第1章 引言

1.1 项目背景

当今各媒体上选秀节目层出不穷，各种各类的参加选秀的选手也是数不胜数。观看选秀节目的观众或多或少会对某些选手有着特别的关注与留意，不少观众甚至会把他们心中的那份关注发上微博，并且各类有影响力的公共微博也会发布一些精彩的选秀视频。同时，由于新浪微博的巨大用户量，许多选秀节目也鼓励观众利用微博对观众自己所支持的选手进行评论、关注甚至投票。因而每位参赛选手在整个选秀比赛中的关注率走势变化以及关注率在不同群体中的分布一定程度上能够体现选手的当红程度、粉丝阶层等信息，例如在今年红透半边天的中国好声音，最终夺冠的选手梁博从倒数第三轮比赛开始人气就开始快速的飙升，在决赛的当晚，微博人气更是达到顶峰，而这也最终预示了他的夺冠。而除了粉丝观众外，许多娱乐圈的公司、媒体等也会对这类信息有强烈的需求。

1.2 项目应用

本项目是基于新浪微博平台，挖取一段时间内新浪微博众多用户针对某个选秀比赛中的选手发表的微博以及发表这些微博的用户信息，通过图表数据分析，以网站的形式展示出指定选手的关注率走势图(从比赛开播起)，该选手相关的热门关键词，以及关注该选手的用户的群体信息，包括年龄分布、地域分布等。通过这样的分析，我们可以知道选手支持者的年龄分布、社会背景、教育水平和地域差异，这些信息对于粉丝而言是一种生活化的信息，增强相互的认同感，能够了解到自己所关注的明星一路的走势，对于一些公司媒体而言，也可以根据根据粉丝类型对选手未来的发展做更好的规划，达到有的放矢。

1.3 项目创新点

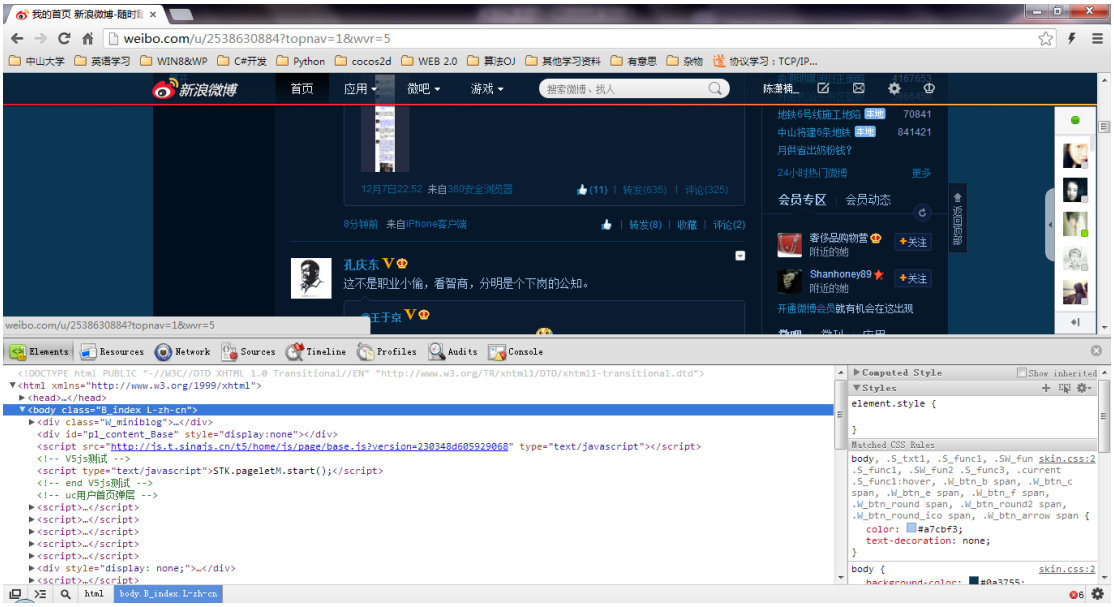
1. 利用图表的形式展现选手的粉丝信息，更加直观，一目了然。
2. 对微博的情况进行一路的跟踪分析，样本数量较大且具有持续性，降低了不确定因素的影响。
3. 结合了当下最流行的社交平台，在以后应用的扩展中更加灵活，空间更大。

第2章 功能需求

2.1 抓取数据

抓取数据的工作分为两个部分。

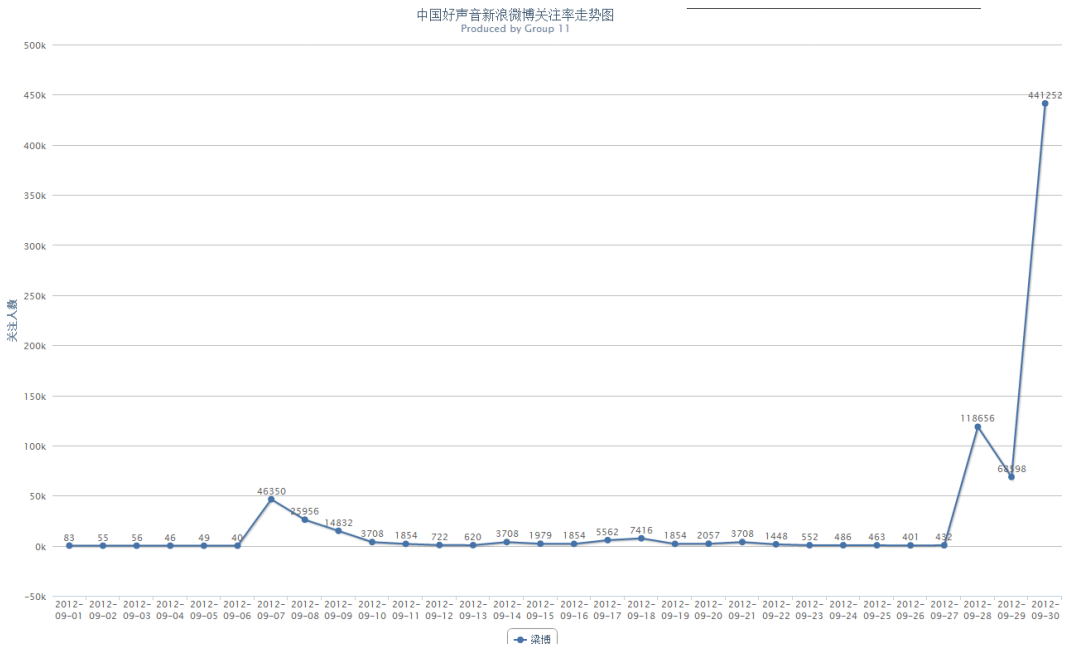
1. 通过网页爬虫，解析网页 HTML 源码，获得微博与评论资料。



2. 通过微博 API，获取用户的个人信息。

2.2 模型呈现

网页爬取完信息后，经过后台的统计与分析，可以得到具体某天的总的关注数，记录在数据库内。需要呈现时，通过访问数据库中已经统计好的数据，在网页上通过图表的方式显示统计结果，如下图，能得到梁博关注率的折线图。

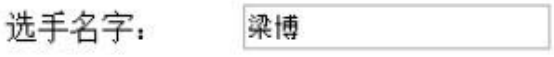


第3章 总体设计

3.1 工作流程

✓ 用户进入应用，输入选秀活动的名字和开播时间，以及播出间隔。

✓ 用户选择想要了解的选手信息。



✓ 点击提交，提交信息。

✓ 系统在之前爬取好相关数据，直接从数据库中取出数据生成图表返回给用户。

✓ 用户在结果页面可选择查看关注率走势变化折线图、关注群体分布图、热门关

键词等，并可以通过一些相关链接去了解活动信息、选手信息以及关键词新闻。

3.2 技术支持

- ✓ Python 第三方库，Splinter.py 是一种自动化回归脚本的方式，可以重复性的回归现有功能，并给出回归测试报告。Python 的世界有一个开源框架 Splinter，可以非常棒的模拟浏览器的行为（从某种意义上也可以说是人的访问点击行为）。Splinter 提供了丰富的 API，可以获取页面的信息，以判断当前的行为所产生的结果。
- ✓ pyquery 库，它是 jQuery 的 Python 实现，可以用于解析 HTML 网页内容，我个人写过的一些抓取网页数据的脚本就是用它来解析 html 获取数据的。可读性非常棒!
- ✓ Pyquery 是一个类似 jquery 的库(主页 http://packages.python.org/，就像作者说的 Hey let’ s make jquery in python)，通过使用 lxml 来处理 xml 和 html. 所以在使用 pyquery 时得先安装 lxml 库。
- ✓ 基于 jQuery 的 highcharts.js，Highcharts 是一个用纯 JavaScript 编写的一个图表库，能够很简单便捷的在 web 网站或是 web 应用程序添加有交互性的图表，并且免费提供给个人学习、个人网站和非商业用途使用。目前 HighCharts 支持的图表类型有曲线图、区域图、柱状图、饼状图、散状点图和综合图表。

第4章 算法设计

我们在获得情感倾向为喜好选手或关注选手音乐的用户的时候采用了朴素贝叶斯过滤算法。

4.1 朴素贝叶斯分类

贝叶斯基本公式：

- ✓ $P(H|T) = \frac{P(T|H)*P(H)}{P(T|H)*P(H)+P(T|M)*P(M)}0$
- ✓ $P(H|T_1, T_2 \dots T_n) = argmax\ P(T_1, T_2 \dots T_n) * P(H) = argmax \prod P(T_i|H) * P(H)$

以提及梁博的微博作为训练集，选取大约 3000 条左右的微博进行人工训练，分成 Hit 和 Miss 两类。再利用 jieba 分词组件对微博进行分词，就可获得 Hit

类别的 Tokens 和 Miss 类别的 Tokens。

如此，我们便可求得先验概率 P(Ti | C)和 P(C)：

P(Ti | C) = (C 分类里的 Ti 出现次数)/(总词数)

P(C) = (C 分类里的总词数)/(总词数)

算得每个 Token 的 P(Ti | C)后，朴素的分类器便基本完成。

对于每条待分类的微博 W，我们使用上述极大似然估计的复合概率公式可计算出 P(C| T1,T2 … Tn),即 P(H|W)和 P(M|W)。两者比较大小即可决定该条微博究竟是 Hit 还是 Miss。

4.2 Laplace 平滑

新关键词出现时假定出现次数为一，防止其概率为 0。

✓ $\frac{N(T=token|H)}{N(all_{tokens})} \longrightarrow \frac{N(T=token|H)+1}{N(all_{tokens})+V}$

V 为整个词库的总词数(不算重复)。

4.3 Tf-idf 加权技术

✓ $tf\text{-}idf(T_i, C) = tf(T_i, C) * idf(T_i)$

✓ $idf(T_i) = log(\frac{N*C\_T_i}{N\_T_i})$

说明： Ti：某一个关键词 Token

C：某一特定分类，即 Hits 或 Miss

N：微博总数

N_Ti：含关键词 Ti 的微博数

tf-idf 是为了去除信息量低的词。我们计算好 tf-idf 排序后选取了信息量前 90%的词来计算概率。

第5章 测试文档

5.1 测试内容

1. 爬虫功能实现。
2. 微博 API 抓取数据。
3. 图表显示用户信息分布。

5.2 测试报告

功能模块	整体功能	**测试员**	吴文杰、陈照	**测试时间**	2012/12/3
前提条件	用户授权	**参考条件**	无	**功能**	整体功能
测试数据	无	**测试目的**	业务流程是否能够正常运作		
操作步骤	**操作描述**	**数据**	**期望结果**	**实际结果**	**测试状态**
Auth 2.0 验证	用户授权应用	Code	授权成功	授权成功	良好
接口调用	调用 weibo API 提取用户信息	用户信息	获取 json 类型的服务器返还数据	获取 json 类型的服务器返还数据	良好
写入文件	API 获取的用户信息写入文件进行保存	无	文件以 utf-8 编码输出用户信息	文件以 utf-8 编码输出用户信息	良好
爬虫	下载得到搜索结果 Html 并解析	搜索结果 Html	得到搜索结果的 Html	Html 正常获取并存储在内存里，因为马上要进行解析处理，所以没有存放到硬盘上	良好
抓取数据	爬取微博搜索结果，得到用户列表	用户列表	得到关注某位明星的一众人群的用户 ID，以作为后期信息	结果正确输入到数据库内，后期可以直接从数据库中提取使用	正常

			统计		
数据库读取	得到生成图表的数据	数据库记录的统计信息	能够获得数据库内相应的记录，如梁博关注率的连续观察值	从数据库中得到了相关的信息，并最终生成了需要的图表	优秀
图表展示	根据得到的记录显示相应图表	无	输出相关图表	输出了一张关于梁博关注率的折线图	优秀

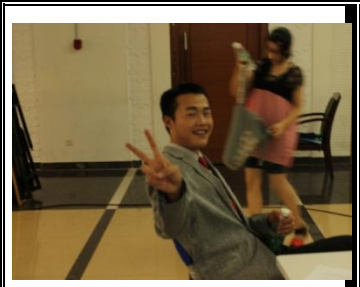
第6章 附录

6.1 资料引用

作者	资料名	资料格式	出版社
Magnus Lie Hetland	Python 基础教程(第二版)	纸质书	人民邮电出版社
Jonathan Chaffer	JQuery 基础教程(第二版)	纸质书	人民邮电出版社
豆丁网分享	Python 抓取页面	电子文档	www.docin.com
豆丁网分享	Python 爬虫参考资料	电子文档	www.docin.com
Yulin Fu	SINA WEIBO CRAWLING	电子文档	
Xuewei Wu	Weibo API	电子文档	

6.2 团队介绍

	队长：赵哲民 个人简介：一位具有非凡创造力与领导能力的 90 后大学生，商业头脑发达，被誉为中国下一位商界奇才，比肩扎克伯克，横扫世界五百强。目前处于待业阶段。
	队员：吴文杰 个人简介：努力向上的程序员学徒。
	队员：陈照 个人简介：一个勤勤恳恳的码农。
	队员：蔡志杰 个人简介：外表粗犷，内心细腻如丝，是一个容易受伤的男人。也正因为他的敏感，给了他难以度量的敏锐和独到的眼光，一次次对 IT 业界市场的预测简直惊为天人。也只有 DOLCE 这样高贵的品牌才能够配得上他。



队员：陈潇楠

个人简介：个性比较低调。

6.3 任务分工

	创意构想	文档撰写	抓取数据	算法分析	页面展示
赵哲民	O				
吴文杰	O	O	O	O	O
陈潇楠	O	O		O	O
陈照	O	O	O	O	
蔡志杰	O				

第7章 项目总结

自 11 月份开始，我们一直在进行 weibo 应用抓取数据项目的开发工作，至此为止已近 1 个月时间，从小组内部创意提出、代码编写，功能测试，以及最后的文档撰写等等。从开始到项目即将结束，一步步走过来。本次项目中，在文档的最后一部分，我们仅对此项目中测试工作进行总结。

一、项目测试进度控制。项目的测试进度主要是按照项目计划进行的，完全按照小组计划要求完成测试任务、提交测试类相关文档，包括测试案例的完善、制定测试计划、执行测试、缺陷跟踪以及 BUG 回归测试等。协调项目的内部测试工作，本此项目中测试小组进行了全面测试工作，认真配合项目工作，共同保证项目质量。项目测试的问题跟踪及处理采用每日进行修改问题回归测试工作，每日同步更新问题跟踪单的模式，按照规划时间完成更新测试。

二、小组内部成员关系处理。在项目工作的这几个月里大家相处融洽，小组内部共同探讨解决问题的方法，向各模块负责人学习模块功能处理方式，向能力较强的队友了解项目中涉及的知识点，两者结合起来进行开发与测试。

三、使用第三方库：此系统中使用了 pyquery 与 highcharts 等第三方库。这些第三方提供的库在很大程度上满足了程序员开发的便利，与软件界面的需求，从而也给软件的操作带来了方便。在今后的项目开发过程中，要继续使用第三方的库。这样一来，无论是针对软

件界面的美观性、友好性来说、易操作性而言，还是针对系统开发效率而言，这都是很好途径。但需要意的是 :在是使用第三方库时 ,要谨慎的选择一些网络中的比较常见的第三方库。

四、设计和编码的区别其实很大的，设计就要考虑到整体的框架，整个软件的各个相关部分，以及更重要的是功能方面！不像编码那样，只要按照别人告诉你怎么做就可以了。这就要求考虑的东西更全面些，更多。

五、个人得失方面。作为此次项目开发的参与者，对于日常的代码开发、任务分配、工作执行、缺陷跟踪、协调内部等能力均得到了进一步提高，理清了项目整个过程中小组的工作过程以及后期的项目移交工作。同时也对相应的业务知识有了更进一步认知。相关应用开发知识方面还需要进一步加强，还要更多的打代码，在练习中进行学习，而并非局限于理论知识。更好的吸收项目经验，做好以后的工作及其他项目的开发工作。