Homework 4

- 编写Web搜索引擎

  目标：

  - 1. 写一个简单Web Crawler，爬取某个网站（如www.cs.zju.edu.cn）的网页；
    - 注意：重复网页的问题；爬取深度的问题
  - 2. 通过命令行进行网页内容检索，并展示网页列表。

Homework 4

- Tips：
- 1. 如何在Eclipse中引入jar包

Homework 4

- Tips:
  - 2. 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

- 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
  - 建索引和检索的简例

Homework 4

- 作业包括：java文件 + 文档
- 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
  - 附加：程序中包含的其他特色或改进，可加分