputStream();  
  
 BufferedOutputStream bos=**new** BufferedOutputStream(**new** FileOutputStream(fimg));  
 **byte**[] b=**new byte**[1024];  
 **int** size;  
 **while** (-1 != (size = is.read(b))) {  
 bos.write(b, 0, size);  
 }  
  
 bos.close();  
 is.close();  
 }  
 }  
  
*// public static void main(String[] args) throws IOException {  
//  
//// String url="http://pic.netbian.com/";//第一页的网址  
//// http://pic.netbian.com/index\_2.html 第2-1037页的网址  
//// String time=sdf.format(new Date());  
//  
// List<String> urllist = urlList();  
// long time1=System.currentTimeMillis();  
// for (String surl : urllist) {  
//  
// System.out.println(surl);  
//  
// String replace = surl.replace("http://www.netbian.com/", "");  
// String spx = replace.replaceAll("/", "");  
// String fdir="f:/webfile/imgs"+ File.separator+spx;//F:\webfile\imgs\2560x1440\  
// File dir=new File(fdir);  
// if (!dir.exists()) dir.mkdirs();  
// String txt=fdir+File.separator+spx;//F:\webfile\imgs\2560x1440\2560x1440  
//  
// List<String> alists=new ArrayList<>();  
//  
// for (int i=1;i<=200;i++){  
// String url="";  
// if(i==1) url=surl;  
// else url=surl+"index\_"+i+".htm";  
// List<String> aList = getManyAdress(url);  
// if(aList!=null) alists.addAll(aList);  
// }  
//  
//  
// String s1=txt+"\_1.txt";//F:\webfile\imgs\2560x1440\2560x1440\_1.txt  
// PrintWriter pw1=new PrintWriter(new FileWriter(s1),true);  
// List<String> singles=new ArrayList<>();  
// String s2=txt+"\_2.txt";//F:\webfile\imgs\2560x1440\2560x1440\_2.txt  
// PrintWriter pw2=new PrintWriter(new FileWriter(s2),true);  
// for (String a1 : alists) {  
// pw1.println(a1);  
// pw1.flush();  
// String singleUrl = getSingleUrl(a1);  
// System.out.println("中型图~~~"+singleUrl);  
// pw2.println(singleUrl);  
// singles.add(singleUrl);  
// pw2.flush();  
// }  
// pw1.close();  
// pw2.close();  
//  
//  
//// String regex= "(http|https)://[\\w+\\.?/?]+\\.[A-Za-z]+";  
//  
// List<String> end=new ArrayList<>();  
// String s3=txt+"\_3.txt";  
// PrintWriter pw3=new PrintWriter(new FileWriter(s3),true);  
// for (String single : singles) {  
// String onePic = getOnePic(single);  
// System.out.println("大型图~~~"+onePic);  
// pw3.println(onePic);  
// end.add(onePic);  
// pw3.flush();  
// }  
// pw3.close();  
// long time2=System.currentTimeMillis();  
// System.out.println("共爬取"+end.size()+"条记录~~~");  
// long sss=(time2-time1)/1000;  
// System.out.println("总共用时（秒）"+sss);  
//// for (String s : end) {  
//// System.out.println(s);  
//// }  
// }  
// }*}

package com.lsh.control;

import java.text.SimpleDateFormat;

import java.util.Date;

import java.util.regex.Matcher;

import java.util.regex.Pattern;

import java.io.BufferedReader;

import java.io.File;

import java.io.FileWriter;

import java.io.InputStreamReader;

import java.io.PrintWriter;

import java.net.\*;

public class GetHttpUrl {

private static SimpleDateFormat sdf=new SimpleDateFormat("yyyy-mm-ddhhmmss");

public static void crawling(String url,String regex,String filename) {

String time=sdf.format(new Date());

String fdir="f:/webfile/"+time;

String urlTxt=fdir+File.separator+filename+"\_url.txt";

String contentTxt=fdir+File.separator+filename+"\_content.txt";

URL url1=null;

URLConnection connection=null;

BufferedReader br=null;

PrintWriter pw1=null;

PrintWriter pw2=null;

Pattern pattern=Pattern.compile(regex);

try {

url1=new URL(url);

connection=url1.openConnection();

File dir=new File(fdir);

if(!dir.exists()) dir.mkdirs();

pw1=new PrintWriter(new FileWriter(urlTxt),true);

pw2=new PrintWriter(new FileWriter(contentTxt),true);

br=new BufferedReader(new InputStreamReader(connection.getInputStream()));

String line=null;

while((line=br.readLine())!=null) {

pw1.println(line);

Matcher matcher = pattern.matcher(line);

while(matcher.find()) {

pw2.println(matcher.group());

}

}

System.out.println("爬取成功~~~");

} catch (Exception e) {

e.printStackTrace();

}finally {

try {

br.close();

pw1.close();

pw2.close();

} catch (Exception e2) {

e2.printStackTrace();

}

}

}

public static void main(String[] args) {

// String url="https://movie.douban.com/";

// String url="https://movie.douban.com/j/search\_subjects?type=tv&tag=国产剧&page\_limit=50&page\_start=0";

// String url="http://www.netbian.com/dongman/index\_3.htm";

String url="http://www.netbian.com/1366x768/index\_2.htm";

String regex= "(http|https)://[\\w+\\.?/?]+\\.[A-Za-z]+";

crawling(url, regex, "douban");

}

}