确定问题

1. 用户最常讨论的事情是什么？为什么是这些话题？

2. 用户在什么时间段最活跃？

3. 什么内容的点赞，转发和评论数最多？

数据导入

使用navicat对数据进行导入，导入时，注意各类数据的格式，如varchat, int等

数据清洗

1. 选择字段

因为用户名和帖子编号已经加密，分析并没有意义，所以在导入navicat时，删除user\_id和post\_id，保留发微时间、点赞数、评论数、转发数和内容；将数值（点赞转发评论）设为int，其他设为varchar，目的是防止排序时，数值型数据按照字符串的排序规则排序

2. 删除重复数据

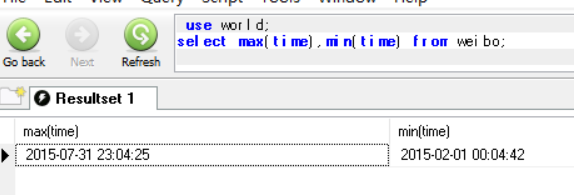
因为所有用户名都是随机加密，所以不存在重复的情况，导入navicat后，一共12200条数据

3. 检查缺失值

在navicat中勾选字段为not null，数据变为11792条

4. 日期数据清洗

数据的时间段为2015-02-01至2015-07-31，检查数据里最大最小时间是否在范围内



最大最小值均在范围内，符合要求

标签云的制作（查找高频词）

1. 复制weibo数据库，将复制的数据库内时间和点赞等日期，数字数据删除，只留下content字段

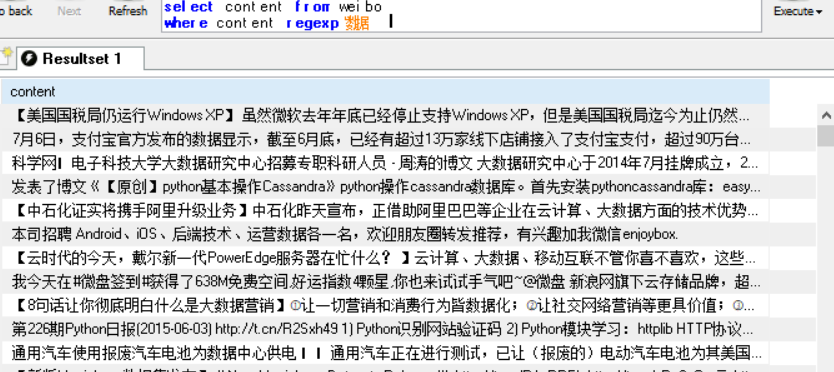
2. 使用图悦进行词频统计

3. 对统计结果排序，并筛选出前30个出现频率最高的词组

4. 清洗统计结果，比如转发，微博，详情等系统自带字眼

5. 标记出不能一眼理解的结果，或词组分割不理想，或疑似系统自带的词组，Sql内使用正则表达式，粗略分析这些词组出现的语境

如



根据查询结果，大数是大数据的未完全拆分，试试手是红包的关联词，所以删去这两处；同时发现，对“体验”、“新浪”、“存储”进行统计，发现都与一个产品相关联——微盘。

6. 整理数据，绘制标签云



分析问题

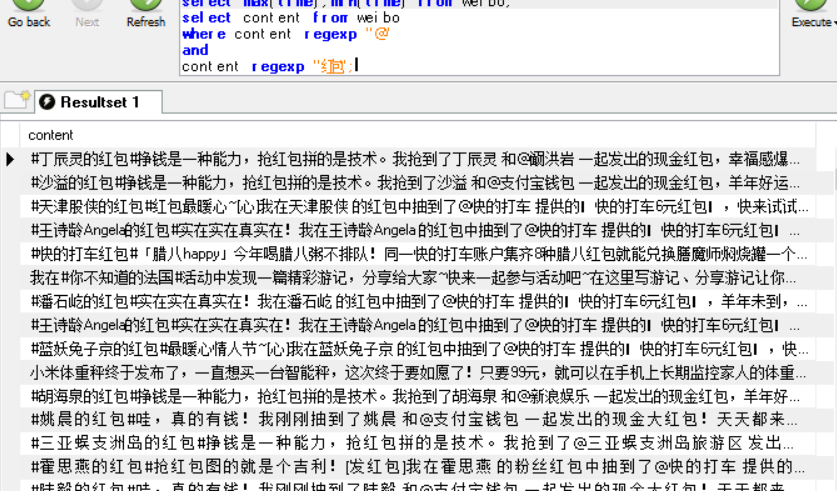
1. 用户最常讨论的事情是什么？为什么是这些话题？

根据标签云，最常讨论的是两个大类，消费方面（红包、代金券、现金等）和互联网方面（技术、设计、软件、智能、网络）

数据的时间是20140701到20141231，并不在过年期间，所以猜测应该是某产品的推广方式，选取带有红包的100条语句进行分析；发现出现频率较高的是某名人+@支付宝钱包的字样，比如沙溢+@支付宝红包；

所以更加肯定是‘产品推广’，按照这样的格式，匹配“红包，@”出现的次数：

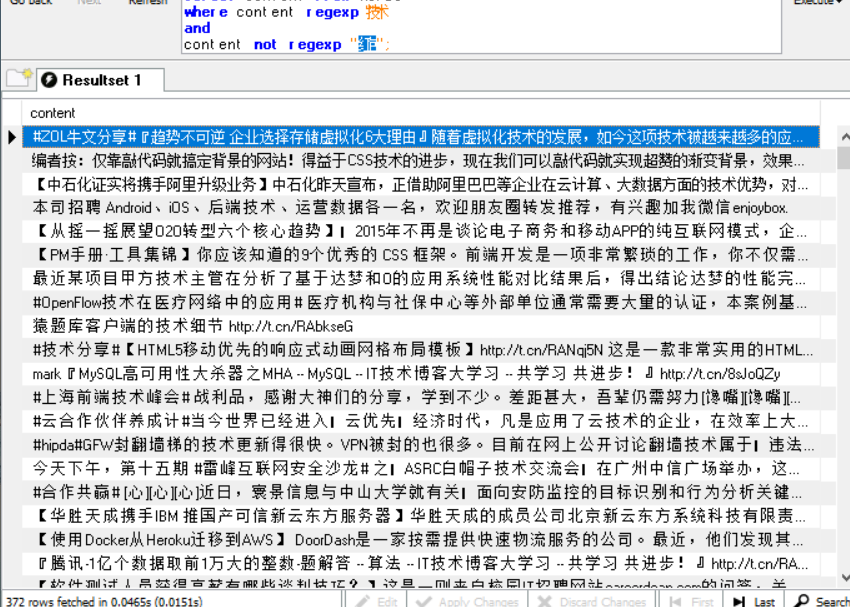
使用Sql的多字段同时模糊匹配：



说明这个时期红包关键词激增，是因为很多产品借助微博这个大平台进行推广。注意到打车这个字眼出现了很多次，根据匹配结果，快的打车在推广方面下了很大功夫。

另一类的话题是互联网领域，出现频率最多的是技术，估计用户可能偏爱互联网技术领域，比如前端、后端技术等，用sql匹配一下数据：

发现技术的匹配项分为两种，一种是互联网相关，一种是和红包相关（抢红包拼的是技术），所以把红包排除，再分析结果：

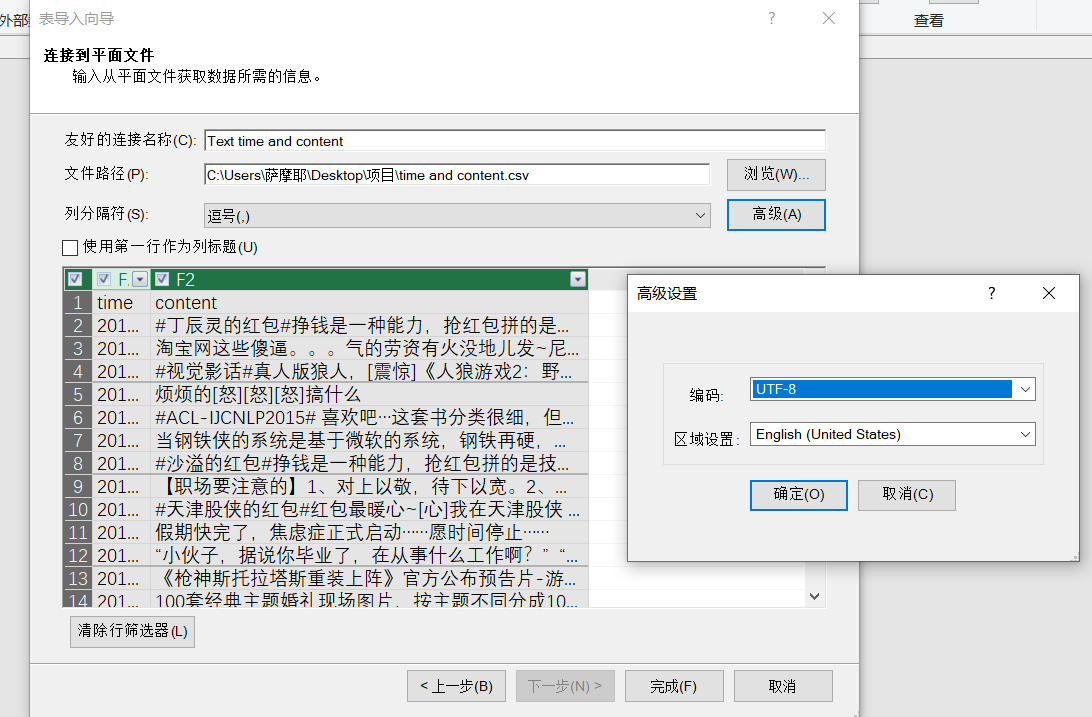


一共372条，这说明有一些用户是互联网领域工作或学习的人群；同样，对数据进行分析，也证明了部分用户确实是和计算机/互联网相关，一共512条记录；



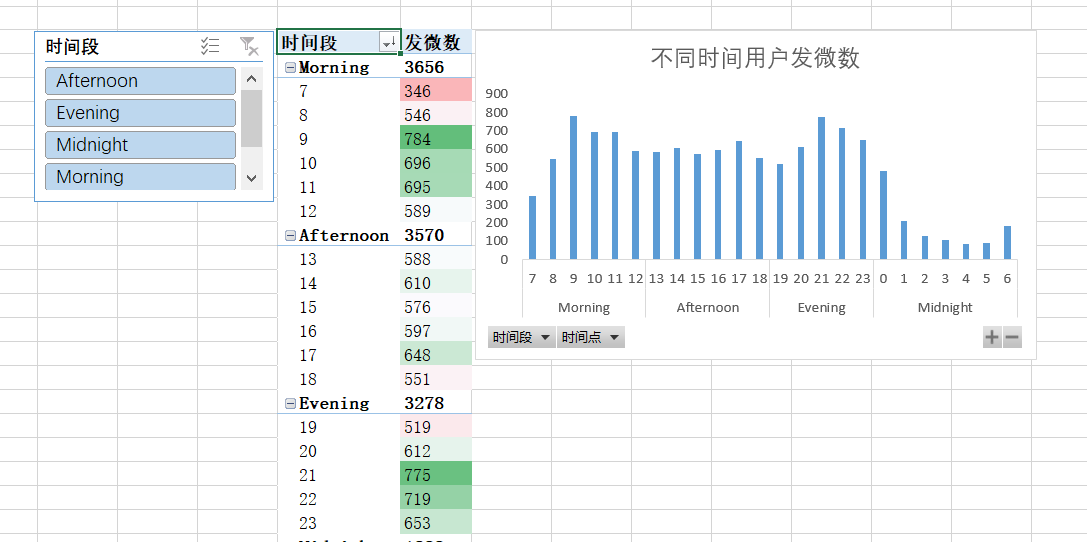
接下来分析，频率最高的几个词组，如红包、打车、技术、数据等，都是在什么时间段出现的，为什么会出现这样的情况：

2. 用户在什么时间段最活跃？



将时间和内容数据导入power pivot，如果中文出现乱码，那么将编码设为UTF-8；

得到结果如图



加入了切片器和数据透视图实现同步变化；发现用户在早上发微博次数最多，具体来看，用户在9-11点和21-23点发博次数最多，晚上发博次数多是因为晚上正是学生和工作者都有空闲的时间段；而早上发博次数较多的原因，可以推测是工作者在9:30上班，有时间看微博发微博，而学生则不那么受影响，早起后发微博也是常见的事情。如果要验证这个猜想，用户的年龄分布可能会有所帮助。

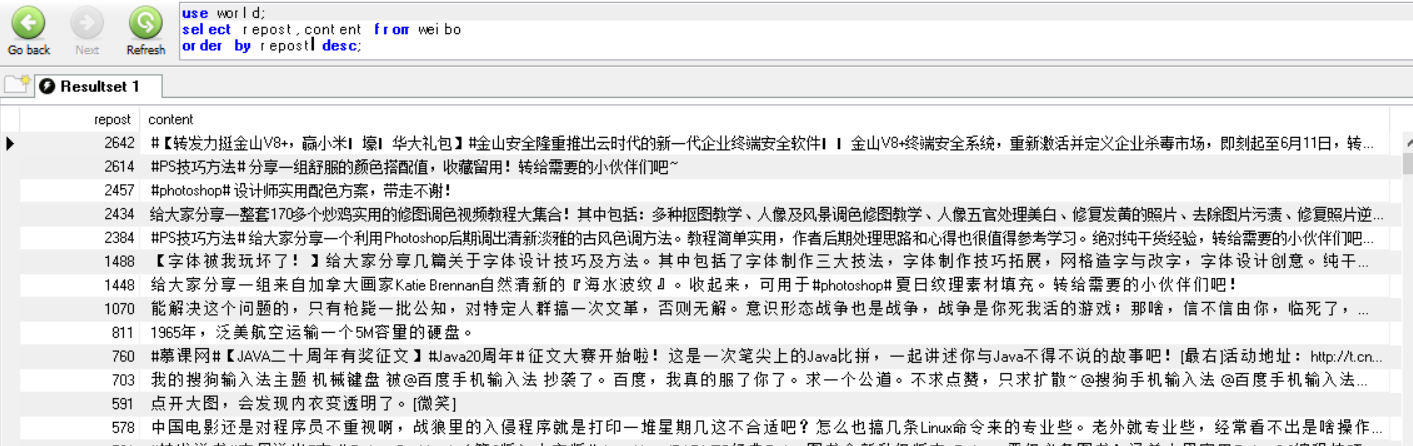
3. 什么内容的点赞，转发和评论数最多？

3.1 点赞数：



发现点赞数最多的八个，五个和办公技能相关，那么可以推测，产生这些数据的用户大多是需要做报告的人员，或者大学生，或者偏业务和运营的工作者。但是凭现有数据很难预测，因为无法得到用户点赞所发生的时间点，有了这个数据，就可以对假设进行分析和验证。

3.2 转发数



转发数超过1000的都是办公技能相关，验证了在“点赞数”中所做的猜想

3.3 评论数



评论数远远高于其他的，是“公车自拍”这种搞笑轻松类短视频，其次就是明星和社会热点问题。而点赞数，转发数很高的微博下评论数很少，这是因为用户只是做一个相应的记录，而不用去评论。