倒排索引和布尔检索

——莎士比亚全集搜索

1511274 朱钰颖 计算机科学与技术

一、实验环境

Java 1.8

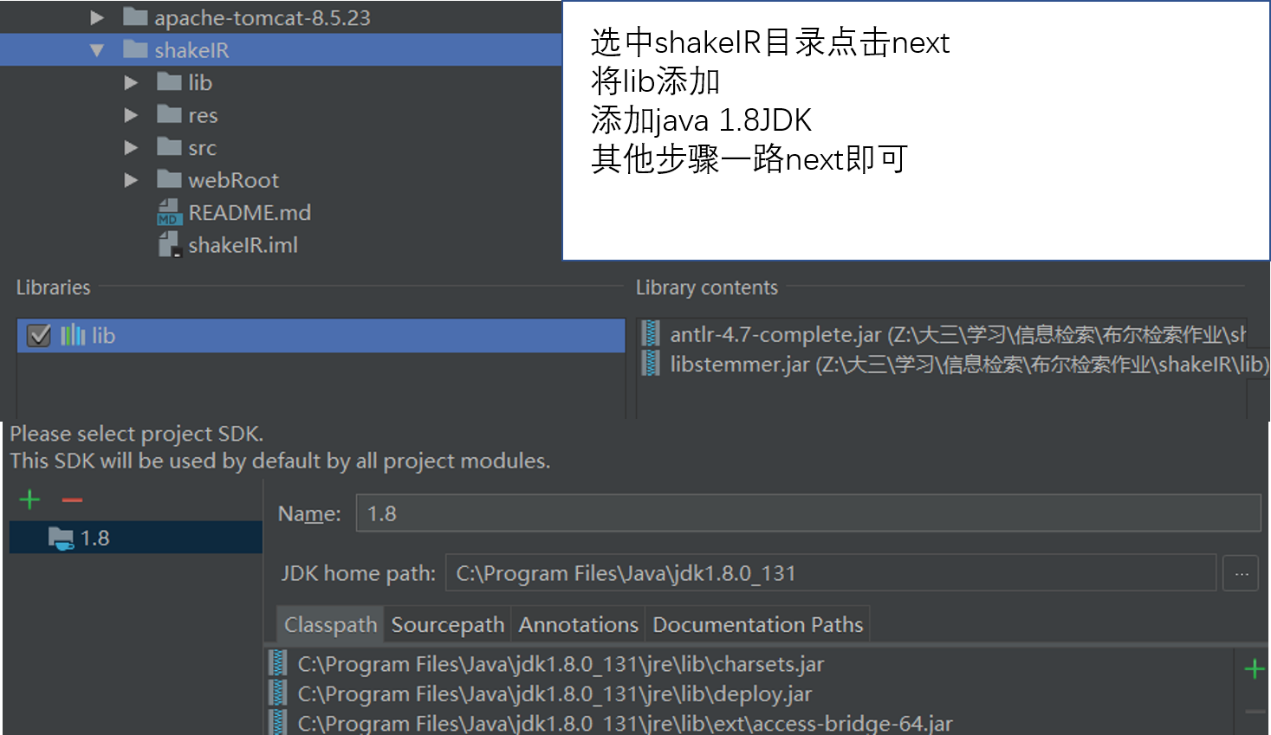
IDE Intellij IDEA Ultimate版本

apache-tomcat-8.5.23

antlr-4.7-complete

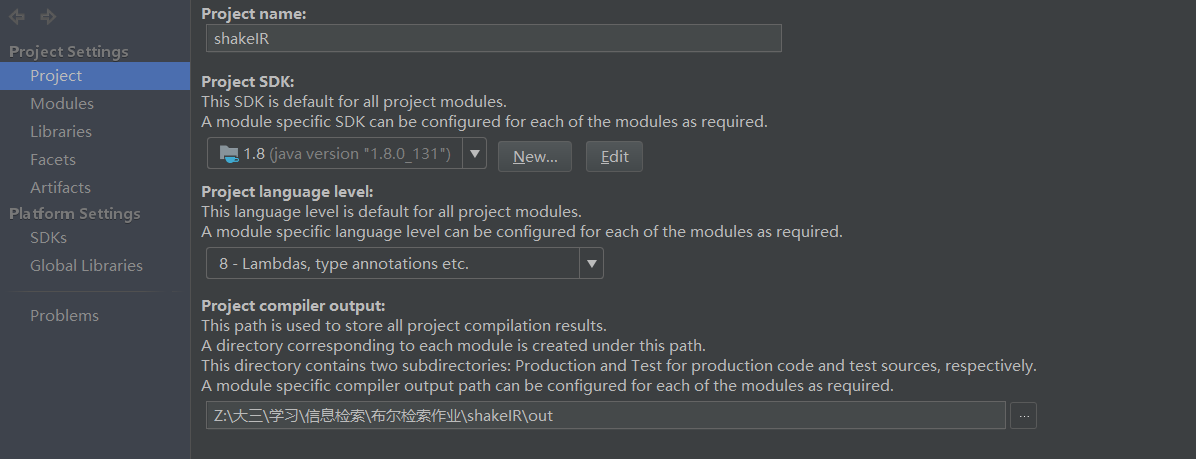
二、环境配置步骤

打开intellij后，import project,初始化步骤如下图所示

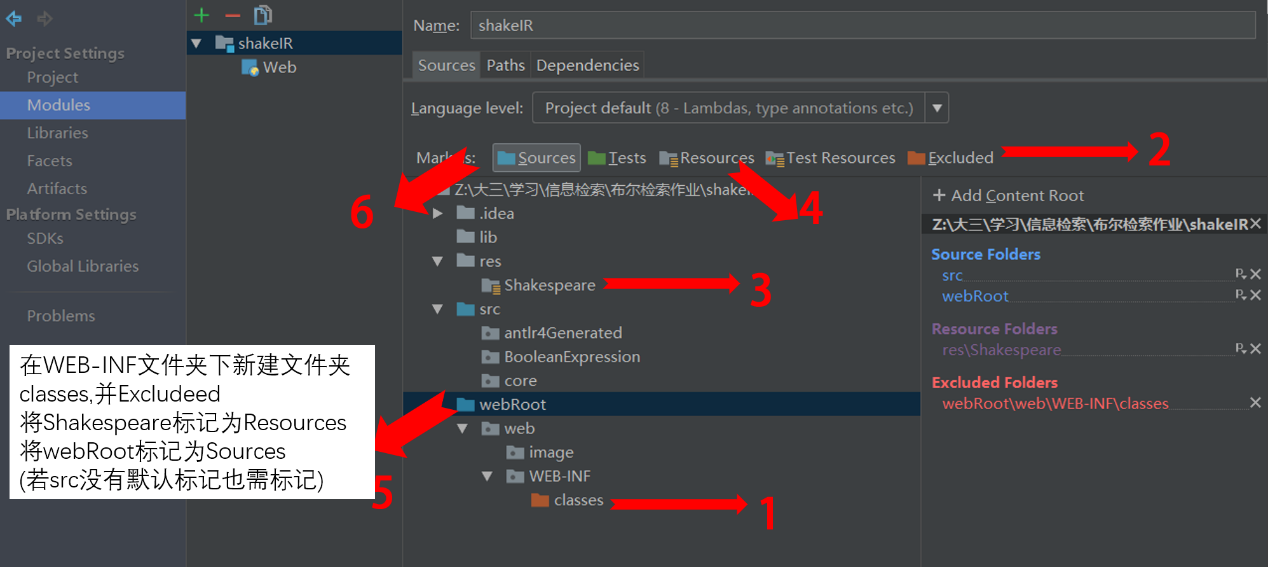


成功进入intellij后，点击File->Project Structure

Project设置如下

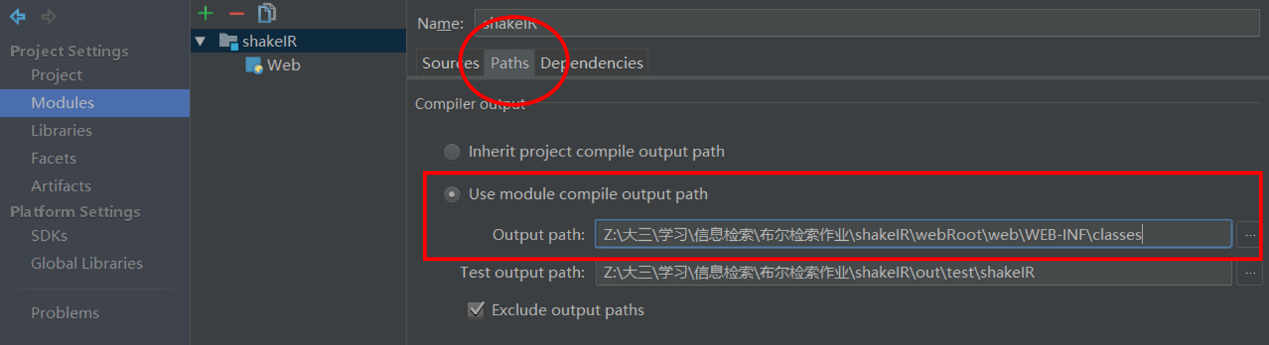


Modules模块设置如下

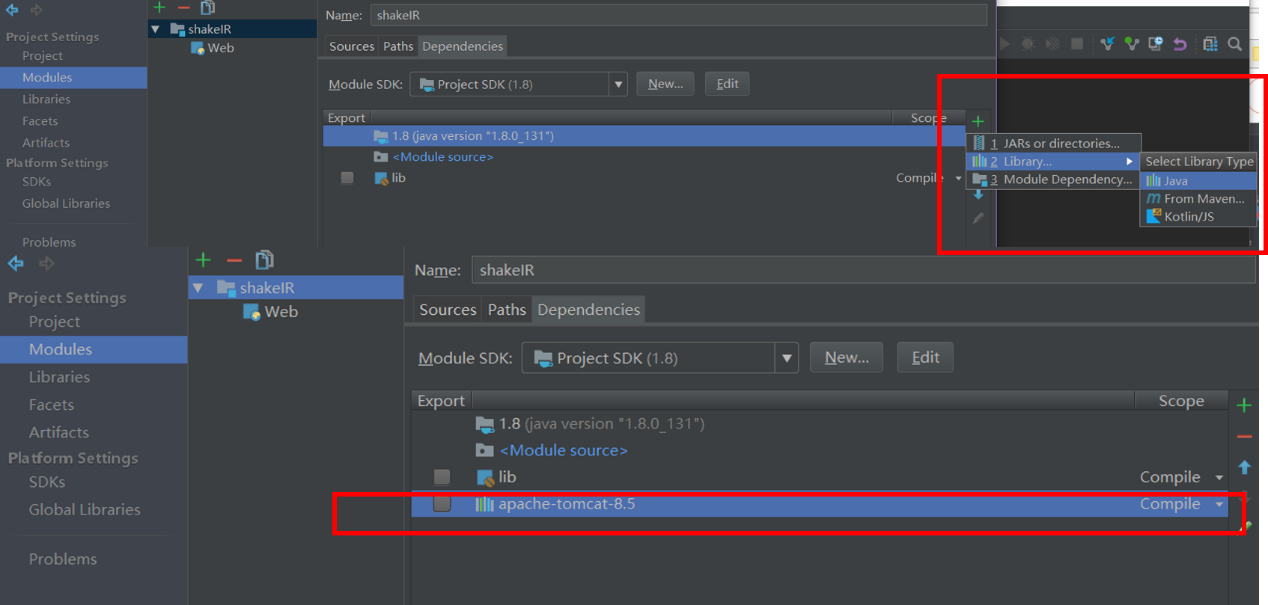


**※更正：将Shakespeare的上层目录res标记为Resources，而不是Shakespeare文件夹**

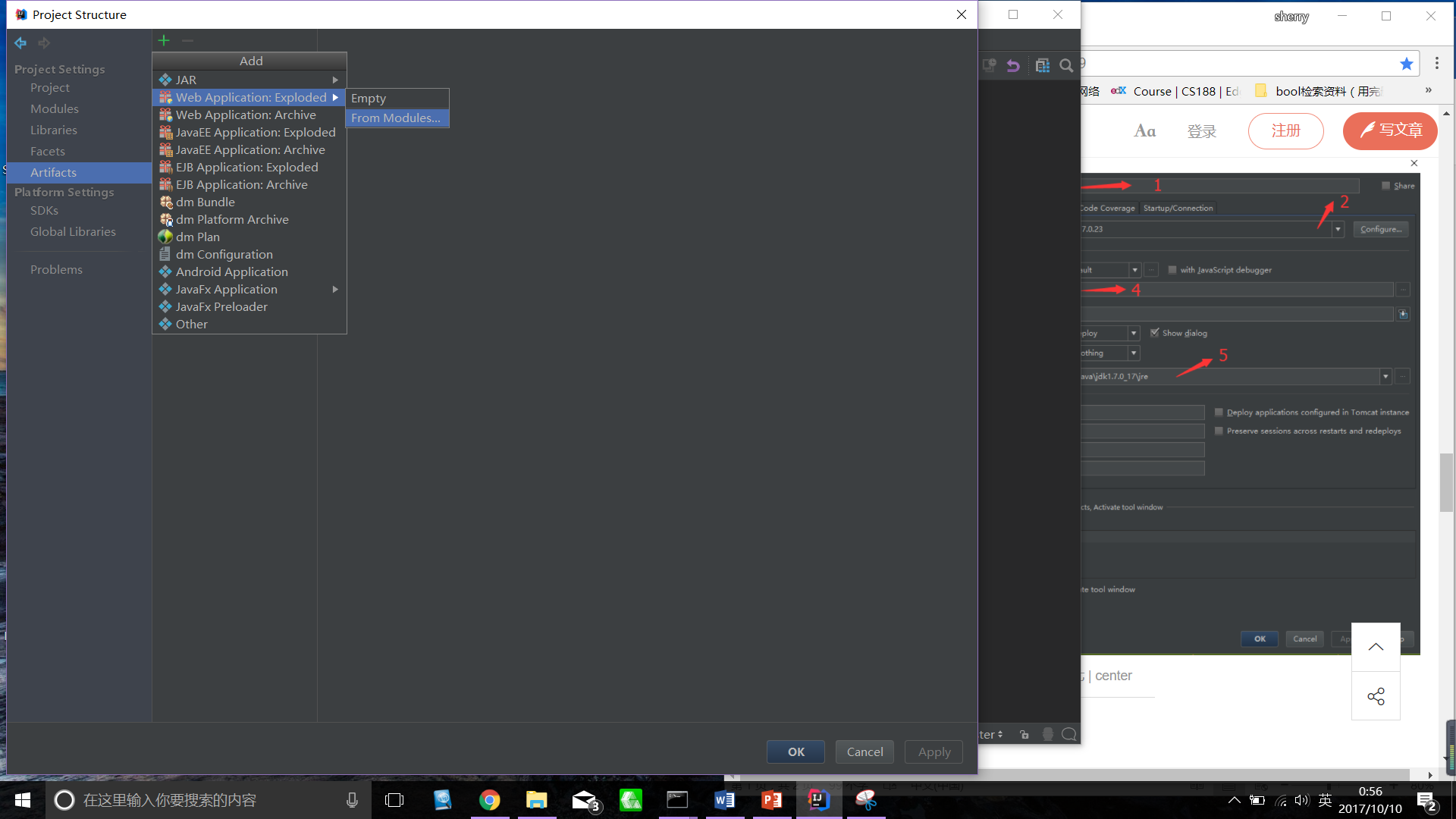
设置Path



设置Dependencies

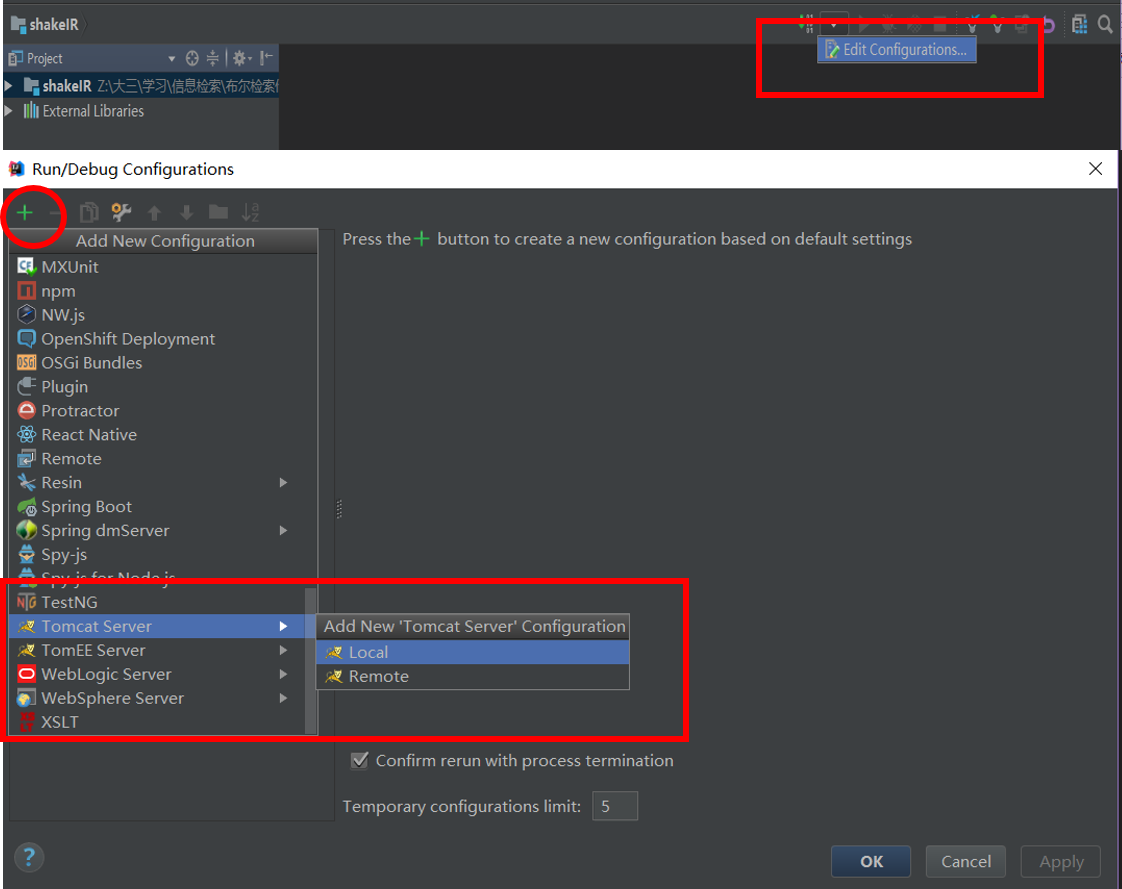


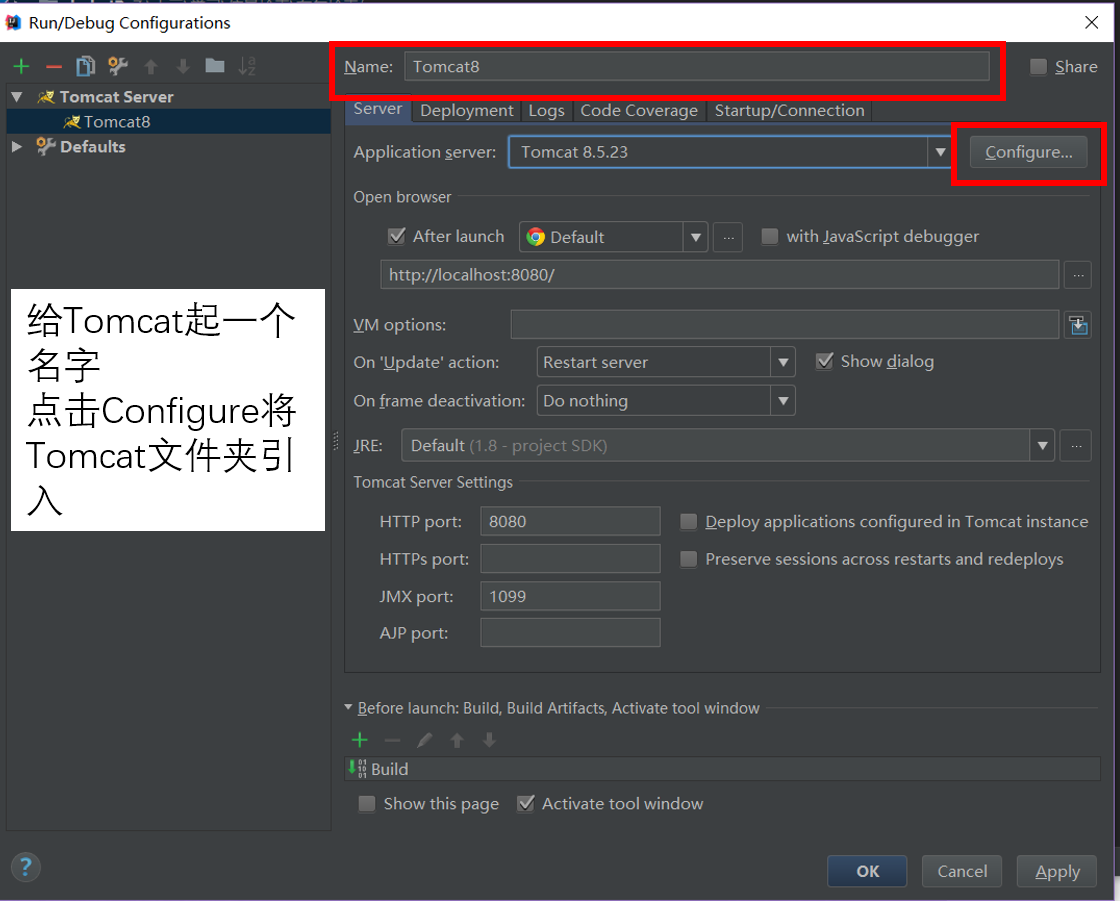
设置Artifacts，如下图所示



配置Tomcat

IDEA右上角点击Edit Configuration，按照图中所示添加一个本地Tomcat

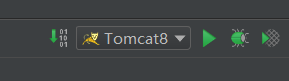




至此，环境配置完毕

三、程序运行

将选项调至Tomcat，点击运行。

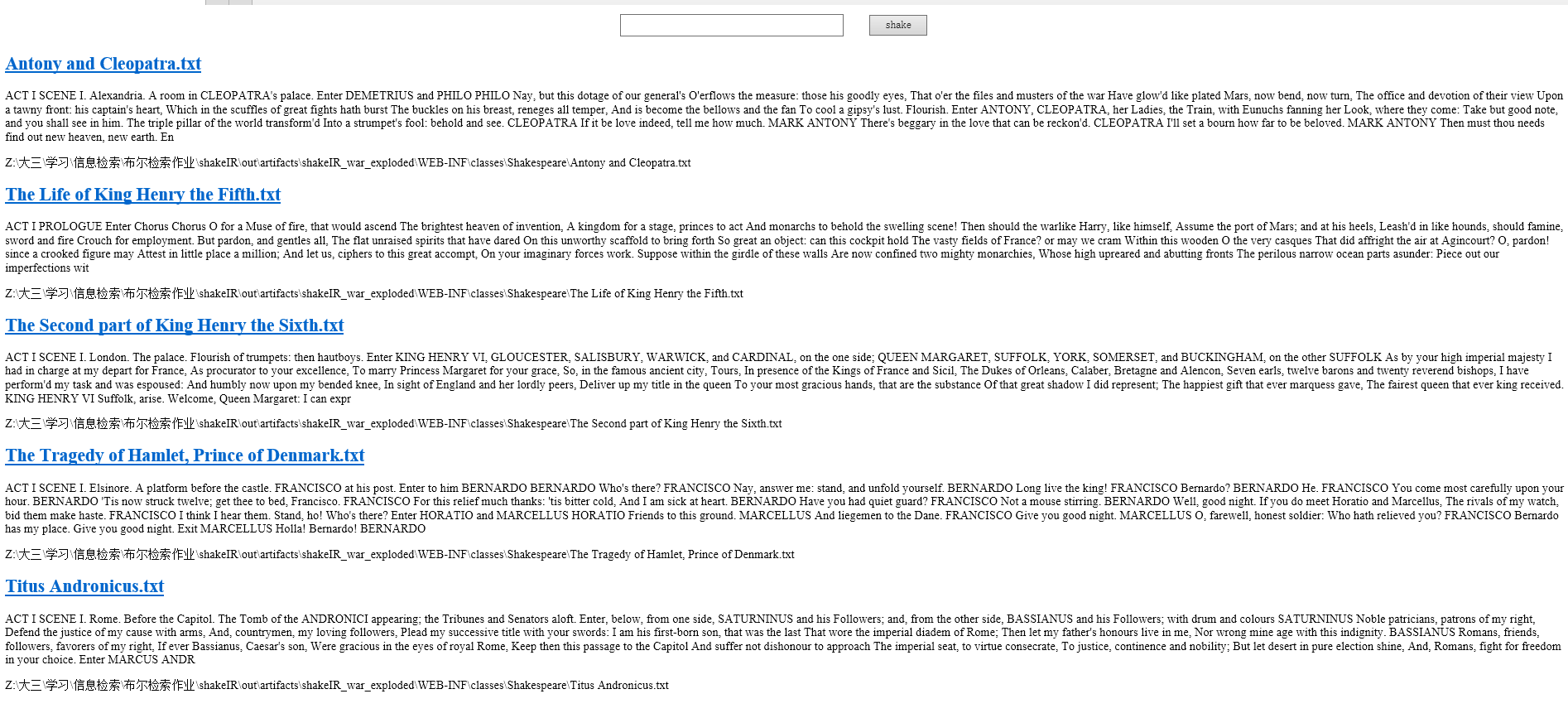


不出意外的话网页会自动跳出，若没有自动跳出，输入<http://localhost:8080>即可打开网页。

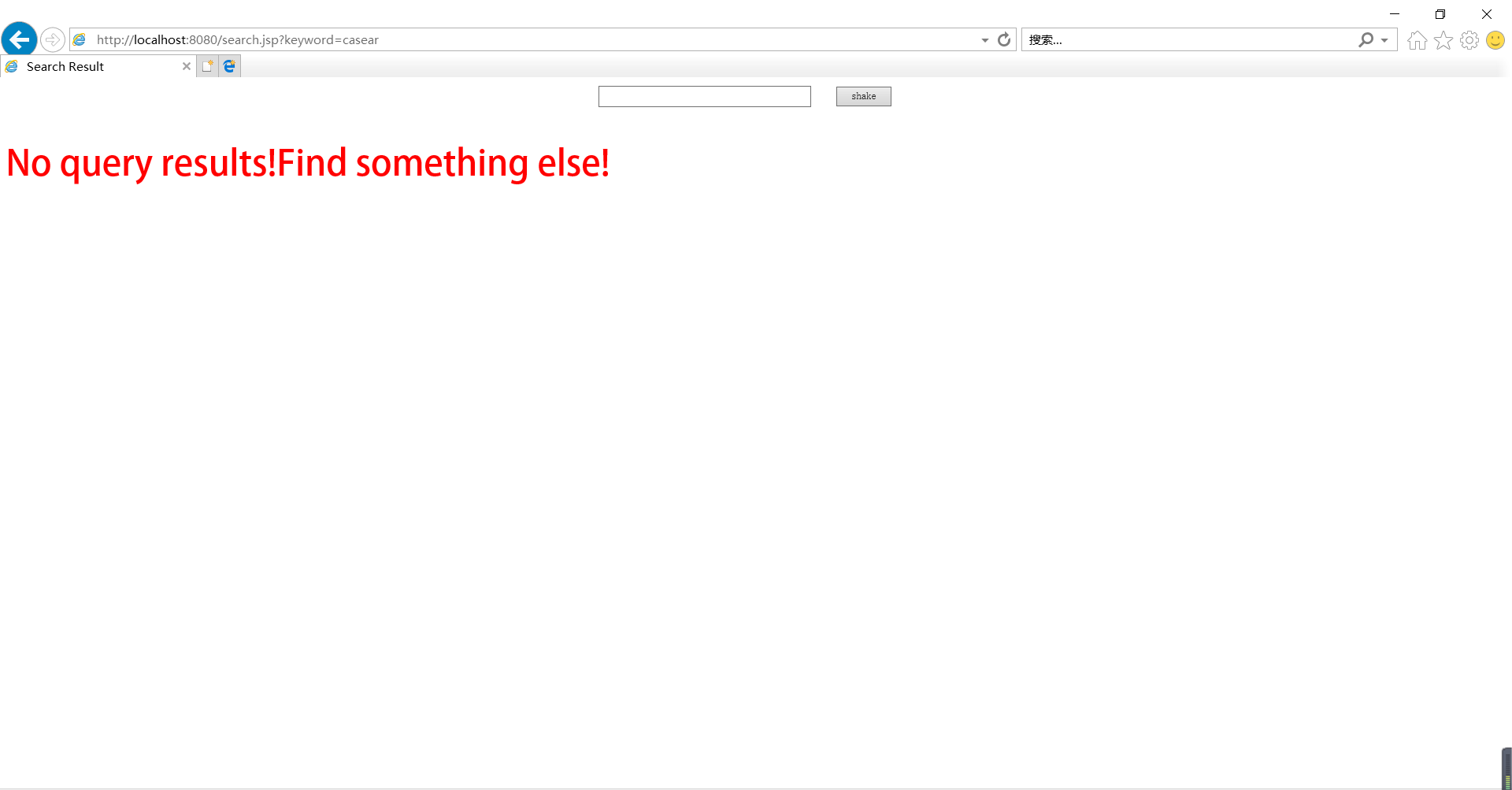
页面如下图所示，在底部文本框输入查询语句，并点击shake即可查询。



结果界面：输入**Brutus AND Caesar AND NOT Calpurnia得到查询结果如下**



在顶部文本框可继续进行查询，若查询不到则显示以下界面：



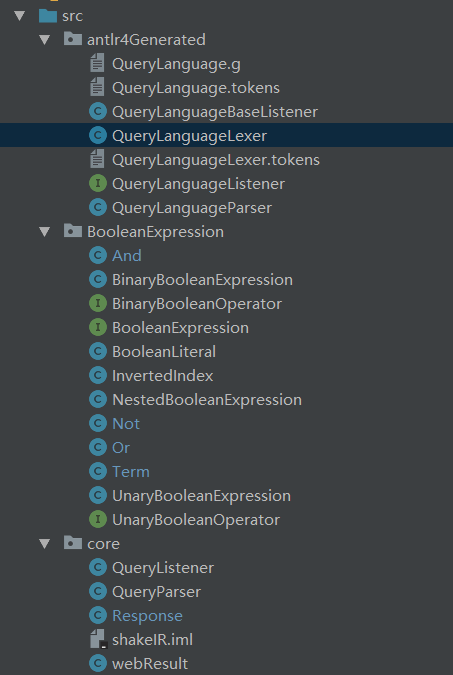
四、程序实现

4.1程序结构

antlr4Generated文件夹中是编写的语法文件和利用antlr语法工具生成的代码。

BooleanExpression 文件夹中是倒排索引和布尔检索所需计算程序

core是核心代码



4.1倒排索引

4.2倒排索引

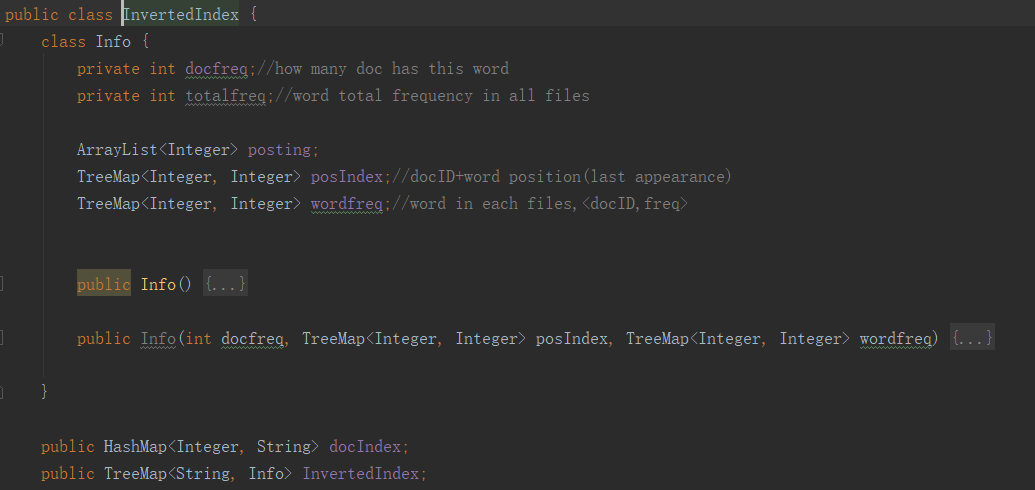
数据结构：InvertedIndex类中有两个变量docIndex和InvertedIndex

docIndex记录键值对<文件编号,文件名字>,使用HashMap

InvertedIndex记录<词语，词语所带的信息Info>，使用TreeMap

Info是一个内部类，包含词语的一些信息。

在本程序中主要用到的变量是倒排表posting，其他变量虽有记录但在布尔查询时并未使用（起初是为了进行优化功能所设）。

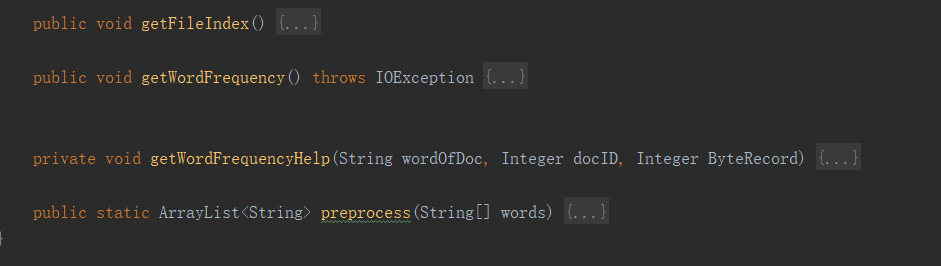


方法：

getFileIndex为文件编号，并储存到数据结构中，且写到了txt文件中。

getWordFrequency获得倒排索引表和词语频率

preprocess利用snowball分词工具对文本进行预处理



4.3查询语句解析

参考资料：

<http://yijun1171.github.io/2015/03/30/ANTLR4学习笔记-语法字典-Grammar-Lexicon/>

<https://theendian.com/blog/antlr-4-lexer-parser-and-listener-with-example-grammar/>

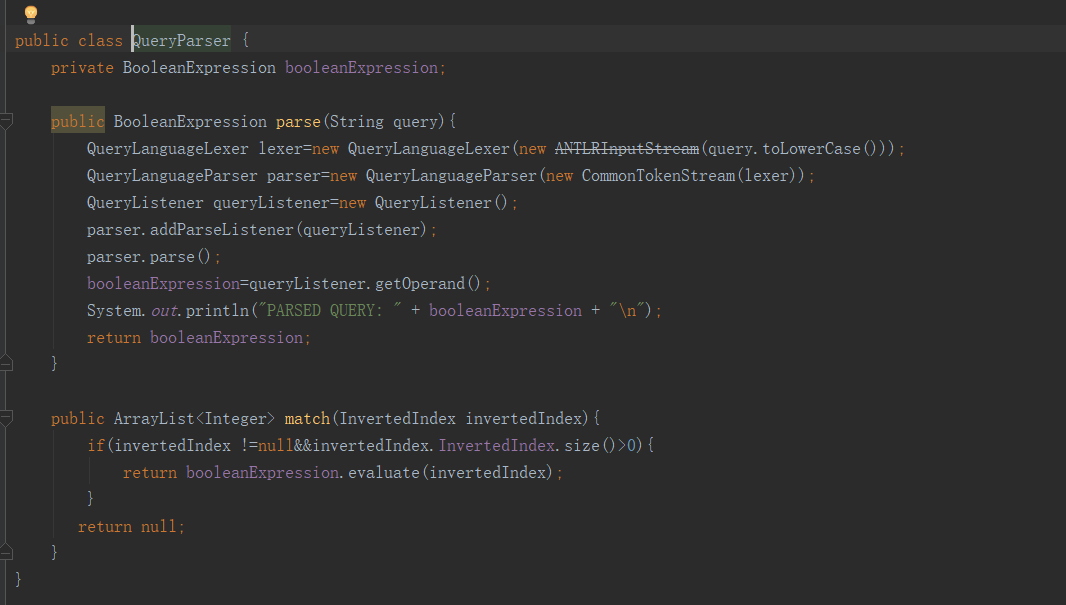
下载ANTLR4包后，编写文法文档.g文件，使用java -jar path/ antlr-4.7-complete.jar /path/xx.g进行编译，即可自动生成该文法的解析器。

文法文档如下图所示：

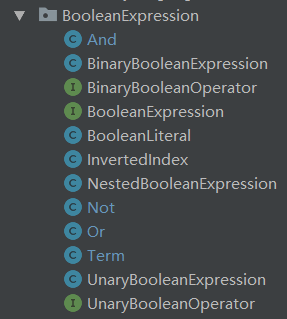


在自动生成的文件中我们只关心QueryLanguageBaseListener，继承这个文件，重写里面的enterxx和exitxx方法，编写QueryListener类。

编写QueryParser类，将query传入此类中，由parse（）函数进行解析，match（）函数进行查询。



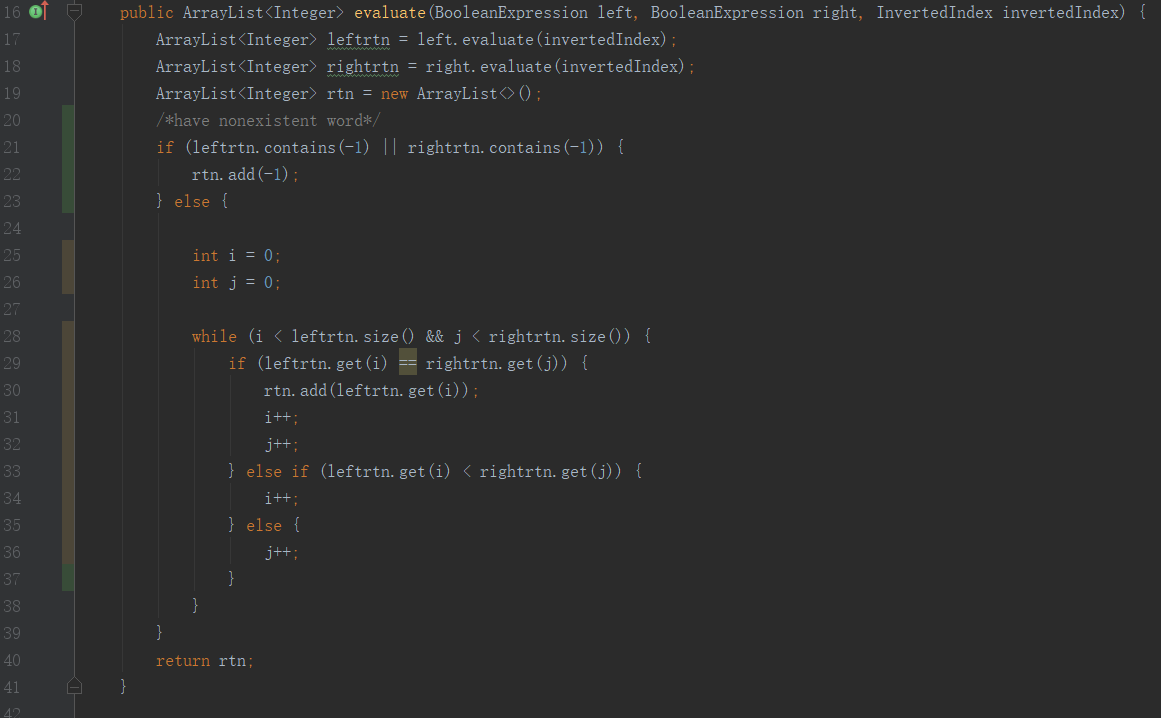
4.4布尔查询



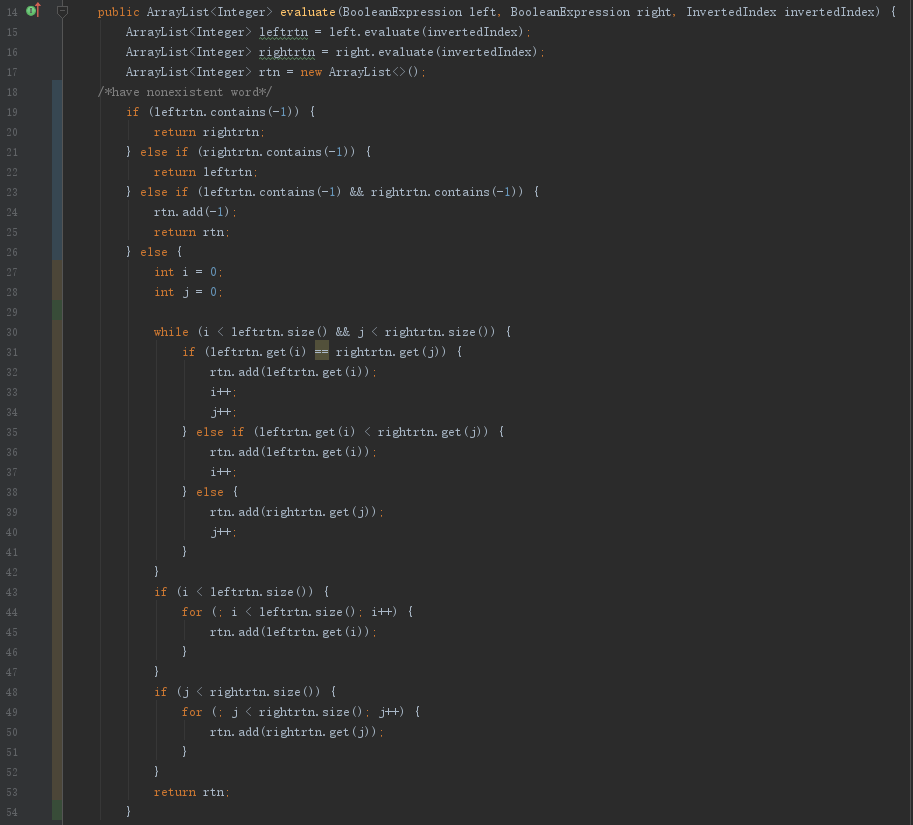
三个接口： BooleanExpression，BinaryBooleanOperator，UnaryBooleanOperator定义布尔表达式，二元运算和一元运算。

其他类实现这三个接口。

And



Or



Not

