**SINA 科技热词可视化展示**

**一 设计目标**

1 服务器上设置爬虫定时任务，借助jsoup工具从新浪科技专题爬取当前页面的相关内容，将每个链接对应的内容以文本文件的格式保存到数据库表中。

2 服务器上设置分词定时任务，读取数据库中爬虫记录，使用分词工具ansj对文本内容进行切分，汇总每个词的出现次数，生成<名词，出现次数>键值对。

3 连接MySQL数据库，将生成的<名词，出现次数>键值对插入到数据表当中。

4 按照出现次数多少对数据表中进行排序，选择出top N的热词，生成一个json文件。

5 新建一个web项目，使用Tomcat和Echarts，在本地服务器地址上对json文件的内容进行可视化展示。

**二 工作流图**

**三 设计与编码**

**1 本项目分成两个部分完成。**

第一部分：网页内容的爬取，分词，入库和生成JSON文件。

第二部分：Tomcat的配置，JSON文件中的数据借助echarts功能在本地进行可视化展示。

**2 实现**（具体参见代码）：

testGetNews.java：使用jsoup对新浪科技（<http://tech.sina.com.cn/>）首页进行爬取，设置一定的爬取规则，将爬取的内容以文本的形式进行存储。

testJdbc.java：Java连接MySQL的一些功能实现，比如数据的查询、插入和删除等。

testLoaddb.java：读取本地目录下的文本文件，分词处理，生成键值对，并插入到MySQL数据表中。

Test.java：第一部分的主程序，执行上述三个文件的相关功能，并在本地生成json文件。

test.jsp：echarts可视化的实现。

**四 MySQL**

**1 创建表：**

Create table tech\_sina

(

hotword varchar(50) not null,

counts int

) primary\_key(‘hotword’)

**2 选择热词**

select hotword,counts from tech\_sina order by counts desc limit 0,20

**五 效果展示**

