开发：

Java 1.8

Maven

Lombok

Hanlp

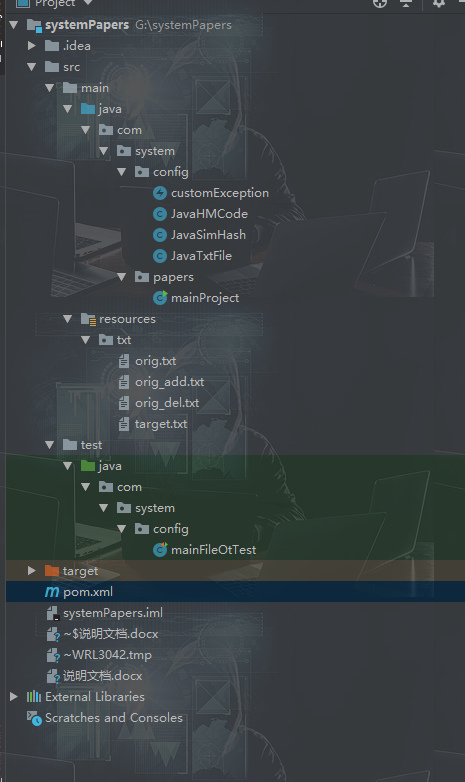
Junit



# 整体流程

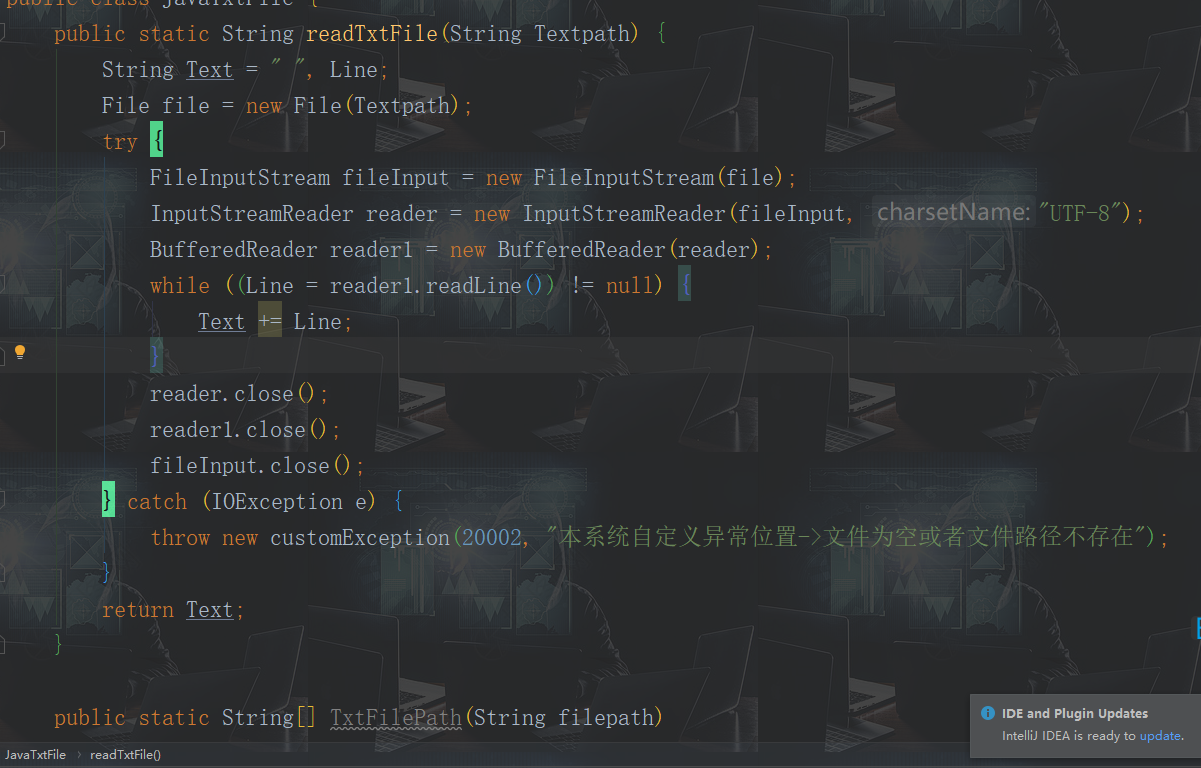
首先读取文本文件并读取文本内容转换成字符串类型-》从中提取关键词-》利用SKA算法计算SimHash的值-》利用海明距离算法求出对应文本内容的相似度-》把结果输出到指定文件

# 项目结构

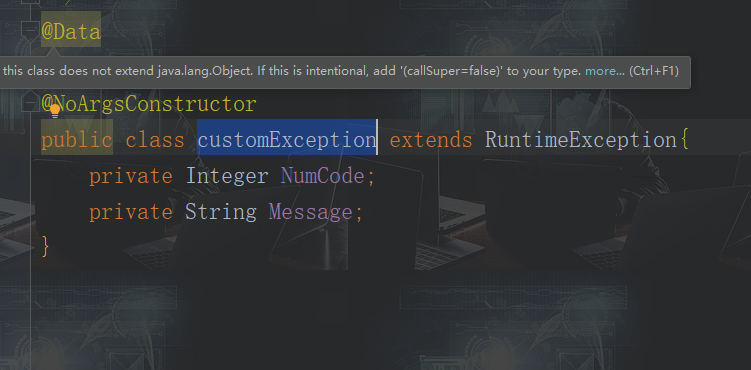


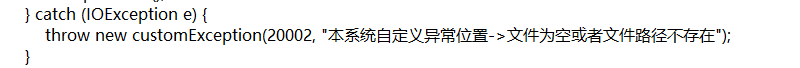
# 项目代码实现：

JavaTxtFile包含读写操作



customException自定义异常处理类





程序段中使用了自定义异常处理类，拥有异常代码和异常信息提醒，让错误位置更醒目

JavaSimHash核心代码模块



类为应用程序提供信息摘要算法的功能，利用SHA算法。接收任意大小的字符串，返回固定长度的哈希值。

getFileHash：传入String，计算出它的hash值，并以字符串形式输出

getJavaSimHash：传入String，计算出它的simHash值，并以字符串形式输出

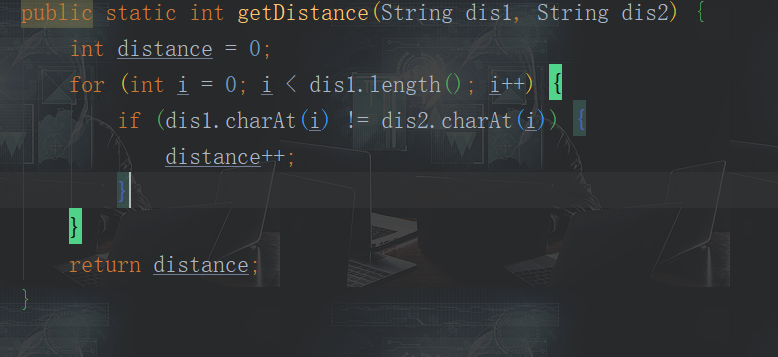


利用hanlp汉语言处理模块进行处理，提出关键词

JavaHMCode海明距离计算类

getDistance：传出两个字符串，计算他们的海明距离

getJavaSimilarity：输入两个 simHash 值，调用 getDistance方法得出海明距离 distance，在由 distance 计算出相似度



Mainproject主程序

读取对应的文件，将文件的内容转化为对应的字符串

由字符串得出对应的 Hash值

由 Hash值求出相似度

把相似度写入最后的结果文件中

# 程序测试类结果

