数据处理过程

一、清除无用数据

- 脏数据

//删除多余的Apply Number 2个
//删除null的专利名 115个
//删除null的分类号 1个
//删除null的申请人 7个
//删除null的发明人 1个
//删除null的公开日期和申请日期 2个

- 处理地理数据
  首先找出所有不同地理数据，根据数据的情况进行处理

//将市匹配到对应的省份

沈阳;89	辽宁
海南;66	
武汉;83	湖北
广州;81	广东
大连;91	辽宁
香港;HK	
台湾;71	
地质矿产部宜昌;42	湖北
西安;87	陕西
东北财经大学;21	辽宁
哈尔滨;93	黑龙江
台湾;TW	台湾
青岛;95	山东
深圳市;44	
杭州市;86	浙江
佛山市;44	广东
烟台市;37	山东
苏州;32	江苏
汕头市;44	
中国;CN	29284
315324浙江省慈溪市周巷镇北片工业区宁波凯波集团有限公司	
江艎??;32	1
????海;31	2
????江;33	1
乌鲁木齐;65	新疆
徐州;32	江苏

```
//影响74行
UPDATE expressionmanager_patent SET province_code = "辽宁;21" WHERE province_code = "沈阳;89";
//影响27行
UPDATE expressionmanager_patent SET province_code = "湖北;42" WHERE province_code = "武汉;83";
//影响97行
UPDATE expressionmanager_patent SET province_code = "广东;44" WHERE province_code = "广州;81";
//影响54行
UPDATE expressionmanager_patent SET province_code = "辽宁;21" WHERE province_code = "大连;91";
//影响1行
UPDATE expressionmanager_patent SET province_code = "湖北;42" WHERE province_code = "地质矿产部宜昌;42";
//影响25行
UPDATE expressionmanager_patent SET province_code = "陕西;61" WHERE province_code = "西安;87";
//影响2行
UPDATE expressionmanager_patent SET province_code = "辽宁;21" WHERE province_code = "东北财经大学;21";
//影响21行
UPDATE expressionmanager_patent SET province_code = "黑龙江;23" WHERE province_code = "哈尔滨;93";
//影响17行
UPDATE expressionmanager_patent SET province_code = "山东;37" WHERE province_code = "青岛;95";
//影响16行
UPDATE expressionmanager_patent SET province_code = "浙江;33" WHERE province_code = "杭州市;86";
//影响6行
UPDATE expressionmanager_patent SET province_code = "广东;44" WHERE province_code = "佛山市;44";
//影响2行
UPDATE expressionmanager_patent SET province_code = "山东;37" WHERE province_code = "烟台市;37";
//影响11行
UPDATE expressionmanager_patent SET province_code = "江苏;32" WHERE province_code = "苏州;32";
//影响1行
UPDATE expressionmanager_patent SET province_code = "浙江;33" WHERE province_code = "315324浙江省慈溪市周巷镇北片工业
//影响1行
UPDATE expressionmanager_patent SET province_code = "江苏;32" WHERE province_code = "江解??;32";
//影响2行
UPDATE expressionmanager_patent SET province_code = "上海;31" WHERE province_code = "????海;31";
//影响1行
UPDATE expressionmanager_patent SET province_code = "浙江;33" WHERE province_code = "????江;33";
//影响1行
```

//处理政治问题
台湾;TW 转为 台湾;71

//处理国家问题

阿联酋;AE	
阿拉伯联合酋长国;AE	阿联酋
也门;YE	
沙特阿拉伯;SA	
格鲁吉亚;GE	
亚美尼亚;AM	
乌兹别克斯坦;UZ	
科威特;KW	
新加坡;SG	
约旦;JO	
马来西亚;MY	
阿富汗;AF	
文莱达鲁萨兰国;BN	
吉尔吉斯斯坦;KG	
阿塞拜疆;AZ	
伊拉克;IQ	
阿曼;OM	
叙利亚;SY	
尼泊尔;NP	
缅甸;MM	
苏联;SU	332
俄国;RU	俄罗斯
俄罗斯;RU	
俄罗斯联邦;RU	俄罗斯
荷兰;NL	
何兰;NL	何兰
南斯拉夫;YU	26

K;H	75

ps:kh是柬埔寨

二、针对每个题目生成对应的静态数据

题目一：专利申请趋势

统计以月份为粒度的每个月专利申请的数量，同时增加需求，统计每个月数量中，不同种类的专利数量

题目二：专利公开趋势

统计以月份为粒度的每个月专利公开的数量，同时增加需求，统计每个月数量中，不同种类的专利数量

题目三：3、4、5题

两步：
一：首先用基本的规则匹配，就是包含有“公司”，“院”，“大学”，“厂”，“所”，“会社”，“中心”，“企业”的类
二：用了NLP中命名实体识别的技术，在第一部的基础上去除掉一些人名

3题：
专利类别分析

CN89103333.5
CN200310102732.3
CN85300127

1= 发明专利申请
2= 实用新型专利申请
3= 外观设计专利申请
8= 进入中国国家阶段的PCT发明专利申请
9= 进入中国国家阶段的PCT实用新型专利申请。

长度	数量	举例
12	800213	CN89103333.5 CN89100995.7
16	12594127	CN200310114879.4 CN200310102732.3
10	31477	CN86100777 CN85300127

根据专利申请号的长度，匹配对应的专利类别，并统计数量

4题：
根据公司统计不同大组小组的数量

5题：
根据公司，统计invente_man这个列下人的数量

雷达图：
统计不同公司的不同类别的数量，查看这个公司在哪些领域比较擅长

关联图

每个专利有classify_node 与 main_classify_node，如果main_classify_node与classify_node的section 不同，这说明这两个领域是有关联的。就可以画出关联图

地理位置图

国家：专利输出到中国的情况
身份：不同省份专利情况