Homework 4

- Web搜索引擎

  目标：

  – 1. 写一个Web爬虫，爬取问答类或课程类网站的网页；
  – 2. 解析网页内容，对内容进行结构化，并存储到文件中；
  – 3. 为内容建立索引；
  – 4. 通过命令行进行内容检索，并展示内容列表

Homework 4

- 网站：
  - 丁香园：https://dxy.com/diseases/
  - 新浪爱问（**理工学科**）：https://iask.sina.com.cn/c/202.html
  - 新浪爱问（医疗健康）：https://iask.sina.com.cn/c/79.html
  - 百度知道
  - 知乎

工程科技类问答

  - CMU课程
    - CMU CS的Faculty: https://www.cs.cmu.edu/directory/all
    - CMU CS的courses: http://coursecatalog.web.cmu.edu/schoolofcomputerscience/courses/
    - 注：
  - MIT课程
    - MIT CS的courses: http://courses.csail.mit.edu/
  - Berkeley课程
    - Berkeley CS的courses: https://eecs.berkeley.edu/academics/courses

Homework 4

- 针对QA，则需要爬取
  - 问题
  - 答案
  - 来源网站
  - 领域

其他更多的内容（如提问者、回复者、点赞数等）

- 针对课程，需要爬取
  - 课程名
  - 课程简介
  - 学校名
  - 领域

Homework 4

- 关键技术：
  - 爬虫
  - 信息抽取
  - 索引建立
  - 查询

Homework 4

- Tips：
- 1. 如何在Eclipse中引入jar包

Homework 4

- Tips
- 2. JAVA爬虫
  - crawler4j
    - https://github.com/yasserg/crawler4j

crawler4j

build passing | maven-central v4.4.0 | chat online

crawler4j is an open source web crawler for Java which provides a simple interface for crawling the Web. Using it, you can setup a multi-threaded web crawler in few minutes.

  - JSOUP
    - https://blog.csdn.net/zbx931197485/article/details/78582407

    - jsoup 是一款 Java 的HTML 解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据，可以看作是java版的jQuery。
      jsoup的主要功能如下：
      从一个URL，文件或字符串中解析HTML；
      使用DOM或CSS选择器来查找、取出数据；
      可操作HTML元素、属性、文本；
      jsoup是基于MIT协议发布的，可放心使用于商业项目。官方网站：http://jsoup.org/

Homework 4

- Tips:
- 3. 利用开源软件对网页中的**正文**进行抽取

(1).Cx-extractor(http://cx-extractor.googlecode.com):基于行块的分布来提取网页中的正文。

– 提取的方法是首先使用Jsoup来获取网页的内容，之后将内容传给cx-extractor，交由其来解析，核心代码如下所示：

```
1 // 通过Jsoup来获取html，在此设置了范文数据包的头部，因为有些网站会屏蔽爬虫。
2 String content = Jsoup.connect("http://www.chinanews.com/gj/2014/11-
    19/6791729.shtml").userAgent("Mozilla/5.0 (jsoup)").get().html();
3 // html_article即为解析出的正文。
4 String html_article = CXTextExtract.parse(content);
```

结果：这个库有时候会有错误，会将不属于正文的内容提取出来，例如一些无关的底部内容，或者一些链接。但性能比较高，约几十毫秒。

Homework 4

- Tips:
- 2. 利用开源软件对网页中的正文进行抽取

  (2).Boilerpipe(http://code.google.com/p/boilerpipe/):

  - 基于网页dom树来解析，内部有多种解析器，比较准确，但是时间在100毫秒左右。
  - 核心代码如下所示：
  - 1 String content = Jsoup.connect("http://www.chinanews.com/gj/2014/11-19/6791729.shtml").userAgent("Mozilla/5.0 (jsoup)").get().html();
  - 2 // 使用Bolierpipe来获取网页正文内容
  - 3 String parse_article = ArticleExtractor.INSTANCE.getText(content);
  - 结果：结果比较准确，性能比稍慢，大约在100毫米左右。

Homework 4

- Tips:
- 2. 利用开源软件对网页中的正文进行抽取

(3). 其他java开源代码

JReadability： https://github.com/wuman/JReadability

Java-readability：https://github.com/basis-technology-corp/Java-readability

JReadability is a Java library that parses HTML as input and returns clean, easy-to-read text.

EXAMPLES

Instantiate the `Readability` class via any one of the provided constructors, depending on where the interested HTML page is from:

```
Readability readability = new Readability(html); // String
Readability readability = new Readability(url, timeoutMillis);  // URL
```

Start content extraction by running:

```
readability.init();
```

The output is clean, readable content in HTML format. You can obtain the output with:

```
String cleanHtml = readability.outerHtml();
```

Homework 4

- 基于jsoup: Java HTML Parser来抽取信息 (如标题等，相同的网站同一个模板)，利用正则表达式来建立模板
  - https://jsoup.org/

```
File input = new File("/tmp/input.html");
Document doc = Jsoup.parse(input, "UTF-8", "http://example.com/");

Elements links = doc.select("a[href]"); // a with href
Elements pngs = doc.select("img[src$=.png]");
  // img with src ending .png

Element masthead = doc.select("div.masthead").first();
  // div with class=masthead

Elements resultLinks = doc.select("h3.r > a"); // direct a after h3
```

Homework 4

- Tips
- 3. 利用Lucene对文本进行索引，并进行检索

Apache Lucene™ is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

http://lucene.apache.org/core/

Homework 4

- 3. 利用Lucene对文本进行索引，并进行检索（输入检索词，查询得到相关的问题（或课程）列表，并显示详细信息。
  - 建索引和检索的简例

Homework 4

- 作业包括：java文件 + 文档 + 数据
- 作业打包上传到ftp homework/homework4下
- 文件：学号_姓名_homework4.rar

Homework 4

- 代码要求：
  - 遵守编程规范，如命名、注释等规范
  - 遵守面向对象的设计原则
  - 考虑异常处理等应用

Homework 4

- 文档要求：
  - 按附件格式样例，至少包括：引用、总体设计、详细设计、测试与运行、总结
  - 包括：数据格式说明
  - 附加：程序中包含的其他特色或改进，可加分
  - 附加：数据的丰富程度，可加分