

## 你用 Python 做过什么有趣的数据挖掘/分析项目？ 修改

我最近刚开始学习 Python，numpy，scipy 等，想做一些数据方面的项目，但是之前又没有这方面的经验。所以想知道大家都做过什么有趣的项目，或者有什么好入手的方向推荐下 修改

[添加评论](#)[分享](#) • [邀请回答](#)[举报](#)

17 个回答

[按投票排序](#)  
140

何明科，不写程序的数据工程师不是好产品经理



贺行舟、刘看山、Blackjett 等人赞同

（注：本文部分回答来源于[你通过什么渠道获取一般人不知道的知识和信息的？ - 何明科的回答](#) 以及 [能利用爬虫技术做到哪些很酷很有趣很有用的事情？ - 何明科的回答](#)）从文中，大家也可以看到一个创业小团队艰辛的摸索过程。从一开始的一个想法，希望通过技术和科学改变世界，到碰巧能赚钱，到因为赚钱快而迷失了方向，到最后回归初心，做自己最喜欢的事情。

第零步：原点，大数据与价值

大概一年多以前，和几个小伙伴均认同一个趋势：觉得通过技术手段获取网上越来越丰富的数据，并基于这些数据做分析及可视化，必能产生有价值的结果，帮助大家改善生活。（大数据被叫烂了，所以用低调的方式来解释我们的初心）

第一步：开工，为基金服务

恰巧和几个基金的朋友（包括对冲基金和VC/PE基金）聊到这个趋势，他们非常认同这个观点并愿意付费，认为可以用这种实时且定量的方式来跟踪一些上市公司或者私有公司旗下的产品，来确定谁是有价值的投资目标。于是立马获得订单并促使我们开干，因为考虑到Python灵活及各类爬虫库的优势，最终选用Python来做数据获取的主体架构；也有新潮的小伙伴使用Go，同时用Go搭建了一个很酷的框架来制造分布式的智能爬虫，应对各种反爬策略。抓取数据主要来自于如下网站：

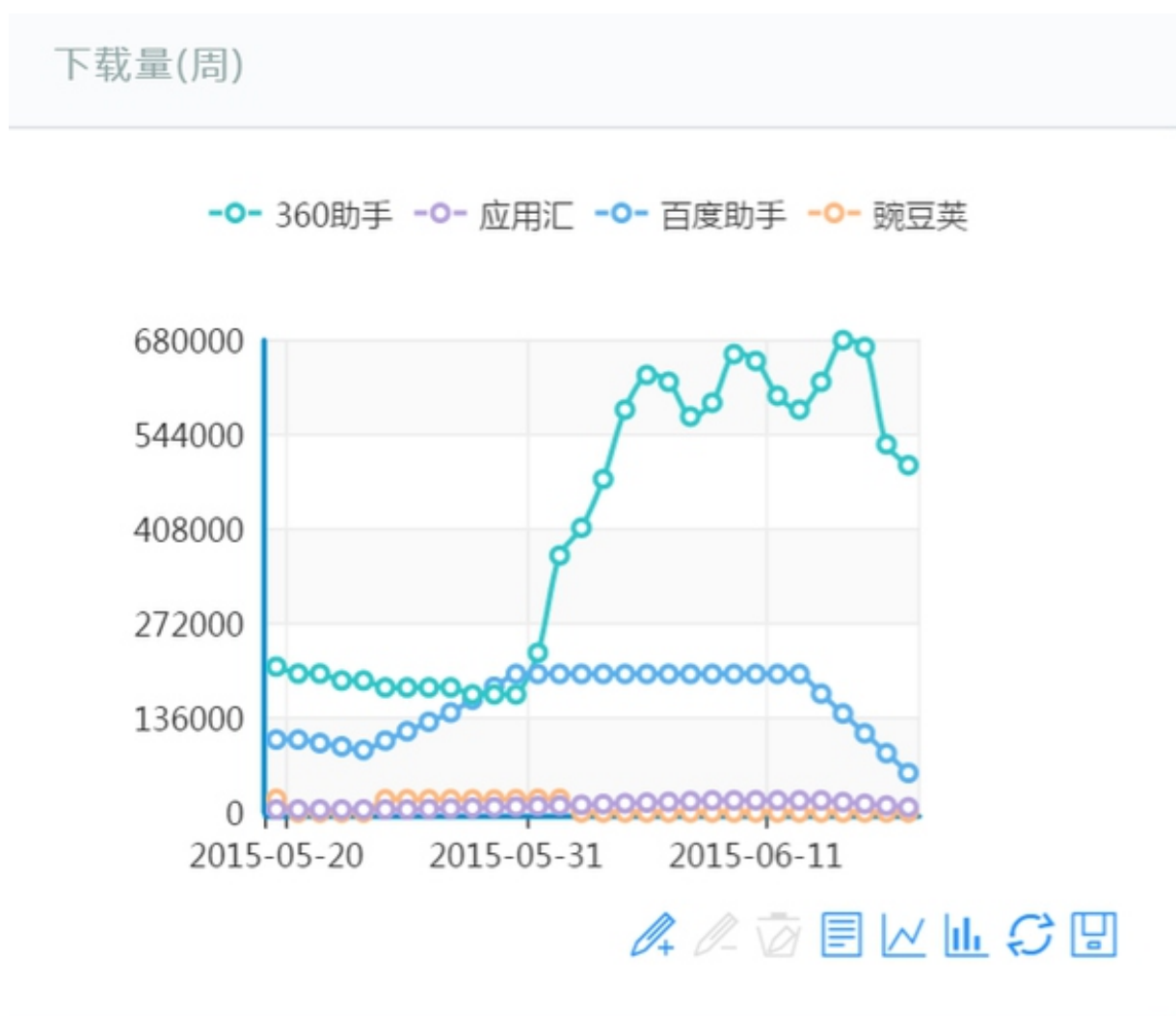
- 各应用商店：获取App的下载量及评论
- 大众点评及美团网：餐饮及各类线下门店消费及评价情况
- 汽车之家及易车：汽车的相关数据
- 58及搜房：房屋租售数据
- 新浪微博：用户的各种发言及舆论
- 财经数据：雪球及各类财经网站
- 宏观数据网站：天气、12306火车、机票网站

最初的产品纯粹是为基金服务。下图是在各个维度找出最有价值的App，各种量级范围内在30天/7天增长最快及评价最好榜单。（顺便吹一下牛，我们这个榜单很早就发现小红书App的快速增长趋势以及在年轻人中的极佳口碑）

增长最快 30天 十万级				
10				
名称	项目	公司	增幅	
养啦	完美宝贝	杭州贝安云科技有限公司	88915.5%	跟踪
恋恋520	unknown		62646.1%	
妈咪神器	unknown		46669.3%	
秒钱	unknown		35886.0%	
财神奇宝	unknown		22954.0%	
不格	unknown		22531.7%	
艺术猫	艺术猫_488322	None	19833.0%	跟踪
国金宝	unknown		19042.0%	
蚂蚁	unknown		18105.6%	

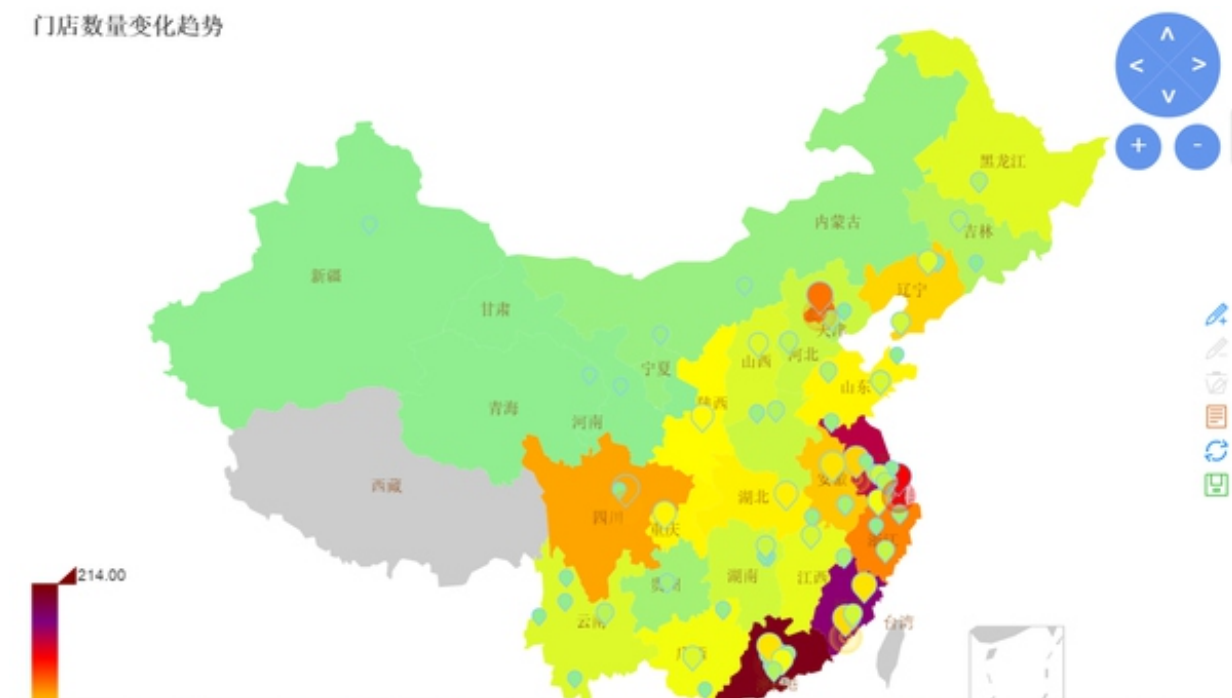
增长最快 7天 十万级				
10				
名称	项目	公司	增幅	
边逛边聊-最全购物时尚平台	unknown		7860.6%	
优享机场	unknown		6639.9%	
新富金融	unknown		3121.0%	
手机定位	unknown		2846.7%	
信融投资	unknown		2166.2%	
时尚圈	unknown		1771.0%	
亿百润	亿百润_488105	None	1720.4%	跟踪
散达决战(正版高	unknown		1235.6%	

下图是对某个App的下载量跟踪，帮着基金做尽职调查。



下图是某上市公司的门店变化情况，帮着基金跟踪TA的增长情况。

门店数量变化趋势



下图是国内各个机场的实时流量，帮着基金跟踪国内出行的实时情况，或许能从一个侧面反映经济是否正在走入下行通道。



## 第二步：扩展思路，开源和分享

为基金服务，虽然给钱爽快，但是也让方向越走越窄。首先，基金希望信息是独享的和封闭的，投资就是投资人之间的零和博弈，公开的信息就迅速会一钱不值，基金最在乎的就是信息的独享及提前量，所以各个基金都希望我们呈现的数据及分析结果能够独家。这样迅速让我们的方向收窄以及工作的趣味性降低，其次，毕竟对于基金而言，能分析的投资对象及方向是非常有限的。而且现阶段，大部分对冲基金里面的分析员的数据分析能力其实很弱：这些分析员里面能用VBA或者能在Excel里面使用矩阵及向量乘法的人几乎可以惊为天人；能写offset函数的人，就应该直接提拔了；大部分人停留在一个个数网页找数据的阶段。所以和他们起来十分费劲，

除了提供一些粗暴的数据，并不能产生太有价值的结果。

在这段迷茫期，本来充满激情的数据分析工作，让大家味如嚼蜡，感觉自己变成了一个外包公司。不过互联网大法好，做技术做互联网的核心思路是分享和开源，我们很快回归到这一点。并且这一点最终让我们做出了改变。有些分析虽然基金不买单，但是对一般的老百姓、对一般的媒体是有价值的，于是我们试着把这些数据分析及结果写出来，发布到知乎上供大家参考。

知乎是个好平台，坚持创作好内容迟早就会被发掘出来。很快一篇用数据分析黄焖鸡米饭为什么火遍全国的回答（黄焖鸡米饭是怎么火起来的？ - 何明科的回答）被知乎日报采用了。



知乎日报团队：何明科 您好，

您在「黄焖鸡米饭是怎么火起来的？」下的回答被「知乎日报」推荐啦。感谢您将精彩回答与千万知乎日报用户分享！

您可以打开知乎日报客户端查看该内容，或访问：[神州大地就这样淹没了黄焖鸡米饭的海洋之中（多图）](#)

为便于站外用户阅读，可能会对回答进行一些编辑工作，希望得到您的谅解。如果您觉得有任何不适的地方，请回复本私信联系我们（老版本移动端不支持回复管理员私信，需要升级到新版或者在电脑端回复。很抱歉给您带来不便）。

打扰了！

爱您的知乎日报团队

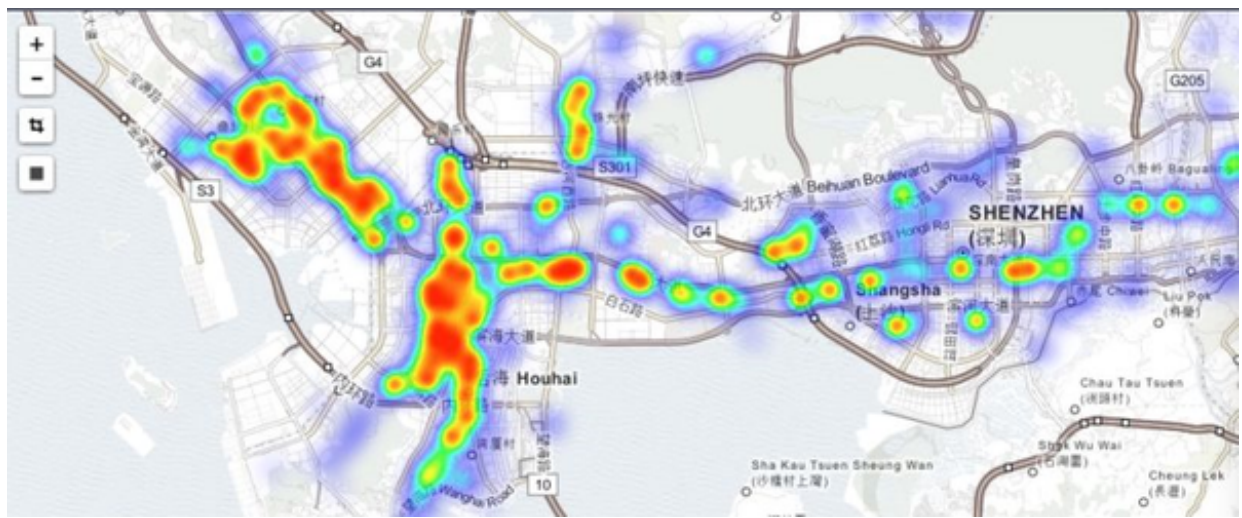
5月23日 15:00

[回复](#) | [删除](#)

这次被“宠幸”让团队兴奋不已，从而坚定了决心，彻底调整了整个思路，回到初心：不以解决基金关注的问题为核心，而以解决用户最关注的生活问题为核心。坚持以数据说话的套路，创作了许多点赞很多的文章并多次被知乎日报采用（所答内容被「知乎日报」选用是什么感觉？ - 何明科的回答），并专注在如下的领域：

- 汽车。比如：一年当中买车的最佳时间为何时？ - 何明科的回答，什么样的车可以被称为神车？ - 何明科的回答
- 餐饮。比如：[为什么麦当劳和肯德基都开始注重现磨咖啡的推广，其优势与星巴克等传统咖啡行业相比在哪里？ - 何明科的回答](#)
- 消费品。比如：口罩（[zhuanlan.zhihu.com/hemi...](#)），尿不湿（[zhuanlan.zhihu.com/hemi...](#)）
- 招聘。比如：[互联网人士年底怎么找工作（zhuanlan.zhihu.com/hemi...）](#)
- 房地产，这个虐心的行业。比如：[深圳的房地产走势（zhuanlan.zhihu.com/hemi...）](#)
- 投融资。比如：[用Python抓取投资条款的数据并做NLP以及数据分析：zhuanlan.zhihu.com/hemi...](#)

还共享了一些和屌丝青年生活最相关的分析及数据。下图是深圳市早晨高峰时段某类人群出行的热点图，通过热点分析，试图找出这类人群的居住和上班的聚集区。



下图反映了在各时间段在深圳科技园附近下车的人群密度。



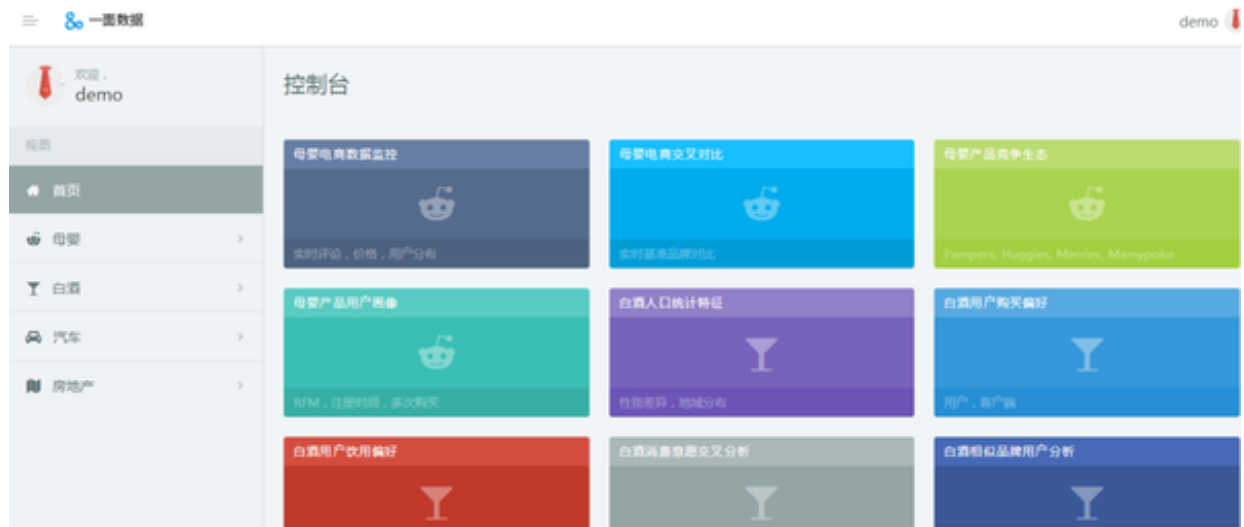
写这些报告，团队没有挣到一分钱，但是整个成就感和满意度大大上升。同时，在Python及各种技术上的积累也提高颇多，数据量级的积累也越发丰富，数据相关的各项技术也在不断加强。同时，顺势扩大了数据源：京东、淘宝等数据也纳入囊中。

### 第三步：扩展客户

在知乎上写这些报告，除了收获知名度，还收获意外之喜，一些知名品牌的消费品公司、汽车公司及互联网公司，主动找我们做一些数据抓取及分析。整个团队没有一个BD，也从来不请客户吃饭。

于是我们顺势做了如下的网站以及一个成熟的Dashboard框架（开发数据监控的Dashboard超有效率），目前主要监控和分析母婴、白酒、汽车及房地产四大行业，都是一些愿意花钱进行深度了解用户以及行业趋势的公司。收入自动上门，很开心！





下图是抓取汽车之家数据，做出BBA（奔驰宝马奥迪）这三大豪华品牌的交叉关注度，帮助品牌及4A公司了解他们用户的忠诚度以及品牌之间迁移的难度。

下图是抓取新浪微博的数据，分析广东白酒的消费场所

下图是抓取新浪微博的数据，分析广东白酒和各类食品的相关度。

除去为以上的品牌合作，我们数据风的文章也越来越受欢迎，曾经一周上了四次知乎日报（不知道 @周源@黄继新 两位大大是否考虑给我们颁发个知乎日报大奖）。另外也有越来越多的知名媒体及出版社找到我们，虽然告知他们我们不写软文而只坚持按照数据结果来发表文章，他们依然表示欢迎。原来非五毛独立立场的数据风也能被媒体喜欢。自此，我们不断成为易车首页经常推荐的专栏。

#### 第四步：尝试功能化平台化产品

降低与高大上基金的合作强度，转而与更接地气的各类品牌合作，让我们团队更贴近客户、更贴近真实需求。于是基于这些需求，我们开始尝试将之前在数据方面的积累给产品化，特别是能做出一些平台级的产品，于是我们开发出两款产品：

##### 第一款：选址应用

选址是现在许多公司头疼的难题，以前完全是拍脑袋。因此我们开发出这样一套工具，帮助公司能够更理性更多维度得选址。

下图，我们抓取多个数据源并完成拼接，根据用户的快递地址，勾画出某时尚品牌用户的住址，帮助其选址在北京开门店。



下图，我们抓取多个数据源并完成拼接，根据大型超市及便利店与某类型餐馆在广州地区的重合情况，帮助某饮料品牌选定最应该进入的零售店面。



## 第二款：数据可视化

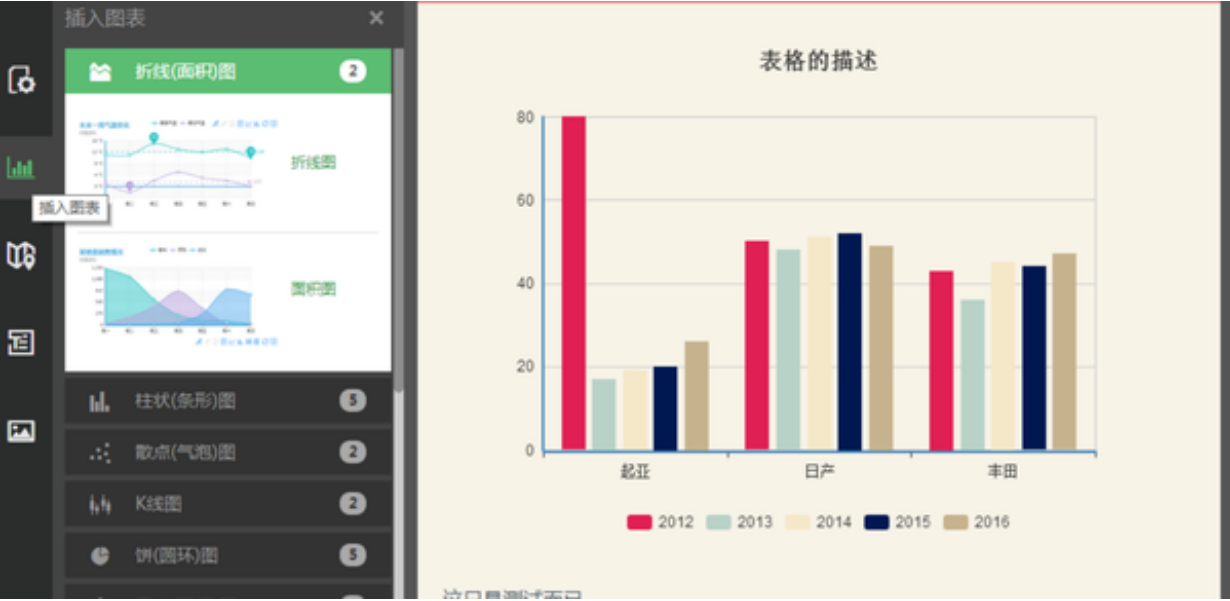
我们在工作中也深刻觉得以前制作图表和展示数据的方式太low、太繁琐，我们希望去改变这个现状，于是开发了一套基于Web来制作图表的工具文图。远有Excel/Powerpoint对标，近有Tableau对标。

下图是文图丰富的案例库及模板库。





下图是简单的使用界面及丰富的图表类型。



下一步的工作：

- 与微信的整合，一键生成适合于微信传播的截图以及公众号格式文章，便于在社交媒体的传播
- 收集更多数据，目前已经覆盖40多家网站，涵盖衣食住行等多个方面
- 将数据SaaS化和开源，便于各类公司及用户使用。（咨询投行等Professional Service人士一定会懂的，你们每年不知道要重复多少遍更新各类宏观微观的经济和行业数据，现在只需要调用KPI）

最后，希望有一天它能部分替代已经在江湖上混迹二三十年的PowerPoint及Excel。

第五步：.....

不可知的未来才是最有趣的。借用并篡改我们投资人的一句话：technology is fun, data is cool and science is sexy。初心未变，希望用数据用技术帮助更多的人生活得更美好。

编辑于 2016-01-22    20 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

诸葛清风，机器学习，信息检索

马骥、Ray Who、DavisH 等人赞同

有两个建议吧。

1. 完成《building machine learning systems with python》书上的所有projects，这本书除了封面其他里面的内容还是挺实用的。中文书名为《机器学习系统设计》

2. 完成kaggle playground和 101上的所有比赛，具体tutorial可以戳

- Getting Started With Python For Data Science [kaggle.com/wiki/Getting...](https://www.kaggle.com/wiki/Getting-Started-With-Python-For-Data-Science)
- Getting Started With Python II Getting Started with Pandas: Kaggle's Titanic Competition [kaggle.com/c/tita...](https://www.kaggle.com/c/titanic-getting-started)

1. 另外补充一个用scikitlearn构建文本挖掘系统的教程，个人觉得写的很好，基本上做一遍大概的流程就很清晰了：[scikit-learn文本挖掘系统学习（已完成）](#)

另外可以看这篇blog: [大数据竞赛平台——Kaggle 入门](#)

分割线补充：

我做过的比较好玩的应该是下载了豆瓣某一个爆照组的所有照片，然后结合发布者ID在其主页上找寻相关信息，然后按照地域进行统计分布，然后在google map上画了出来... 不过这个就没什么含金量了，现在在水推荐系统。

ps: 我也在入门中，欢迎一起探讨^\_^

编辑于 2015-03-27    17 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲  
20

王喆，蓝色光标集团，Senior Data Analyst，THU

cigarNdevil、玖號、tan chao 等人赞同



▼

巴西世界杯，为了找到一个高效的赌球方法，用python写了蒙特卡洛方法的赌球模拟实验，验证各种策略下的赌球盈利水平。

最终结果是在没有先验知识的情况下，无论何种赌球策略，在赌球次数足够多的情况下都不可能盈利。

不甘心，想通过各大博彩公司的博彩赔率差值来盈利，又用scrapy写了爬虫实时把各大博彩公司的即时赔率爬下来，一个简单的贪心就能求出利润的最大值，发现只要你能够在这些博彩公司开户，就完全有可能利用赔率的差值盈利！

项目地址：[wzhe06/soccerbet · GitHub](#)

关键是你没法开户啊。。国内参与博彩也是要被查水表的呀。

发布于 2015-03-28    3 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲  
160

grapeot，人生苦短 我用Python

d小z、山人周、Hp Dai 等人赞同



▼

用python在知乎上爬了400万答案，用numpy做数据分析，发现对高赞同答案贡献最大的话题是 脂溢性皮炎 维持关系 和 寝室神器。所以我机智地提了个问题：[想要和室友维持关系，有什么寝室神器可以治疗脂溢性皮炎？](#) - 数据挖掘 三个话题都占全一定超多人关注！！

然后问题立马被关闭了。妈蛋。

=====严肃的分割线=====

好吧换了个靠谱的算法，现在发现放在问题里最有利于吸引关注的知乎话题是：程序员，搞笑，平面设计，英语，和个人成长。而关注的人最多的话题是：电影，生活，音乐，互联网和创业。感觉二者并不交叉有点意料之外情理之中。。大家声称对什么感兴趣并不代表他们真的感兴趣。。

编辑于 2015-03-27    31 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲

46

王立武, 吉他手

若谷、顾琪、lisa wang 等人赞同

▼

用python把全校学生的证件照爬了下来, 加入自己的人脸数据库。通过人工标定初始的几个样本, 利用有监督的机器学习方法, 开发了一套颜值评分系统, 为全校所有人的颜值都打了分。

编辑于 2015-03-27    17 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲

265

薛昆Kelvin, uqer.io    优矿网 投资界uber

Spirit\_Dongdong、诺墨不科、靳彬昂 等人赞同

▼

还有什么比写个交易策略给自己赚钱更有意思呢? 推荐入门可以看

- 量化分析师的Python日记【第1天: 谁来给我讲讲Python? 】
- 量化分析师的Python日记【第2天: 再接着介绍一下Python呗】
- 量化分析师的Python日记【第3天: 一大波金融Library来袭之numpy篇】
- 量化分析师的Python日记【第4天: 一大波金融Library来袭之scipy篇】
- 量化分析师的Python日记【第5天: 数据处理的瑞士军刀pandas】

然后可以在社区克隆一个别人的策略研究一下, 对照《building machine learning systems with python》做一个自己的股市情感分析模型

发布于 2015-03-27    11 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲

4

何史提, 数据挖掘民工

李文静、付剑、邱国栋 等人赞同

▼

谢邀。

Document retrieval。我用Latent semantic indexing, 现在在想怎样用LDA（这方面知乎的高人很多, 不献丑了。）。

发布于 2015-03-24    2 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲

4

邵成

向儿回皮皮、张少敏、卢生 等人赞同

▼

去打kaggle吧

发布于 2015-03-23    4 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲

11

LIKE, Software Engineer

柳梦璃、钱生生、王魁 等人赞同

▼

看到有广泛报道都说有女性CEO在高管的公司一般比没有的公司表现要好, 而且中国女性高管领先亚洲: 今年3月8日国际妇女节前夕, 国际招聘专家瀚纳仕在其《2015年瀚纳仕亚洲薪酬指南》中指出, 尽管情况有所改善, 但性别多元性仍是企业面对的一个严峻问题。在职场性别多元性方面, 中国仍领先亚洲。中国担任高级管理职位的女性占比要高于亚洲其他国家。纵观整个亚洲, 瀚纳仕发现, 在中国36%的管理职位由女性担任。该数字与去年同期相比没有变化, 与29%的亚洲平均水平相比占据明显优势。紧随中国的分别是马来西亚（34%, 较去年的29%有所上升）、中国香港（31%, 较去年的33%有所下降）以及新加坡（27%, 与去年同期持平）。

一般这类问题都会引导人们去思考：如果我们去投资这些女性作为核心高管主导的公司，我们会获得更高的收益吗？

第一个挑战是找到合适的数据源。最理想的是我可以找到一个时不时可以保持更新的数据源，最后我们找到了wind并且写了一个爬虫爬下来了所有公司的高管以及其性别。从中我们挑选出来了一部分有代表性的是以女性为一把手（具体职位是法定代表人+总经理/总裁）的公司的股票：

股票代码	major	姓名	职位	性别	出生年份	教育背景
002464.XSHE	0	方幼玲	总经理 法定代表人	女	1959年	硕士
300405.XSHE	0	姜艳	总经理 法定代表人	女	1961年	本科
002599.XSHE	0	栗延秋	总经理 法定代表人	女	1969年	中职
601799.XSHG	0	周晓萍	总经理 法定代表人	女	1961年	硕士
002592.XSHE	0	顾瑜	总经理 法定代表人	女	1954年	专科
300249.XSHE	0	张菀	总经理 法定代表人	女	1962年	硕士
002537.XSHE	0	刘国平	总经理 法定代表人	女	1963年	硕士
603519.XSHG	0	卢凤仙	总经理 法定代表人	女	1956年	高中
600615.XSHG	0	涂建敏	总经理 法定代表人	女	1967年	专科
002227.XSHE	0	廖晓霞	总经理 法定代表人	女	1961年	本科
300304.XSHE	0	付红玲	总经理 法定代表人	女	1969年	专科
002256.XSHE	0	汤薇东	总经理 法定代表人	女	1968年	本科
300087.XSHE	0	张琴	总经理 法定代表人	女	1963年	硕士
600246.XSHG	0	李虹	首席执行官 法定代表人	女	1958年	硕士
600323.XSHG	0	金铎	总经理 法定代表人	女	1966年	硕士
000705.XSHE	0	宋逸婷	总经理 法定代表人	女	1957年	本科
002439.XSHE	0	王佳	总经理 法定代表人	女	1969年	博士
000790.XSHE	0	周蕴瑾	总裁 法定代表人	女	1974年	本科
000637.XSHE	0	刘华	总经理 法定代表人	女	1961年	硕士
002697.XSHE	0	曹世如	总经理 法定代表人	女	1952年	中职
600252.XSHG	0	许淑清	总经理 法定代表人 代董事会秘书	女	1958年	本科
600220.XSHG	0	陈丽芬	总经理 法定代表人	女	1959年	本科
000651.XSHE	0	董明珠	总裁 法定代表人	女	1954年	硕士

说明：0为上市公司上报自己职位的最高顺位，也是筛选的依据之一。XSHG为沪市股票，XSHE为深市股票。

很遗憾的是我们无法找到每一位女性总经理的入职时间，否则可以继续分析女性CEO的每年增加情况。

下一步是什么？不如我们开始分析一下单独一家上述公司的情况，这样子也可以更简单的开始研究并且让我们对数据有一个初步的了解。我选取了非常出名的董明珠女士的格力电器，以格力的股价为例，并且我们从网上爬取了关于董明珠女士的新闻加在一起，使用Ricequant的IPython Notebook研究平台，用几行Python pandas代码+matplotlib出来的结果：



说明：Ricequant的数据最早始于2005年，晚于董明珠最早上任业务一把手的时间。

亲们有空可以看看董女士的故事，挺励志的。她带领格力不断走向新的高度，收获无数赞誉和肯定。

之后我们将这个列表中的股票代码整合成一个列表传入最新上线的ricequant python SDK中，写了一个简单的python交易策略进行回测。

买入的具体逻辑：**2010年**基本全体同权重买入构建投资组合，还没上市的股票，在IPO当天任性买入（如果不涨停的话），因为策略默认该**21只股票**大名单（不是踢足球的国家队那个）是值得考验的股票。**2010年**这个年份是综合考虑高管任期时长和一把手名单大部分上市公司的上市时间的结果。

在[www.ricequant.com](http://www.ricequant.com) 上在线策略回测截图如下：

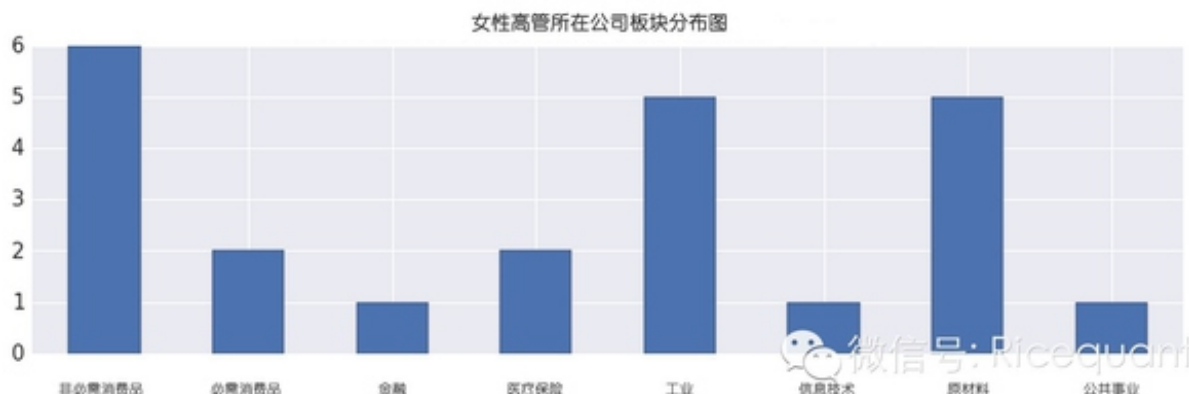


回测结果分析：回测的结果貌似一直压着基准收益诶。。收益方面一直压制基准策略，表明这比长期买入沪深300机智得多。当然，由于是买入长期持有，自然是会有一个很高的Beta，最大回撤（MaxDrawDown)也是挺大的。



在整理的过程中，印象最深的是女性高管已经有1000+人，应该能占到10%，而且从05年(米筐最早的回测年份)开始，女性进入高管的人数也有增长，可惜的是，由于暂时没有整理完整的女性任职高管的数据(比如具体的任职日期)，所以并不能完全清晰地展现这个趋势。

也好奇到底这份投资组合的女性高管的公司都是哪些板块的，以造成比较强悍的表现，依然使用pandas和matplotlib画图：



由图可知，消费者（6+2）版块表现抢眼，工业、原材料也表现不俗，都为5人。

在目前这个系列的探讨中，我们已经获得了基本完整的A股上市公司高管名单数据，包括年龄、学历、公司职位、性别，也欢迎对社会学及女性主义有兴趣的童鞋私信一起来研究。我们只是做了一个最初步的研究，还是有很多地方可以更加改进和讨论。

---- 硬广时间开始 ----

And Welcome to [www.ricequant.com](http://www.ricequant.com) ，拥抱云端的量化分析和交易。

---- 硬广时间结束 ----

发布于 2015-09-12    1 条评论    感谢    分享    收藏 • 没有帮助 • 举报 • 禁止转载

据数，顺其自然！ 微信：13475153511

早川橘下、凌风、刘志强 等人赞同

转载请联系我微信！

楼下答案里埋伏着各种大神！但是楼主既然是刚开始学习，我觉得还是接触一些相对简单的实践案例最好，同样是新手所以把最近写的一篇发布在微信公号的文章贴出来，希望对你有帮助！大神们一定轻拍：)

#先预热下#

许许多多的人都非常容易爱上Python这门语言。自从1991年诞生以来，Python现在已经成为最受欢迎的动态编程语言之一，尤其进入21世纪以来，Python在行业应用和学术研究中进行科学计算的势头也越来越迅猛。

——《Python for Data Analysis》（Wes Mckinney）

Python不仅在编程方面有强大的实力，而且由于不断改进的第三方库，Python在数据处理方面也越来越突出；近年来，非常火爆的机器学习(Machine Learning)以及前沿的自然语言处理(Natural Language Processing )也选择Python作为基础工具。所以要想在数据科学领域有所进步的话，了解学习Python看来还是有所必要的。本文通过简单案例，分享Python在数据处理方面的实际应用，属于基础学习范畴，希望刚刚接触Python学习的新手们能通过应用去实际问题从而巩固掌握Python操作，在这里与大家相互学习，也希望大神们轻拍：)

Without further ado, let's get started!

### #进入正题#

本文使用Python2.7版本，操作在集成开发环境Spyder中进行；选择的数据集，是大名鼎鼎的鸢尾花数据集iris.csv，数据集网上公开请自行下载！

1. 数据集截图如下图1：

该数据集包含数据有150行\*5列。前4列分别是：花萼的长度、宽度，花瓣的长度、宽度；最后一列是花的分类，总共分3类。

图1.iris数据集截图

2. 读入数据，代码如下：

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import * #机器学习库
np.random.seed(123) #设置随机数种子

iris=pd.read_csv("C:\\Users\\Administrator\\Desktop\\iris.csv",header=False)
#操作请注意:输入文件实际路径
print iris.shape #输出数据维度
print iris.head() #查看前5行
```

输出结果如下：

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import * #机器学习库
np.random.seed(123) #设置随机数种子

iris=pd.read_csv("C:\\Users\\Administrator\\Desktop\\iris.csv",header=False)
#操作请注意:输入文件实际路径
print iris.shape #输出数据维度
print iris.head() #查看前5行
(150, 5)
   SepalLength  SepalWidth  PetalLength  PetalWidth      Name
```



```

train_data,test_data,train_target,test_target=cross_validation.train_test_split(data,
                                         target,test_size=0.24,random_state=0) #分成训练集、测试集（占0.24）
clf=tree.DecisionTreeClassifier(criterion='gini', max_depth=6,
                               min_samples_split=5) #CART算法
clf_fit=clf.fit(train_data, train_target) #开始fit
#print clf_fit
train_est=clf.predict(train_data) #预测训练集
test_est=clf.predict(test_data) #预测测试集
sum=0
for i in range(36):
    if test_est[i]==test_target[i]:
        sum=sum+1
print 'test_accuracy=', "%.2f%%"%(sum*1.0/36*100) #测试集预测正确率

sum=0
for i in range(114):
    if train_est[i]==train_target[i]:
        sum=sum+1
print 'tarin_accuracy=', "%.2f%%"%(sum*1.0/114*100) #训练集预测正确率

```

输出结果：

```

target=iris["Name"] #目标变量
data=iris.ix[:,1:4] #自变量
train_data,test_data,train_target,test_target=cross_validation.train_test_split(data,
                                         target,test_size=0.24,random_state=0) #分成训练集、测试集（占0.24）
clf=tree.DecisionTreeClassifier(criterion='gini', max_depth=6,
                               min_samples_split=5) #CART算法
clf_fit=clf.fit(train_data, train_target) #开始fit
#print clf_fit
train_est=clf.predict(train_data) #预测训练集
test_est=clf.predict(test_data) #预测测试集
sum=0
for i in range(36):
    if test_est[i]==test_target[i]:
        sum=sum+1

print 'test_accuracy=', "%.2f%%"%(sum*1.0/36*100) #测试集预测正确率

sum=0
for i in range(114):
    if train_est[i]==train_target[i]:
        sum=sum+1

print 'tarin_accuracy=', "%.2f%%"%(sum*1.0/114*100) #训练集预测正确率

test_accuracy= 97.22%
tarin_accuracy= 98.25%

```

利用CART算法对iris数据集建立模型，并预测结果，同时输出训练集测试集的预测正确率。相关说明及代码含义均在代码中已注释。

最后：以上仅为小例子，实际问题要比这个复杂的多：数据清洗、模型选择、调参等等！因为网上教程给出

完整数据、代码及结果的资料不多，所以把自己的浅薄经验分享给大家希望对你有帮助！

码字也挺不容易的，求各位点个赞吧=￣ω￣=

END

编辑于 2015-11-11    5 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 禁止转载

▲ 何宜晖，初学Python

9

扒哥、齐思、玖玖 等人赞同



▼ 占坑

学习machine learning in action 与 programming collective intelligence 中，这两本都是用python写的，有些有趣的小项目，等我搞点什么回来答~

发布于 2015-05-19    添加评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲ 匿名用户

7

玖號、设置、coyg road 等人赞同

▼ 在自学练习《building machine learning systems with python》里面的项目时，妹子说我认真的样子好帅，然后接下来的事情都很有趣。

发布于 2015-03-27    2 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲ 裴小浩，关注推荐、广告算法

7

成三驾、zhaoqing PENG、李志文 等人赞同



▼ 先占个楼，东西还没有完成，半个月后更新，20150327。

-----  
2015-04-12更新

更新的时间比预计的时间晚了两天。

（1）网页抓取

工作以来，我一直希望工作中可以有搞爬虫的机会，奈何一直没有等到。有一天突然在知乎上看到一个问题，说如何找到知乎中简洁精辟的回复，感觉很有意思，于是就自己做了一下。

通过scrapy抓取了一下知乎的精华问题的每一个答案，然后通过简单的策略过滤出来了一些精简的答案。

网页查看抓取结果：[戳我查看抓取结果](#)

（2）app开发

抓取了之后，发现android平台有人专门针对知乎的精短回复做了一个app。由于我刚好最近在自学iOS开发，所以就顺便做了一个app，还没有放到AppStore，先放张图上来。





## 你最孤独的时刻是什么？

4811 赞 毕业离校的时候，游戏里掉了个好东西，扭头炫耀的时候才发现整个寝室就剩下了自己，当时的感觉真像挨了一记闷拳啊

## 为什么有些事对别人来说只是举手之劳可他们却不愿帮忙？

4777 赞 我只是不喜欢别人自己能完成的举手之劳，却叫我来代劳。

## 如何评价「阿里上市是中国的悲哀，因为阿里最大股东是日本软银」的论调？

4745 赞 这种文章稍微改造下就可以改造成《马云：用美国日本人的钱养活中国人的公司——华尔街惊呼上当了，不转不是中国人》

## 在书店如何跟女生搭讪？

4695 赞 你就说：……美女，喜欢书吗？美女说：喜欢…… 你就说：叔也喜欢你～

## 猫的智商有多高？

4626 每次我看完AV以后，它都会这样妩媚地

编辑于 2015-04-12    3 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲  
19

张铁匠，数据挖掘爱好者，数学渣

韩三皮、高权奥、匠艺成 等人赞同



—————答案更新，我把原先比赛做pre的视频上传了欢迎观摩拍砖！—————

b站视频传送门：[自制机器学习系统-MMpython演示](#)

—————以下是原答案—————

仅仅就题主的题目，说一说我自己的一个故事：

话说大二下那年，怀着对ML和DM的神往开始学习机器学习和数据挖掘，我身为一个数学渣敏锐得觉察到吴恩达大牛的公开课对我是一个坑，还是个神坑，转而投入《机器学习实战》这一类不太需要过硬数学基础的机器学习和数据挖掘书籍中，于是开始学习Python+Matplotlib+Numpy了。

大三上适逢学校申报大学生创新项目，抱着当炮灰的心态交了一个关于强化学习算法验证的申报书，没想到过了，开始做项目的时候，指导老师对我说“你这个，要用matlab啊，不然就用Octave呀，会用不？？”然后我就

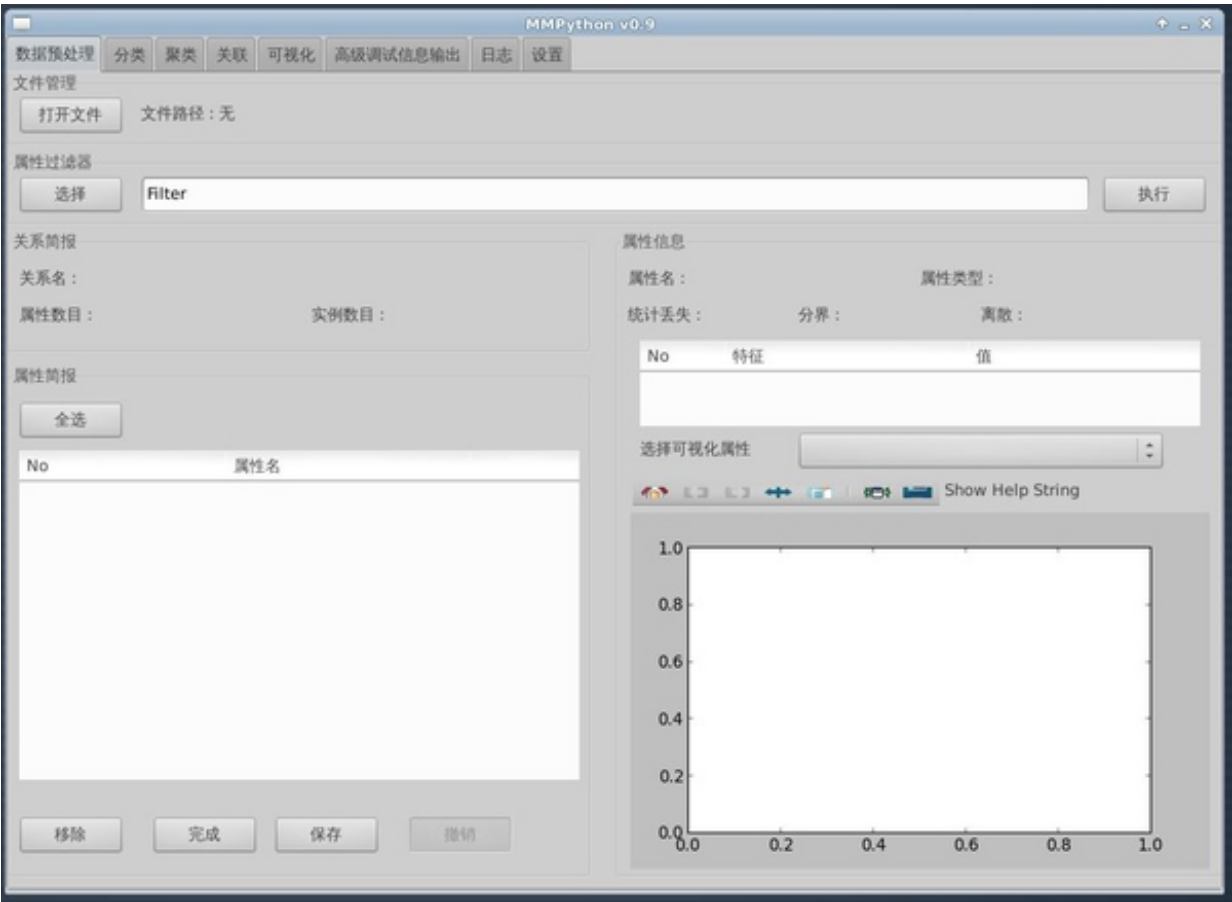
老老实实回去查matlab和Octave的资料。。。。。。不出所料，几个通信的同学都说matlab是个神坑，迅速浏览了官网和文档之后，我也觉得是神坑，遂弃而不学（其实是太懒了）。

然后又学长给我说可以用Weka来做这个项目的实验，我接着去google weka是个神马东西，一看尼玛是用Java写的，遂逃。。。（我也是醉了）。

为什么不用正在学的Matplotlib+Numpy呢？因为我当时觉得Matplotlib+Numpy没有一个成熟的可用的现成系统给我用，每一次都要我自己写脚本，尼玛好麻烦。

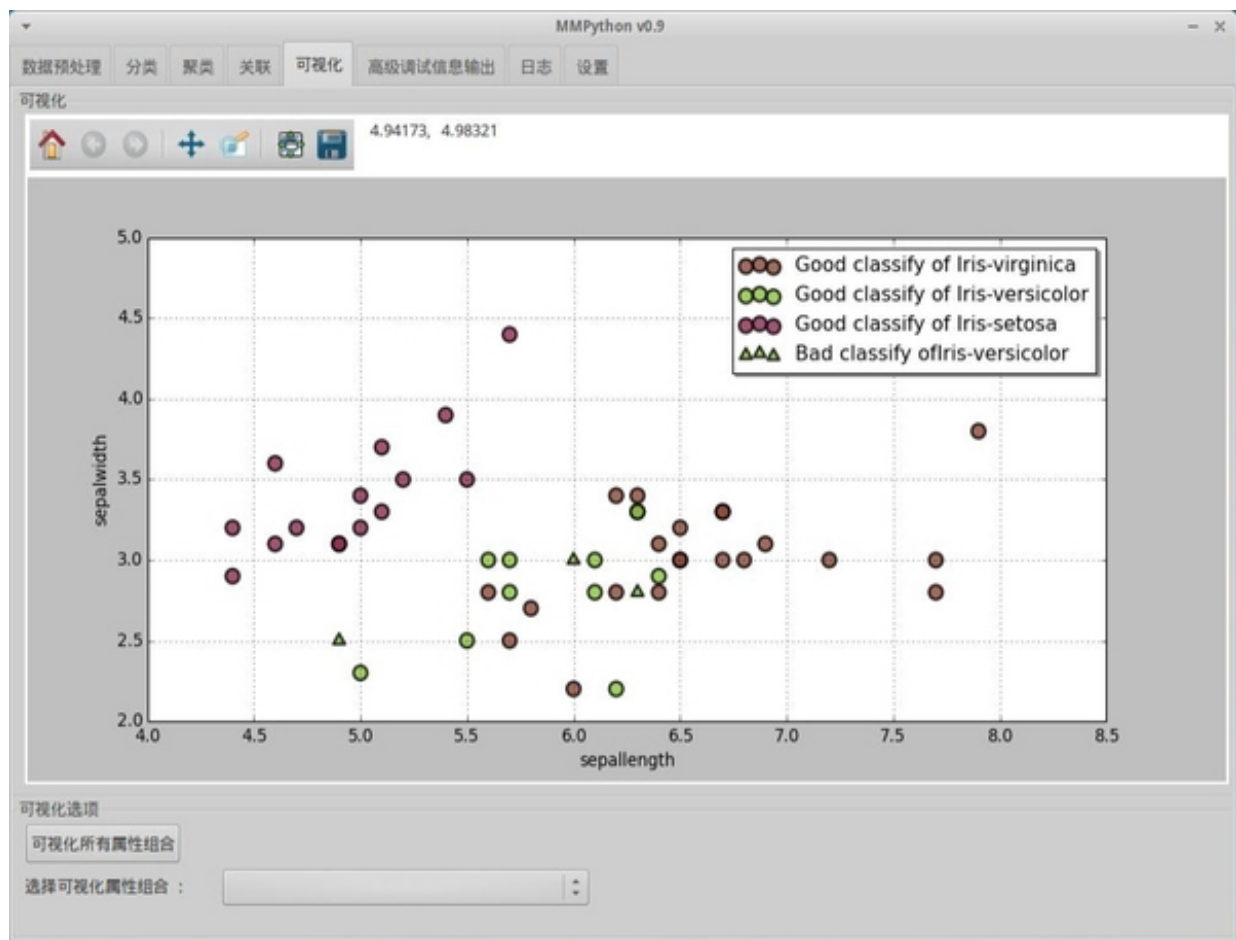
于是我就坑了指导老师有半年时间，到了中期检查的时候我觉得deadline快要到了，感觉慌得裤子都着火了，一气之下就想，“既然matlab是个坑，老子就写一个缩小版的matlab自己用就够咯”  
前期的懒惰直接给自己开了一个新坑，那一个月的目标就是用Python写一个有图形界面的数据挖掘工具出来。

一个月之后，下图这个什么鬼就诞生了：



当然，最初版本的界面没有这么好看，这个界面完全是致敬（jie）敬（jian）weka做的，我没有看过weka的源码，也不知道它的实现逻辑。然后经过很多次的修改和重写，这个工具可以分类，聚类，关联，处理的数据类型主要是数值型，当然可以通过导入与处理函数把字符型化归为数值型，这个工具可以导入算法，比如我给你一个分类算法统一遵循的“参数表范式”，只要你按照参数规则用Python写你的分类算法，这个工具就可以识别出来。

忘了说，这个工具用Numpy实现数学运算（矩阵运算），用Matplotlib实现可视化，可视化就是下面这个样子：



没错，上面这个图就是对UCI IRIS数据集的分类结果，算法用的是朴素的KNN。

然后到了今年，这个项目已经结题了，因为这个工具越写越大越写越丰富，到最后倒成了这个项目的主要成果。

这个工具主要是用：

Numpy做矩阵运算输出

matplotlib做绘图

wxPython写的界面

今年六月份被学校征用去参加挑战杯省赛了，作为一个即将毕业成为校友的我明明就是一个去交流经验的，没想到还得了一等奖进入了国赛审查阶段。本来无心插柳的一件事，没想到后来还会得到一些不错的发展。

总的说来，其实生活中处处都有ML和DM的用武之地，比如说我在挑战杯评审的时候给评委演示的就是“通过近十年中国男足比赛情况看中国男足属于世界几流球队？”这种激（mei）动（you）人（jie）心（cao）的问题。在学校里呢我没事就帮生科的几个孩纸录数据。

学习scipy和numpy这些库并不代表你以后只能用它写脚本来做数据科学的任务了，C/Java/php这些坑里面都有可以调用和接入Python的门，所以不仅仅是局限于Python，其他的工具和技术也可以和Python结合使得ML和DM更加得心应手。

以上，如有纰漏请各位大牛指正，希望对题主有帮助。

编辑于 2016-01-20    5 条评论    感谢    分享    收藏 · 没有帮助 · 举报 · 作者保留权利

▲  
0

▼

1oscar，一个纯粹的人，唯爱技术与动漫

11



编辑于 2016-01-22    添加评论    感谢    分享    收藏 • 没有帮助 • 举报 • 作者保留权利

▲  
2

▼

孙行者，罗马之所以为罗马，全凭风雨！

王琦、青仙 赞同



先占个楼。正在学习nodejs。准备学习完了nodejs，再学python。当然，紧接着就是人工智能.....，也就是，有趣的数据挖掘/分析项目。


编辑于 2015-04-05    添加评论    感谢    分享    收藏 • 没有帮助 • 举报 • 作者保留权利

▲  
1

▼

星逍L，逢人不说人间事，便是世间无事人。

猜花星 赞同



把学校开的所有课程爬了一遍，筛选出两分制(通过or not)不算GPA的水课。。似乎算不上数据挖掘...

编辑于 2015-09-24    1 条评论    感谢    分享    收藏 • 没有帮助 • 举报 • 作者保留权利

