PaperPass专业版检测报告
简明打印版

比对结果（相似度）：
总　体：23 % *（总体相似度是指本地库、互联网的综合比对结果）*
本地库：17 % *（本地库相似度是指论文与学术期刊、学位论文、会议论文数据库的比对结果）*
期刊库：13 % *（期刊库相似度是指论文与学术期刊库的比对结果）*
学位库：14 % *（学位库相似度是指论文与学位论文库的比对结果）*
会议库： 3 % *（会议库相似度是指论文与会议论文库的比对结果）*
互联网：9 % *（互联网相似度是指论文与互联网资源的比对结果）*

编　号：592482CFD8A531BH2
版　本：专业版
标　题：京津冀安全大数据平台
作　者：ltp
长　度：16778 字符(不计空格)
句子数：514句
时　间：2017-5-24 2:43:27
比对库：学术期刊、学位论文（硕博库）、会议论文、互联网资源
查真伪：http://www.paperpass.com/check

句子相似度分布图：

本地库相似资源列表（学术期刊、学位论文、会议论文）：
1. **相似度：6 % 篇名:《基于Mediawiki的学科信息门户建设》**
   来源：学术期刊 《现代图书情报技术》 2007年12期 作者: 杨雁 刘志
2. **相似度：5 % 篇名:《基于Mediawiki的学科信息门户建设》**
   来源：学位论文 湘潭大学 2009 作者: 杨雁
3. **相似度：3 % 篇名:《基于MediaWiki的故障处理知识库研究与实现》**
   来源：学位论文 天津大学 2014 作者: 田宇航
4. **相似度：3 % 篇名:《维基在中学德育中的应用探索——基于上海市洋泾中学的维基平台建设实践研究》**
   来源：学位论文 华东师范大学 2009 作者: 乔骁杰
5. **相似度：2 % 篇名:《基于Wiki的远程协作学习平台的研究与实践》**
   来源：学术期刊 《天津职业院校联合学报》 2012年12期 作者: 沈伊海

6. 相似度：2％ 篇名:《WIKI技术在教学中的应用研究》
   来源：学术期刊 《吉林师范大学学报（自然科学版）》 2007年2期 作者: 李艳
7. 相似度：2％ 篇名:《基于Spring MVC框架和TRBAC访问控制模型的工作流系统的设计》
   来源：学位论文 合肥工业大学 2014 作者: 王旺
8. 相似度：1％ 篇名:《中国航空网络的复杂性研究》
   来源：学位论文 南京信息工程大学 2010 作者: 王俊超
9. 相似度：1％ 篇名:《节点移动性对MANETs网络拓扑特征的影响》
   来源：学术期刊 《计算机工程与应用》 2014年9期 作者: 冯慧芳 王梦茹
10. 相似度：1％ 篇名:《基于维基的政府知识管理应用——以杭州海关为例》
    来源：学位论文 浙江工业大学 2009 作者: 蒋晨
11. 相似度：1％ 篇名:《唐山市区公交系统静态网络性质的研究》
    来源：学术期刊 《唐山师范学院学报》 2010年5期 作者: 李先铭 崔乃忠 孙立萍 杨俊锋 刘艳春
12. 相似度：1％ 篇名:《微博用户关系网络的结构研究与聚类分析》
    来源：学术期刊 《复杂系统与复杂性科学》 2013年2期 作者: 杨凯 张宁
13. 相似度：1％ 篇名:《基于wiki的主题学习的应用研究》
    来源：学位论文 上海师范大学 2008 作者: 林丽
14. 相似度：1％ 篇名:《最近邻演化网络模型》
    来源：学术期刊 《中国科技信息》 2015年2期 作者: 徐玉章 朱磊
15. 相似度：1％ 篇名:《基于Scrapy框架的新闻实时抓取及处理系统的设计与实现》
    来源：学位论文 南开大学 2012 作者: 林伟坚
16. 相似度：1％ 篇名:《基于网络中心性的城市轨道交通应急救援站选址研究》
    来源：学位论文 北京交通大学 2014 作者: 李刚
17. 相似度：1％ 篇名:《构造无线传感器网络的小世界效应研究》
    来源：学术期刊 《计算机工程与应用》 2010年2期 作者: 唐鹭 洪月华 伍华健
18. 相似度：1％ 篇名:《基于MVC模式Web应用框架的研究开发和应用》
    来源：学位论文 南京工业大学 2009 作者: 王向中
19. 相似度：1％ 篇名:《微博用户关系网络演化特性的初步研究》
    来源：学位论文 北京化工大学 2013 作者: 高民胜

互联网相似资源列表：

1. 相似度：5％ 标题：《Scrapy研究探索2 - ¯ -_lishk_- ¯ -...》
   http://blog.csdn.net/lishk314/article/details/44243581
2. 相似度：3％ 标题：《初试Scrapy（五）—通过下载保存美女图片来学习下Spider中间...》
   http://blog.csdn.net/xj178926426/article/details/53915757
3. 相似度：2％ 标题：《“泛社交”时代，典型社交应用差异化发展-中共中央网络安全和信息化领导...》
   http://www.cac.gov.cn/2017-01/22/c_1120362548.htm
4. 相似度：1％ 标题：《推荐12个最好的JavaScript图形绘制库_HTML5中国_...》
   http://www.html5cn.org/article-7332-1.html

全文简明报告：

1.1 项目的背景

京津冀地区化工厂、加油站、交通事故对民众日常生活和出行的安全均具有一定威胁。 {59%：京津冀地区作为中国的重要城市群，其重要性不言而喻。} 北京作为中国的政治中心、首都，在中国的地位十分重要，其常住人口2172.9万人； 天津是中国的直辖市，毗邻北京，常住人口1562.12万； 河北省环绕天津北京，全省总人口7185万，总面积18.88万平方千米，河北保定的雄县、安新、容城三地称为雄安新区， {45%：位于京津冀腹地，将作为中国的经济中心，发展潜力极大。} {64%：三地共称京津冀，是我国重要的经济圈之一。} {54%：从安全与发展的关系看，安全稳定是经济社会发展的基础。} 京津冀的发展需要安全稳定的区域环境，特别是需要城市安全作支撑[1]。 {44%：民众作为城市的支撑者，民众的安全也是不可忽视的重要方面。} 而威胁民众日常生活以及出行安全的很大一部分外在因素在于城市具有安全隐患的地区。 {42%：包括化工厂、加油站以及城市的事故多发路段。}

化工厂多处于港口、市郊等地点，距离人口密集地方较远，但是其存储的化工原料一旦泄露或者发生爆炸，影响范围广，可控性小，危害极大。 {56%：2015年8月12日晚，天津滨海新区天津港的瑞海公司危险品仓库发生火灾爆炸事故，} {61%：造成165人遇难，798人受伤，造成直接经济损失60多亿元。} 化工厂因其特殊性，监管较严格，不易发生事故。 然而一旦发生事故，对民众的影响将会非常大。 加油站分布在道路两侧，是人们出行路途中经常会遇见的危险源。 同时因为民众对于汽油的刚需，加油站距离民居距离较近，一旦加油站发生火灾或者爆炸，对出行的民众和附近居民都会造成不小的安全威胁。 {100%：2001年7月23日下午3时，郑州标准石化有限公司商城路加油站发生爆炸，导致4人死亡，} {97%：1人重伤，10人轻伤和轻微伤，爆炸造成直接财产损失近20万元。} 在人们的日常生活中，出行是必不可少的。 很大一部分人会选择自驾出行，但是自驾出行过程中偶尔会出现车祸。 其原因一部分是因为驾驶员未遵守交通规则，还有一部分是因为道路崎岖或者道路设计不合理。{98%：2004年1月至11月，全国共发生道路交通事故470019起，造成96870人死亡、435740人受伤。} 交通事故在人们日常生活中安全受到威胁的主要方面之一。

从对民众造成威胁发生的概率看，交通事故无疑是概率最高的，加油站事故次之，化工厂事故再次。 然而从每次事故造成的损失来看，化工厂事故造成的损失最大，齐次是加油站，再次是交通事故。 {42%：综合来看，对民众日常生活影响最大的是交通事故。} 齐次是加油站事故，其发生概率小于交通事故，但是其一旦发生经济损失要大于交通事故。 化工厂事故的发生概率远小于前两者，但是其造成的损失堪称巨额，也不能忽略。

目前还没有针对京津冀化工厂、加油站、事故多发路段的专门的安全平台。 高德地图对于京津冀的地理信息（布局、路网、地块）展示的很详细，但是并没有对于京津冀地区具有安全威胁的地区进行有针对性的标注。 微博对于京津冀的各种信息的传播非常有效迅速，但是并不能让我们更加系统地了解民众对于京津冀地区安全的言论走向以及关注点。

1.2 项目的意义

本项目构建了京津冀安全平台。 聚焦影响京津冀民众安全的化工厂、加油站以及事故多发路段，构建了京津冀安全地图，并展示在平台上。 让人们能够对于威胁京津冀民众安全的地点的分布有个直观地了解。 同时我们还用微博上人们对于京津冀的相关评论数据构建了地域-事件-人-行为的网络， 并分析了其基本属性，结果展示在平台上，供人们查看，使人们对于京津冀相关的网络话题有一个直观地了解。

1.3 章节结构

本章我们介绍了本项目的背景及意义，总体介绍了项目的来源和平台能够为人们提供的帮助。 第二章我们将

介绍相关研究，介绍我们的平台的数据来源以及使用的工具。 第三章我们将介绍数据的获取，包括数据的需求分析，地理信息的获取和社交网络信息的获取。 第四章我们将对获取的数据进行进一步处理及分析，包括安全地图的构建和网络的构建与分析。 第五章我们想对平台进行全面的描述，包括平台的需求分析、平台的设计、平台的实现以及效果展示。 最后一章为小结与展望，我们想介绍该平台未来的拓展以及致谢。

第二章 相关研究

2.1 领域现状

京津冀安全地图的构建需要得到京津冀的地理信息，所以我们需要在网络地图中获取这些信息。 {70%：备选的地图有高德地图、百度地图和腾讯地图。} {41%：道路信息在网络地图中很关键，所以道路改变在网络地图中的更新速度很重要。} 通过简单的对比，我们发现高德地图的道路数据更新是最快的； 同时poi数据来源方面，高德地图有自己完整的采集更新团队，而百度地图的数据来源是长地万方公司、腾讯地图是四维图新公司。综合以上两点，我觉得选择高德地图作为京津冀地理信息数据来源。

高德地图对于京津冀地理信息的阐释很详尽，但是并非是针对京津冀地区的，缺少对京津冀地区地理信息足够的针对性。 同时网络地图没有民众的舆论信息。

{95%：微信朋友圈、qq空间、微博同属于综合性的社交应用，但在社交关系的紧密度、用户属性及地域特征上存在较大差异。} {100%：微信朋友圈是相对封闭的个人社区，分享的信息偏向朋友之间的交互，微博是基于社交关系来进行信息传播的公开平台，} {100%：用户关注的内容越来越倾向于基于兴趣的垂直细分领域，qq空间则介于两者之间；} {100%：从用户特征来看，微信朋友圈用户渗透率高，除低龄（6-9岁）、低学历人群（小学及以下学历）外，} {95%：个群体网民对微信朋友圈的使用率无显著差异；} {95%：无线城市网民、10-19岁网民对qq空间的使用率明显较高，产品用户下沉效果明显；} {100%：微博用户特征更为明显，一线城市网民、女性网民、20-29岁网民、本科及以上学历网民、} {65%：城镇网民对微博的使用率明显高于其他群体[中国互联网络发展状况统计报告2017]。}

新浪微博是一线城市网民舆论的主要平台，京津冀安全方面的新闻传播迅速，但是对于地理位置的定位不够明确， 且不能够让人们对于京津冀相关话题的分布有更加准确的了解。

2.2 技术现状

2.2.1 Mediawiki平台与SpringMVC

Web2.0技术的发展使得网络空间更具有互动性与协作性，加强的个人的互动积极性与参与感[1]而在web2.0技术中， 维基则被认为是典型的代表[2]，并被描述为协作项目的理想在线平台[3]。 {69%：维基的内容有广大的民众自由贡献，对知识的协同创作和知识的共享有很大帮助，同时在知识的组织和传播利用中也起到了重要作用。}

{100%：Mediawiki是全球最著名的，运行于PHP+MySQL环境的wiki知识库引擎。} {98%：从从2002年2月25日被作为维基百科全书的系统软件，并有大量其他应用实例，如目前国内的天下维客、维库等站点都采用这套系统。} {94%：目前Mediawiki的开发得到维基媒体基金会的支持，为Mediawiki的用户提供了良好的系统开发保障和技术支持[4]。}

Mediawiki的主要特点有一下四点：

{100%： （1）经受过重量级应用的考验，功能丰富而且易于安装。} {100%： 全世界最大的wiki项目维基百科全书是使用Mediawiki的成功范例，数据量、访问量都超级庞大。} {100%： Mediawiki的功能非常丰富而且拥有大量扩展插件可供选择安装，支持多语言版本，充分满足知识站点的需要，中文支持较好。} {100%： 运行环境要求很低，安装过程简洁，即使新手也可以迅速建立自己的站点。} {98%： 同时，Mediawiki作为目前应用最广的wiki程序，数以万计的网站在使用它，很容易找到范例站点和相关的技术支持。}

{100%： （2）持续开发，程序特性功能不断完善，保证未来的支持。} {99%： Mediawiki虽然作为开源软件，但由于它是维基媒体基金会支持的开源项目，因此在功能、性能、安全方面将不断优化，版本也将不断升级更新。}

{100%： （3）满足wiki的重要特征，良好的个性化设置。} {98%： 保留网页每一次改动的版本，即使参与者将整个页面删掉，管理者也会很方便地从记录中恢复最正确的页面版本，这使得开放性编辑成为可能。} {97%： 能够自动产生链接，在进行文本编辑时，只要编辑双中括号中的内容（如} “ [[条目]] ” ），将自动产生链接。 {100%： 允许使用模板，方便对相同内容的重复使用、更新。} {96%： 支持分类，并根据分类在不同的文章之间自动产生关联。} {100%： 允许每个用户自行选择系统外观。}

{94%： （4）拥有页面保护、IP禁止技术，管理员可以对一些主要页面（如首页、已经相当完善的词条、标准模板等）用保护技术将内容锁定，} {100%： 这样，其他人员就不能再对这些内容作编辑修改。}

MVC框架也可以作为平台的架构工具进行使用。 典型的代表是SpringMVC框架。 {63%： MVC模式（Model View Controller，简称MVC）是软件工程中的一种软件架构模式，把软件系统分为模型（Model）、视图（View）和控制器（Controller）三个部分[5]。} Model对象包含数据； {93%： View对象负责显示有模型包含的数据，用于与用户交互；} {88%： Controller对象是介于Model与View之间的桥梁，它可以分发和处理用户的请求，选择适当的视图用于显示模型包含的数据返回给用户。}

{68%： SpringMVC框架提供了构建 Web应用程序的全功能 MVC模块[6]，是一种高度可配置的 MVC框架，} {85%： 可以定制本地化和主图解析，并提供多种视图技术，实现了控制器、模型对象、分派器以及处理程序对象的多角色分离[7]，} 这种分离让它们更容易进行定制。

从平台的扩展性来看，SpringMVC框架数据和视图分离，添加页面方便，但是需要由开发者进行添加； {43%： 而Mediawiki自带编辑语言，可以有用户进行页面编辑，无疑降低了编辑的门槛。} 从开发难度来看，Mediawiki由php与MySQL构建，有完整的教程，其安装简单，配置过程仅仅需要回答几个配置问题即完成，平台架构起来相对简单； 而SpringMVC需要对javaweb开发以及SpringMVC框架本身进行系统的学习，学习周期较长。{40%： 而从语言特点来看，Mediawiki使用PHP语言，支持热部署，但是连接数据库的速度慢；} 而SpringMVC热部署能力弱，但是其jdbc数据库连接模块连接数据库的速度远快于php。

本项目的京津冀安全地图的构建需要用到的数据为 json格式，可以直接由 JavaScript从后台读取， 不需要连接数据库，而 php的热部署功能也能够保证平台的运行不会因为修改而停止。 故本平台选择Mediawiki作为开发平台。

2.2.2 d3.js

{41%： d3.js（data drive document）是一个基于数据的文档操作的JavaScript库[8]，它通过数据加载、数据绑定、分析元素转换和大量元素操作来进行数据的可视化。} 与excel不同，它给用户提供了自定义的映射规则。 根据用户的需求不同，用户可以自己决定图形的映射规则，如颜色、大小等。 D3不支持旧版浏览器，这样可以使得其代码更加干净。 {40%： D3在处理SVG上表现很好，这是万维网（world wide web）规范的指定网络矢量图形标准[9]。} {71%： SVG严格遵守XML语法并使用文本格式的描述语言进行图像内容描述。} {51%： 它是一个不受分辨率影响的矢量图形格式[10]。}

Chart.js是和d3.js同类型的可视化JavaScript库，其大小小于d3.js。 {92%： 且建立在HTML5 Canvas的基础上，目前它支持6中图表类型（折线图，条形图，雷达图，饼图，柱状图和极地区域区）。} {91%： 而且，它是一个独立的包，不依赖第三方的JavaScript库。}

{100%： Highcharts JS 是一个制作图表的纯 Javascript 类库，主要特性如下：} 兼容性： 兼容当今所有的浏览器，包括 iPhone、IE 和火狐等等； 对个人用户完全免 费； 纯JS，无BS； 支持大部分的图表类型： {100%： 直线图，曲线图、区域图、区域曲线图、柱状图、饼装图、散布图；} 跨语言： 不管是 PHP、Asp.net 还是 Java 都可以使用。

以上三者均为纯JavaScript图形库，不依赖后台语言。 {57%： 而Chart.js仅支持六种（折线图，条形图，雷达图，饼图，柱状图和极地区域区）图表类型，自由性要弱于另外两种JavaScript库；} Highcharts.js的主要功能为画图，对数据的控制力度要弱于chart.js以及d3.js； 另外，d3.js虽然对数据的控制更强一些、且对于用户的自由性更高，但是其绘图是基于SVG，SVG对于大量数据的处理速度要慢与canvas。

本项目是平台性项目，该部分库主要用于实现京津冀安全地图，对加载速度要求不高。 综合以上对比，d3.js无疑是最好的选择。

2.2.3 scrapy

scrapy是目前比较主流的开源爬虫框架[11]。 本框架用python编写，并基于twist框架---基于事件驱动的网络引擎框架。 Scrapy的各个组件及工作流程如下所示：

1. Scrapy Engine

{100%： 引擎负责控制数据流在系统中所有组件中流动，并在相应动作发生时触发事件。}

2. 调度器(Scheduler)

{100%： 调度器从引擎接受request并将他们入队，以便之后引擎请求他们时提供给引擎。}

3. 下载器(Downloader)

{100%： 下载器负责获取页面数据并提供给引擎，而后提供给spider。}

4. Spiders

{100%： Spider是Scrapy用户编写用于分析response并提取item(即获取到的item)或额外跟进的URL的类。} 每

个spider负责处理一个特定(或一些)网站。 更多内容请看 Spiders 。

5. Item Pipeline

Item Pipeline负责处理被spider提取出来的item。 {100％： 典型的处理有清理、 验证及持久化(例如存取到数据库中)。}

6. 下载器中间件(Downloader middlewares)

{100％： 下载器中间件是在引擎及下载器之间的特定钩子(specific hook)，处理Downloader传递给引擎的response。} {100％： 其提供了一个简便的机制，通过插入自定义代码来扩展Scrapy功能。}

7. Spider中间件(Spider middlewares)

Spider中间件是在引擎及Spider之间的特定钩子(specific hook)，处理spider的输入(response)和输出(items及requests)。 {100％： 其提供了一个简便的机制，通过插入自定义代码来扩展Scrapy功能。}

其工作流程如下：

{100％： 1. 引擎打开一个网站(open a domain)，找到处理该网站的Spider并向该spider请求第一个要爬取的URL(s)。}

{100％： 2. 引擎从Spider中获取到第一个要爬取的URL并在调度器(Scheduler)以Request调度。}

{100％： 3. 引擎向调度器请求下一个要爬取的URL。}

{100％： 4. 调度器返回下一个要爬取的URL给引擎，引擎将URL通过下载中间件(请求(request)方向)转发给下载器(Downloader)。}

{100％： 5. 一旦页面下载完毕，下载器生成一个该页面的Response，并将其通过下载中间件(返回(response)方向)发送给引擎。}

{100％： 6. 引擎从下载器中接收到Response并通过Spider中间件(输入方向)发送给Spider处理。}

7. {100％： Spider处理Response并返回爬取到的Item及(跟进的)新的Request给引擎。}

8. 引擎将(Spider返回的)爬取到的Item给Item Pipeline，将(Spider返回的)Request给调度器。

{100％： 9. (从第二步)重复直到调度器中没有更多地request，引擎关闭该网站。}

从其工作流程可以看出，该框架流程化的爬取过程可以让我们清晰的完成爬虫各个部分的工作，同时其允许用户自定义中间件让我们实现很多自定义的处理。 非常适合入门用户大量爬取数据。

2.3 本章小结

本章中，我们分析了项目要使用的几个工具的优缺点以及为什么要使用。 包括平台框架Mediawiki，可视化工具库d3.js和爬虫框架scrapy。 其中Mediawiki框架因其为php开发对数据库的读取速度较慢，但是安装简单，扩展性好。 我们的平台中京津冀安全地图不会从数据库读取数据，数据格式全部为json格式存在文件系统中。 网络的分析也不会再平台中进行，平台用来展示结果。 故其数据库读取慢的缺点可以忽略。 D3.js对于数据的操作以及可视化的效果都非常优秀，同时入门简单，样例较多。 {47%：唯一的问题是，其v3版本基于svg进行可视化，对于大量的数据显示能力不如canvas。} 我们的项目京津冀地图要用到d3.js，其数据量不大。 网络的计算并不会在平台上进行，故d3.js的能力约束不会对平台造成负担。 Scrapy爬虫框架需要进行一段时间的学习，其结构化的爬取流程以及可以定制的中间件非常方便，使得我们能够灵活的处理目标网站的反扒机制。

第三章 数据获取

3.1 数据获取整体设计

京津冀安全大数据平台所需要的数据分为两个部分， 京津冀地理信息（即京津冀物理空间信息）以及京津冀地区在社交网络上的相关数据（即京津冀网络空间信息）， {54%：通过 scrapy框架进行爬取，整体爬取流程如图3.1所示。}

京津冀地理信息的数据用于构建京津冀安全地图，来自高德地图。 高德地图拥有京津冀详细的地理信息数据，且准确性可以保证。 我们需要用于构建京津冀安全地图的数据包括京津冀边界、京津冀的 路网数据、京津冀化工厂、加油站、事故多发路段坐标。 爬取后需要对数据进行清洗，首先去除重复信息，将道路名重复且坐标重复的信息删除； 然后需要对数据进行整理，去除无用数据，有些数据只包含道路名，但是并没有坐标； {50%：然后统一格式，将数据格式转换为GeoJson格式，便于进行地图的绘制。}

{47%：京津冀地区在社交网络上的相关数据用于构建京津冀舆论网络。} 来自新浪微博。 {48%：新浪微博拥有大量的用户，话题传播非常迅速。} 我们爬取的数据包括新浪微博关于京津冀以及京津冀安全方面的数据和主题词下用户的评论数据和用户的信息。 爬取后需要对数据进行清洗，首先合并相同主题微博，然后统一用户，将重复id的用户合并。 {46%：然后去除无用的数据，关键词缺失或者用户信息缺失的数据进行删除。} 完成后将数据存储在本地。

{62%：图3.1 数据爬取顶层数据流图}

3.2 地理信息获取

地理信息包括三部分，京津冀边界坐标、京津冀路网信息以及威胁京津冀安全的地理位置坐标。 {50%：这三部分构成了京津冀的安全地图。} 三种数据的爬取方法各不相同。

3.2.1 京津冀边界坐标获取

京津冀边界的坐标获取相对简单，可以直接使用高德地图提供的JavaScriptAPI中的绘制行政区划边界的功能。以朝阳区为例，逻辑如图3.2所示。 {47%：首先连接地图，设置高德地图的中心以及缩放级别等基本信息。} 完成后加载行政划区插件，实例

{61%：图3.2 行政区边界坐标获取流程}

化插件，执行查询命令并获取朝阳区边界坐标，完成后将坐标保存在本地。 通过百度得到京津冀行政区名称列表，从列表中获取行政区名称，依次执行上述逻辑，得到所有的地图边界坐标。 因为是通过JavaScript在浏览器上爬取的坐标，没有遇到ip被ban的情况。 数据的完整性也有一定的保障。

下面将爬取的行政区坐标与名称对整理成geoJson格式。 GeoJSON是一种对各种地理数据结构进行编码的格式[12]，它可以表示点、线、多边形等图形。 这种格式是标准的JavaScript绘制地图数据格式，d3.js对于地图的绘制使用的数据格式主要为GeoJson格式。 其标准格式如图3.3所示。

图3.3 GeoJson标准格式

京津冀边界数据属于多边型，故其geometry属性内的type属性为Polygon，同时其还有name属性，值为各个行政区的名字。 另外还要加上id属性，以区分各个行政区。 获得的行政区边界坐标写在geometry属性内的coordinate属性中。 {45%：至此，京津冀边界坐标格式化完成。}

3.2.2 京津冀路网数据获取

京津冀路网数据的数据量较大，使用和京津冀边界数据获取相同方法的话，效率太低，可操作性太差。 故此处使用python的scrapy爬虫框架爬取京津冀路网数据。 首先获取道路名称，道路名称来自安居客网站。 然后通过获得的道路名称为keyword来访问高德地图的数据接口获得道路坐标。

下面介绍道路名称获取方式。 道路名称的获得来源是图吧网站（http: //www.mapbar.com）。 {47%：图吧公司全名北京图吧科技，是国内最专业的电子地图服务提供商。} 道路来源url为http: //poi.mapbar.com/tianjin/G70/，其面如图3.4所示。 其url中的tianjin即为天津。 同理，beijin即为北京。 因我们所需城市道路名仅限京津冀三地，故此处简单的复制粘贴即完成道路名获取。

图3.4 图吧天津道路名

下面介绍高德地图的数据接口。 高德地图的数据接口如图3.5所示。 其中传递的参数由 号连接，返回格式为json。 下面介绍几个重要参数，参数中city参数表示城市定位，keywords表示请求关键词，在这里添加道路名称，然后修改城市名称即可获得相应道路的信息。

图3.5 高德地图接口示例

下面介绍scrapy爬取道路坐标的逻辑。 先介绍 scrapy的调度过程，如图3.6所示，调度器先从 spiders中获取 request，然后经过 downloaderMiddlewares发给 Downloader， Downloader从网络中获得 response传给 spiders进行处理，完成后将 items发给 item pipeline进行进一步处理（保存或者丢弃）。

图3.6 scrapy工作流程

在这里我们稍微修改一下其执行过程。 因为我们的道路坐标需要被整理成GeoJson格式，故我们将舍弃itempipeline模块，在spiders中直接将爬取的数据进行处理并保存。

道路坐标的爬取逻辑如图3.7所示。

图3.7 道路坐标的基本爬取逻辑

其中高德地图返回的json中status如果为1则为正常，如果不是1，则说明请求出现了错误。 实际操作发现，每爬取100条道路数据，高德地图的服务器端就会封锁本机ip。 这时候就要进行代理ip的修改。 为了实现代理ip的修改，我们在DownloaderMiddleware中加入了httpProxyMiddleware中间件，并在settings.py中注册。 其主要功能是在遇见 status为6的时候切换代理 ip，并且在代理 ip均不能用时，在 66 ip. com、 httpdaili. com等免费代理 ip网站中爬取可用 ip并保存供接下使用。

此部分数据不同于京津冀边界数据，通过高德地图的 api获得的 json数据中， 地理坐标在 json[’ data ’][’ poi_ list ’]中 typecode属性为190301内 domain_ list块内的第四个元素的 value属性中， 如图3.8所示。

图3.8 高德地图返回数据示例

返回的数据中，有部分数据发生value属性为空的情况，这部分数据经检验均为非要道的道路坐标，故处理方式为将数据舍弃，道路名称删除。 另外，此数据在进行geoJson格式转换时的geometry属性内的type属性为MultiLineString。

3.2.3 京津冀加油站、化工厂、事故多发路段坐标获取

京津冀的加油站、化工厂坐标在高德地图上有标注，故此两地的坐标可以通过高德地图的api获得。 方法同上一部分的路网数据获取一样，此处不做赘述。 而事故多发路段在高德地图中没有标注，通过调查发现，各地事故多发路段随着时间的不同，其分布也不同。 {40％：因为事故多发路段的定义为如某路段交通事故次数相对较多事故造成的伤亡情况相对较为严重就可以认定为事故多发路段。} 所以这种路段并不固定。 因此我通过一段时间的搜集，获得了一些历史事故多发路段的位置，因位置分布较少，故采用手动确定坐标。

另外这里的位置坐标在GeoJson中geometry属性内的type属性为point。

3.3 社交网络信息获取

{46％：社交网络数据的来源是新浪微博，爬取流程如图3.9所示。} 首先通过selenium中的webdriver登录微博，以获取微博的cookie，将cookie存入redis。 Selenium是一系列的网页自动化测试工具，并且被用在很多工业项目中[13][14]，包括Selenium webDriver和Selenium IDE。 Selenium webDriver可以用来创建健壮的、基于浏览器的自动化套件或测试。 而Selenium IDE用来创建快速的bug测试脚本[15]。 这里用webDriver与chrome浏览器来模拟微博登录以获取cookie。 {42％：而redis是一款开源的，基于内存的数据结构存储器，可用作数据库，缓存和消息代理。} {41％：它支持的数据结构，包括字符串，散列，列表，集合等。} 这里用做存储维护cookie和维护request队列。

{67％：图3.9 微博数据爬取流程}

{96％：而mongodb是一个基于分布式文件存储的数据库。} 其功能丰富，数据结构较mysql要松散，类似于json，且部署方便。 Mongodb非常适合网站实时数据处理。 故对于相对于mysql，它更适合存储通过scrapy爬取的微博信息。

{46%：微博信息的主题爬取类似于高德地图的webapi。} 即通过访问微博的search功能url，将keyword与查询时间间隔拼接在url中，获得response，提取数据。 url的格式如图3.10所示。

{63%：图3.10 微博主题信息爬取url}

获得的response经过xpath的筛选，获得微博的ID、微博内容、点赞数、转载数、评论数等item传入pipeline。 {95%：Xpath是即为XML路径语言，它是一种用来确定XML（标准通用标记语言的子集）文档中某部分位置的语言。} Scrapy应用xpath来在response中寻找相应的item。 在pipeline中，将相应的item存入mongodb中。 而相应主题下的微博的评论者以及转发者是后面构建关系网络的关键，这里会在微博的评论和转发中找到作者的主页， 并查询其关注者和被关注者，同时以 edge的形式保存 id（例如1944388654l002357894代表 id为1944388654的人关注了id为1002357894的人）， 同时对找到的关注者和被关注着继续进行查询及保存，这样就得到了微博用户的关注与被关注数据。

因为对于微博的访问存在cookie，多以对于爬虫的访问，微博的反爬虫机制并不会频繁触发。 同时，因为代理ip会拖慢爬取速度，故本部分并没有使用httpProxyMiddleware。

爬取后的数据结构如图3.11所示。

图3.11 mongoDB中的微博数据结构

3.4 本章小结

本章介绍了数据获取的整体设计、地理信息获取过程以及方法、社交网络信息获取的过程以及方法。 其中地理信息的获取分为三个部分， 分别是京津冀边界坐标的获取、京津冀路网数据的获取、京津冀化工厂、加油站以及事故多发路段的获取， {43%：京津冀边界坐标的获取是通过高德地图的 JavaScriptAPI的绘制行政区划边界功能获得；} 京津冀路网数据是通过高德地图的webapi获取，为了解决ip被禁的情况，通过httpProxyMiddleware来更换代理ip以及爬取免费可用的代理ip； 京津冀加油站和化工厂坐标的获取同京津冀路网获取方法相同，而 京津冀事故多发路段的坐标获取则因为能得到的数据量过小而通过手动获取。 而社交网络信息的获取具体为通过 Selenium webdriver模拟微博登录获得 cookie， 保存在 redis中，通过和高德地图 webapi相似的方法拼接关键字访问微博的搜索功能， 获得相应关键字下的微博信息、发布者信息以及评论信息等，并通过 scrapy的 pipeline将数据保存在 mongoDB中， 同时在主题微博下爬取用户信息，通过有向 edge的方式保存关注与被关注关系，并在关注者与被关注者中继续进行关注者被关注者的爬取， {54%：得到的信息继续保存，获得微博用户的关注与被关注数据。}

第四章 数据分析

4.1 安全地图构建

D3.js内部自带地图绘制函数，地图的绘制比较方便，具体流程见图4.1。

图4.1 安全地图绘制流程

首先从本地取出京津冀地理信息的json数据，具体格式为GeoJson格式，格式的详情见第三章3.2小结地理信息获取。 {50%：检查数据格式，对不符合规则的格式进行修改调整。}

然后用d3.js进行地图的绘制，具体实现如下：

1） 设置用于地图展示的svg长宽。 var width = 1500; var height = 1500;

{ 41 %： 2） 设定映射函数，用于将地图实际坐标映射到svg中。 }

projection = d3.geo.mercator()

center([117.19， 39.14])//设置映射中心坐标

scale(15000)//设置缩放比例

translate([width/2， height/2]); //设置映射中心

var path = d3.geo.path()

projection(projection); //将投影函数应用在地图中

3） svg画板的创建。

var svg = d3.select(" body ").append(" svg ")//在dom中创建svg

attr(" width " ， width)

attr(" height " ， height)//设置长宽

append(" g ")

attr(" transform " ， " translate(0， 0) "); //无偏移

4） 读取京津冀地理信息json数据。

d3.json(" ./json/beijingR.json " ， function(error， root) {//绘图位置}

{ 46 %： 5） 在4）中的绘图位置中进行地图绘制，方法如下所示： }

svg.selectAll(" path ")

data(root.features)//features是GeoJson中的属性

enter()

append(" path ")//绘制

attr(”stroke”, ”#000”)//设置线条颜色

attr(”stroke-width”, 0.3)设置线条宽度

attr(”fill”, function(d, i){

// return color[0];

return ’none’

})设置颜色

attr(”d”, path)

至此，基本地图绘制完成。 效果如图4.2所示。

图4.2 安全地图效果展示---天津

此处地图为天津市部分地图，其中绿色代表事故多发路段，浅蓝色代表化工厂，橘黄色代表加油站。

细节调整指的是增加地图背景色，调整路网路线的宽度以及化工厂、加油站、事故多发路段的标注点的颜色大小的调整。 目的是增加这三者在地图上的鲜明程度，以及地图的可看性。

至此，京津冀安全地图绘制完成。

4.2 微博数据的处理与分析

{45%：微博数据的处理包括微博用户关系网络的构建、相关属性的计算分析以及验证。} 本小结流程如图4.3所示。

{58%：图4.3 微博数据分析及验证流程}

4.2.1 关系网络的构建

{54%：关系网络的构建数据来源于微博用户。} {48%：数据爬取过程见第三章3.3社交网络信息的获取。} 数据格式为svg，内部分为两个条目，分别是source和target，内容为用户id。 每一条数据代表source内相应id的用户关注了target内相应id 的用户。 将svg文件导入到gephi中进行网络属性的分析以及网络的可视化。 Gephi是一款开源的图形和网络分析软件，可以用来计算网络的属性以及网络节点的可视化、操纵等[16]。 同时其支持大数据节点的计算。 导入数据后，调整其布局，gephi支持openord布局，这是一种基于力导向布局算法，支持多核、并行，速度快，效果明显。 调整后的布局（部分数据）如图4.4所示。

{55%：图4.4 微博用户关系网络布局（部分）}

此布局为部分用户关系构建的布局，数据量为166292条边，1650个节点。 考虑到gephi的节点展示对于过多节点不够明显，故用该示例展示其布局。 其中红色的为边，从图中红色的深浅可以明显看出此关系网络分为两个部分，因此初步推测该网络具有社区。 所挑选的数据为关注以及被关注用户不同层级均匀抽样选取的，所以此布局能够代表所有用户关系的总体布局。

4.2.2 属性的计算及分析

首先通过gephi计算其度分布，如图4.5所示。 {50%：通过其分布能看出其符合幂率分布，即少数节点拥有较大的度分布，而大部分节点拥有较小的度分布。} {41%：节点的度指的是和其连接的节点的连线的个数，而微博关系网络是有向图，} {56%：故有入度和出度，入度指的是指向该节点的线的个数，} 出度指的是该节点指出的线的个数。 拥有入度多的节点代表其有更高的接纳度或者更受欢迎，拥有出度多的节点说明其影响力更大，有更多的朋友。 {51%：微博用户关系网络的入度和出度分布如图4.64.7所示，从图中可以看出，} {46%：微博关系网络的出度幂率特性要高于入度幂率特性。} {41%：这由微博的关注功能决定的，普通用户可以关注任意其他用户，这使得微博关系网络中用户的出度要高于入度。}

{67%：图4.5 微博关系网络的度分布}

{67%：图4.6 微博用户网络入度分布}

{65%：图4.7 微博用户网络出度分布}

{51%：微博的小世界特性从两个方面来分析，分别是平均最短路径长度和聚类系数。}

{81%：网络中节点i与j之间的距离dij定义为这两个节点之间最短路径的边的个数。} {87%：网络中任意两个节点的距离的最大值称为网络直径。} {91%：网络的平均路径长度L定义为任意两个节点之间距离的平均值，即有：}

$$L = 1/[0.5N(N+1)\sum i \geq j dij]$$

其中，N为网络的节点数。 {92%：研究发现，尽管许多实际的复杂网络的节点数很大，但网络的平均路径长度却很小。}

{51%：本网络的平均最短路径长度为3.18，对于微博这样的社交网络来说，这个数值是很小的，即微博社交网络具有较短的平均最短路径长度。} 说明了微博用户之间平均通过3-4个人就能彼此建立联系。 而其网络直径为9，即最多通过九个人，距离最远的两个用户就能彼此联系，鉴于其平均最短路径长度为3.18，这个数值应该属于极个别的情况。

{54%：在现实社会中，你会发现你的两个朋友可能也是朋友，这种属性即为网络的聚类属性。} {63%：假设网络的一个节点i与k个节点相联系，这k个节点即为节点i的邻居，显然在这k个节点之间最多有k(k-1)/2条边。} {69%：而这k个节点实际相互联系的边数E和总的可能联系的边数之比即为i的聚类系数Ci，即：}

$$Ci = 2Ei/[ki(ki-1)]$$

{83%：所有节点的聚类系数的平均值即为该网络的聚类系数。} 其取值从0到1。 {52%：当网络完全连通

时，即网络中任意两个节点均有连接时，聚类系数为1。｝｛54％：一个完全随机的网络含有N个节点时，其聚类系数随着N的增大而趋近于O(N-1)。｝

｛67％：而微博用户网络的聚类系数为0.331。｝ 经调查，和微博用户网络相似的随机网络的聚类系数小于0.1，这说明微博用户网络具有较大的聚类系数， 证明某用户的粉丝之间也很可能存在关注关系。

4.2.3 结果的验证

本文选取了2017年五月15日央视新闻微博发出的一条天津中医药大学宿舍楼起火的微博，分析了其传播特征以及传播速度。 如图4.8所示

图4.8 央视微博的转发时间趋势

在微博发出之后就立刻受到了100多条的转发，经过两天之后转发量逐渐下降。 而其总转发人数超过1000，其中天津用户的转发量达到了34.3%，位居榜首。 ｛43％：值得注意的是天津用户微博转发的平津值在全国为1.6%，见图4.9。｝

图4.9 央视微博各地区转发占比

｛45％：（其中，黄色代表本条微博的转发量，蓝色代表各地平均转发量）｝

这说明微博的消息对于涉及到本地区的人来说比其他微博更重要。 而图4.8可以看出，微博上一条消息传播呈现出爆炸式增长，而以比其增长速率低的多的速度回落。 而一条信息从发布到引起舆论事件的速度也快的惊人，这一点可以从微博用户网络的无尺度特性来解释。

｛40％：微博用户网络的小世界特性使得消息的传播快速而广泛，如图4.10所示，｝ 微博的入度更多的用户（央视新闻）在消息传播过程中扮演着很关键的角色， ｛43％：新闻由央视微博发出后，经过上百用户的转发，占总转发量的94%。｝ 而这些二级转发者的微博被转发的则寥寥无几。 ｛40％：这从侧门说明了微博用户网络的两个特性，以及微博用户网络中大v用户和知名用户在信息传播过程中扮演者重要的角色。｝

图4.10 央视微博转发关系

4.3 本章小结

本章中，我们介绍了安全地图的构建过程以及其效果图。 通过 d3. js的地图绘制功能，将京津冀地图加以绘制， 同时加入路网以及威胁京津冀民众安全的加油站、化工厂和事故多发路段的坐标以及绘制， 并展示了其效果图。

同时我们着重介绍了微博用户网络的分析以及验证。 ｛42％：通过gephi构建了微博用户网络，以及计算了微博用户网络的无标度属性和小世界属性。｝ 这两个属性说明了微博上消息传播的过程呈现迅速而广泛，同时某些关键节点（微博大v、知名人士等）在消息传播过程中扮演者关键的角色。 另外我们还发现了微博内容涉及到的地点的人会对该微博投入更多的关注。

第五章 平台实现

5.1 平台需求分析

该平台搭建的目的是用来给京津冀民众查看，给生活在标注有潜在危险地点的民众以安全提醒。 同时借助Mediawiki良好的扩展性，该平台还可以给民众提供一个针对京津冀的维基平台。 民众可以自行上传京津冀的任何安全信息。 也可以对错误信息进行修改。 具体需求如图5.1所示。

{57%： 图5.1 京津冀安全平台用例图}

{46%： 本平台面向两个角色，平台的查看者和平台的建设者。} 查看者可以查看京津冀各省市的安全地图，同时还可以查看平台对京津冀各省市收录的信息，如省市简介、基本情况等。 而平台的建设者可以修改平台上的错误信息，同时可以创建关于京津冀的其他页面。 本平台还有管理员角色，此部分功能由Mediawiki框架提供，这里不做赘述。

5.2 平台设计

得益于Mediawiki框架的强大功能，本平台不必涉及到数据库存取的部分以及前后端数据交互部分，这能让我们专注于平台页面。 我们的平台需要设计的是页面布局以及页面链接逻辑等。 平台的状态图如图5.2所示。 打开平台显示主页，主页显示京津冀的总体信息以及京津冀地图。 点击地图某一地区，跳转到该地区主页，地区主页主体显示该地区的安全地图，地图具体格式见第四章4.1安全地图构建。 同时还有该地区的基本信息，包括简介、历史等链接，点击后进入具体介绍页面。 用户可以点击Mediawiki自带的编辑按钮进入编辑页面进行错误修改，也可以在搜索栏搜索新页面，不存在的话可以创建新页面进行新页面编辑。

{48%： 图5.2 京津冀安全平台页面跳转状态图}

5.3 平台实现

Mediawiki平台支持自带的语言进行页面编辑，因此本平台的编辑者角色需要对Mediawiki平台的文本编辑语言有一定的了解。 可喜的是，其文本编辑语言很容易学习。 而与此同时，Mediawiki平台也支持html代码的内嵌，故我们可以使用d3.js，这让我们能够将京津冀安全地图嵌入Mediawiki平台上。 值得注意的是，d3.js所引用的json文件需要在Mediawiki所在的www目录下，并且与正常的json文件引用有一些不同。 区别如下所示：

正常的文件引用： d3.json(”./json/tianjin.json”, function(error, root) {}

Mediawiki中的文件引用： d3.json(”../json/tianjin.json”, function(error, root) {}

这是因为，Mediawiki的页面加载是通过index.php中的函数加载的，而json文件在index.php的同级目录中，所以需要向上翻到二级目录中。

平台主页如图5.3所示：

{56%： 图5.3 京津冀安全大数据平台主页}

通过点击各地区，跳转到相应的省或者市的主页，省、市的主页包括其安全地图， 即地区边界、路网以及对

当地有威胁的地区标注，还有该地区的一些导航，包括该地区的简介、大事记以及生活导航的链接， 点击链接即可跳转到相应界面查看信息。 以天津为例，如图5.4所示。

图5.4 天津安全地图样例

其中不同颜色的点代表不同的地区标注，紫色的两个点分别为天津812事件以及2017.5.15日天津医科大学着火地点。 {40％：右下角的按钮点击后会跳转到相应的页面，页面如图5.5-图5.7所示。}

图5.5 天津简介样例

图5.6 天津大事记样例

图5.7 生活导航样例（其导航可为外链可为自创页面）

需要说明的是，出于安全考虑，天津安全地图中平台建设者只能增加事件标注点，而不能修改地图本身， 因为其json文件在平台上不可访问，而另外三个页面平台建设者可以自由添加以及改正。 每个页面点击左上角的图标就可以返回主页面，修改可以通过点击右上角“阅读”字样旁边的“编辑”字样来进行编辑。

5.4 本章小结

本章介绍了京津冀安全平台的需求分析、设计与实现，得益于Mediawiki的强大功能，我们只要专注于平台的跳转逻辑就可以了。 在平台中，平台的查看者可以通过点击地图、点击按钮查看不同的页面，而平台的建设者也可以通过Mediawiki带有的编辑功能进行页面添加、页面编辑、错误修改等。

值得一提的是，平台的维护者即为Mediawiki的管理员，因为这是Mediawiki本身的功能，这里不做赘述。 而在各地区的安全地图中，平台的建设者不能进行地图的修改，只能进行标注，因为地图的json文件被保存在Mediawiki所在的目录下，平台建设者看不见。

第六章 小结与展望

6.1 未来的展望

{44％：本项目包括两个部分，微博用户关系网络的分析以及京津冀安全平台的构建。}

对于网络方面，受限于数据，微博用户关系只是限定在过去一段时间，网络的属性可能会随着时间的变化而变化， 但是因为微博用户关注方式的固定，我们得出的结论不会发生本质性的变化。

至于京津冀安全平台方面，我的工作只是搭建了京津冀安全地图以及设计了该平台的基本页面逻辑， 这个平台的定位是维基性质的京津冀安全信息平台，其未来的建设不可限量， 可以包含京津冀安全的各个方面的信息。{50％：但是其要依赖于京津冀民众的自发建设。} 我相信，这样一个关乎京津冀民众生活以及出行的平台，会收到欢迎，它未来一定能够成为京津冀民众生活中不可缺少的一部分。