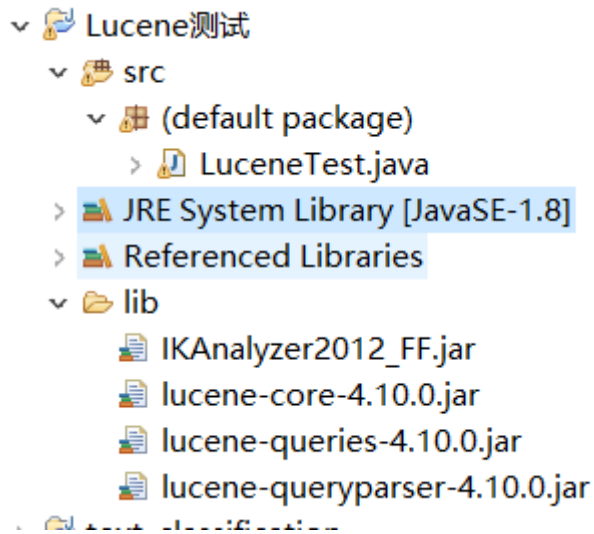
Homework 作业补充

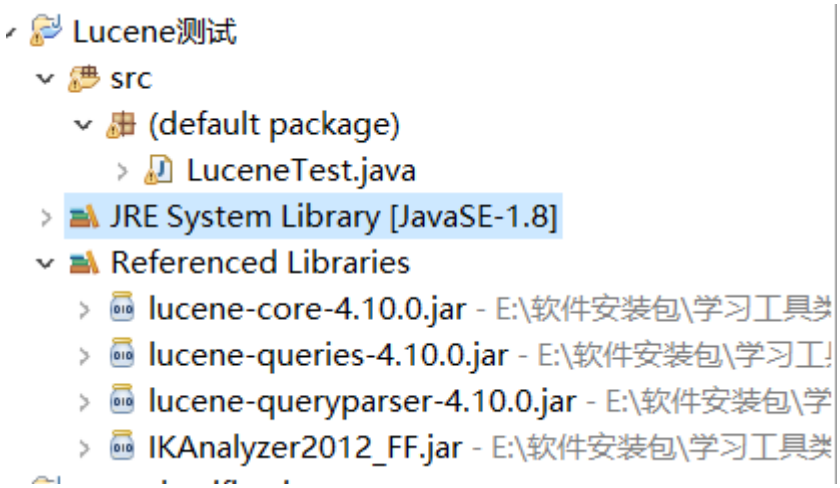
一，在 Eclipse 引入 jar 包

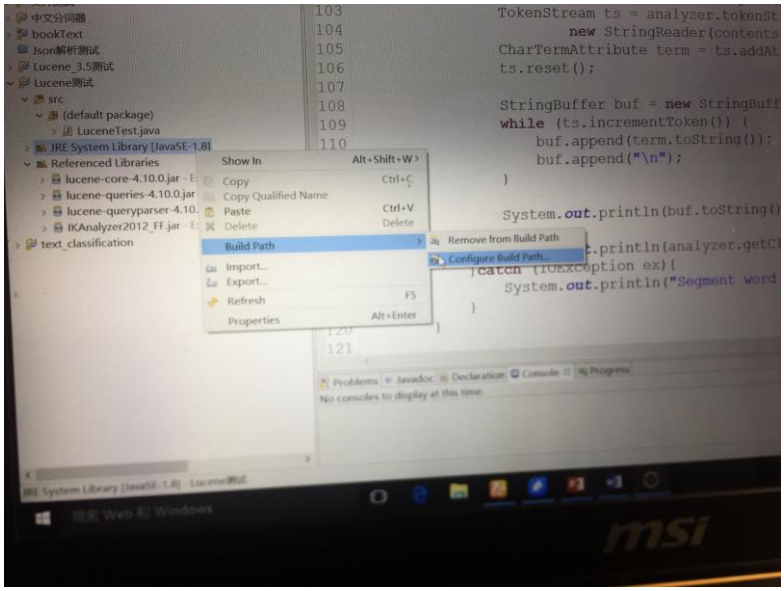
1， 下载需要的 jar 包。在 lucene 中，需要 Lucene-core-4.10.jar 等三个 jar 包以及 IKAnalyzer2012FF.jar 分词器 jar 包。本次作业所有的 jar 包均会直接提供。

2， 在工程中新建文件夹，命名为 lib（右单击项目，new->folder）。将下载的 jar 包复制粘贴到此文件夹中。

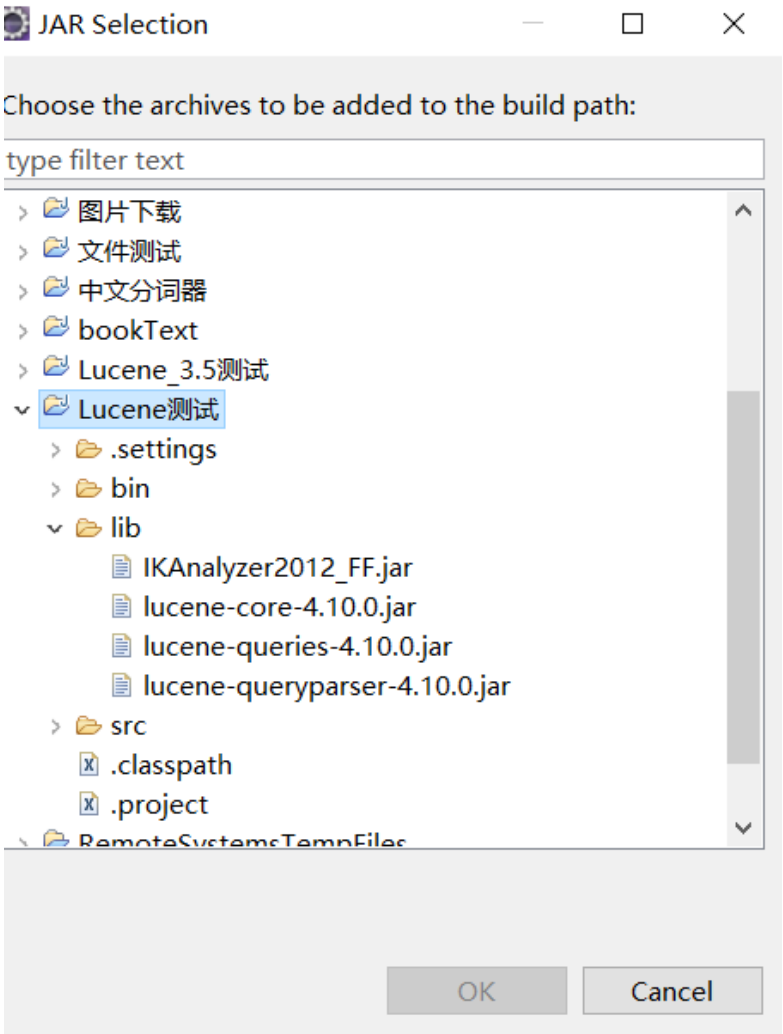


3， 在新建项目中的 JRE System Library 中右单击，选择 Build Path，再选择 configure Build Path。





3， 选择 Add　JARs（第一个选项）。将项目中的 lib 添加入工程中。如下图所示。



至此，导入 jar 包工作完毕。可以直接使用 jar 包工具

二，Lucene 使用

1， 本作业会提供一个简单的 Lucene 的基础工程，同学们使用 File->import->General/existing projects into workspace 即可将工程导入 eclipse 中。

2， Lucene 的使用主要分为两部分，创建索引和使用索引查询。其中创建索引需要使用中文分词，但这个不需要考虑，工具包已经提供好（IKAnalyzer2012_FF.jar）代码直接按照样本中使用即可。

2.1 在创建索引过程中，首先提供的是一个目录，用于存储索引；

2.2 然后建立索引写文件入口 IndexWriter；（这一部分直接按照样本就行了）

2.3 创建索引过程中会使用 Field，这些 Field 就是用于查询的依据。一般直接使用 TextField 即可。样本中还使用 StringField，这个用法是只有查询内容与原文全匹配才符合要求。用于 id 等索引。关键词一般都用 TextField。

2.4 Document 是索引的基本单位。在搜索过程中会先指定 field，然后 Lucene 在不同的 Document 进行搜索，直到搜索结果数目达到要求。在样本中因为测试缘故，只有一个 document。

2.5 在搜索过程中，首先建立 IndexSearcher，指定 field，关键词，通过一系列固定的操作即可完成搜索。

三，网页解析

1， 使用 Biolerpipe 工具进行解析。添加 jar 包过程同上，jar 包已经提供，并且基础的 demo 也已经提供

2， 这个只用于抽取正文，使用起来也十分方便。（代码其实我也是从网上综合了几个参考，复制粘贴的。经过测试，可以使用）。如果在以后的工作中，要抽取具体的内容，比如页面中某些信息，建议使用 jsoup，定位会更加精确（当然现在不需要掌握太复杂的功能。还有的网页 jsoup 也解析不了，就需要使用正则表达式进行爬取）。