来源： Alfred   链接：

http://mp.weixin.qq.com/s/mW7PTofuCOQrW5e34Ei2Pw

最近几天干啥都不来劲，昨晚偶然了解到Python里的itchat包，它已经完成了webchat的个人账号API接口，使爬取个人微信信息更加方便。鉴于自己很早之前就想知道诸如自己微信好友性别比例都来自哪个城市之类的问题，于是乎玩心一起，打算爬一下自己的微信。

首先，在终端安装一下itchat包。pip install itchat

安装完成后导入包，再登陆自己的微信。过程中会生产一个登陆二维码，扫码之后即可登陆。登陆成功后，把自己好友的相关信息爬下来。

import itchat  
itchat.login()  
#爬取自己好友相关信息， 返回一个json文件  
friends = itchat.get\_friends(update=True)[0:]

有了上面的friends数据，我们就可以来做分析啦。

**1. 自己微信好友的男女比例**

仔细观察了一下返回的数据结构，发现”性别“是存放在一个字典里面的，key是”Sex“，男性值为1，女性为2，其他是不明性别的（就是没有填的）。可以写个循环获取想要的性别数据，得到自己微信好友的性别比例。

#初始化计数器  
male = female = other = 0  
#friends[0]是自己的信息，所以要从friends[1]开始  
for i in friends[1:]:  
    sex = i["Sex"]  
    if sex == 1:  
        male += 1  
    elif sex == 2:  
        female += 1  
    else:  
        other +=1  
#计算朋友总数  
total = len(friends[1:])  
#打印出自己的好友性别比例  
print("男性好友： %.2f%%" % (float(male)/total\*100) + "\n" +  
"女性好友： %.2f%%" % (float(female) / total \* 100) + "\n" +  
"不明性别好友： %.2f%%" % (float(other) / total \* 100))

Getting uuid of QR code.

Downloading QR code.

Please scan the QR code to log in.

Please press confirm on your phone.

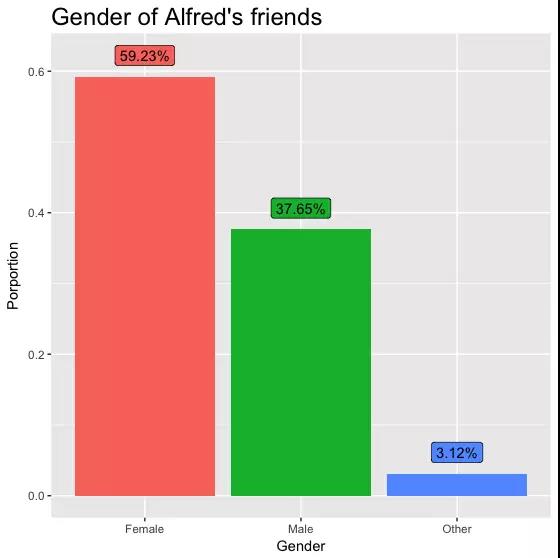
Loading the contact, this may take a little while.

Login successfully as 夏敏捷

男性好友： 54.96%

女性好友： 29.01%

不明性别好友： 16.03%



**2. 自己微信好友的城市分布**

再仔细观察friends列表，发现里面还包含了好友昵称、省份、城市、个人简介等等的数据，刚好可以用来分析好友城市分布，最好的方式是定义一个函数把数据都爬下来，存到数据框里，再进行分析。

#定义一个函数，用来爬取各个变量  
def get\_var(var):  
    variable = []  
    for i in friends:  
        value = i[var]  
        variable.append(value)  
    return variable  
#调用函数得到各变量，并把数据存到csv文件中，保存到桌面  
NickName = get\_var("NickName")  
Sex = get\_var('Sex')  
Province = get\_var('Province')  
City = get\_var('City')  
Signature = get\_var('Signature')  
from pandas import DataFrame  
data = {'NickName': NickName, 'Sex': Sex, 'Province': Province,  
        'City': City, 'Signature': Signature}  
frame = DataFrame(data)  
frame.to\_csv('data.csv', index=True)

以上便得到一个叫data的csv桌面文件， 用R打开并简单做一下数据预处理，得到如下(涉及隐私的已被预处理）：

        City NickName Province Sex                                    Signature

1              Alfred           男 一个专注挖矿的，煎得一手好牛排#stay focused.

2     马鞍山     月\*     安徽  女            你太虚伪了

3 花地玛堂区     M\*\*\*     澳门  女      双子，天秤，白羊

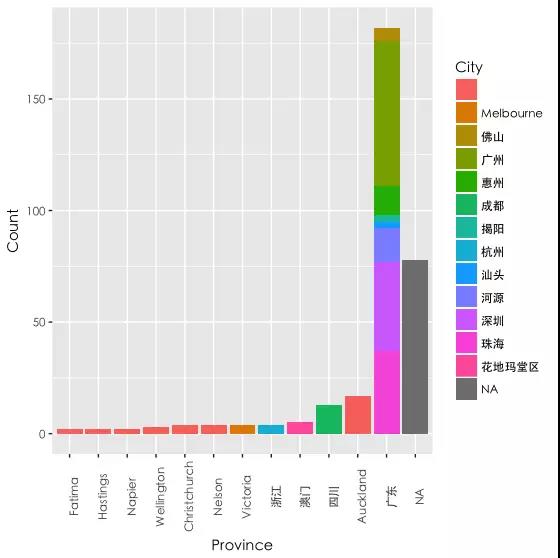
4             嘉\*\*\*           女

5       广州     西\*     广东  女

6       珠海 浪\*\*\*\*     广东  男             满目青山空念远，不如惜取眼前人。

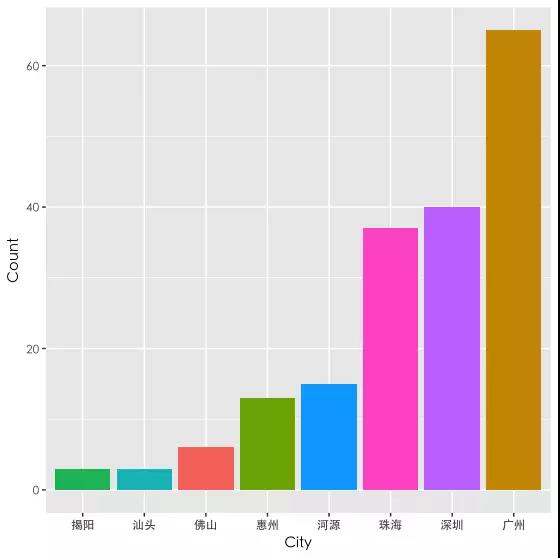
# ... with \*\*\* more rows

接着先根据省份、城市进行数据的分组和聚合，选择排名前二十的，利用ggplot2包画出如下的城市分布图：



看来我大部分的朋友都是在广东的，其中广东的朋友大部分集中在广、深、珠，第二名是在奥克兰， 接着是四川、澳门等。灰色的NA值是指没有设置自己所在地的朋友，一共有70多人。另外，在国外的朋友由于微信的设置问题（很多是直接跳过省份，只有城市可以选择的），很多国外的城市被误当成了省份。

再来一张图看看自己微信朋友在广东的具体分布（取前八）：



**3. 自己微信好友个性签名的自定义词云图**

好玩的来了。之前已经爬下了每个好友的个性签名，刚好可以分析一下大伙儿个性签名时使用的高频词语是什么，顺便可以做个词云图（云图制作方法：[逼格满满的文字云快速制作！）](http://mp.weixin.qq.com/s?__biz=MzI4MTQ2NjU5NA==&mid=2247486136&idx=1&sn=916084ea89fc2f062cc34932e0eb18b2&chksm=eba98935dcde00238fb4821625117deda3c3cdbb2ee28be5449c30de4444c5a0b63ac2f81b05&scene=21" \l "wechat_redirect" \t "_blank)

先把原先爬下来的个性签名（Signature）打印出来，发现有很多本来是表情的，变成了emoji、span、class等等这些无关紧要的词，需要先替换掉，另外，还有类似<>/= 之类的符号，也需要写个简单的正则替换掉，再把所有拼起来，得到text字串。

import re  
siglist = []  
for i in friends:  
    signature = i["Signature"].strip().replace("span","").replace("class","").replace("emoji","")  
    rep = re.compile("1f\d+\w\*|[<>/=]")  
    signature = rep.sub("", signature)  
    siglist.append(signature)  
text = "".join(siglist)

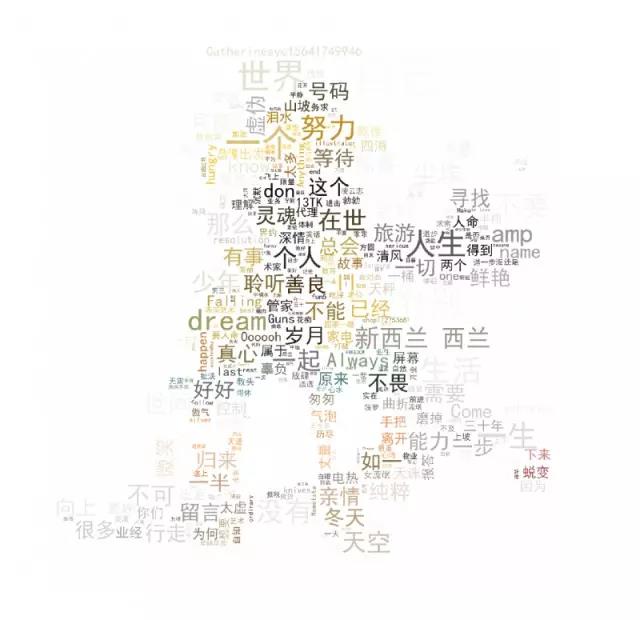
接着就可以把结巴分词这个包搞进来分词。

import jieba  
wordlist = jieba.cut(text, cut\_all=True)  
word\_space\_split = " ".join(wordlist)

终于可以进入画图阶段了。可以根据自己想要的图片、形状、颜色画出相似的图形（在这里，我使用的是我的头像，当然，为了颜色可以更加鲜艳使最后画出的词云图更加好看易辨，我先对自己的头像用PS做了一点小处理）。为此，我们需要把matplotlib、wordcloud、numpy、PIL等包搞进来。

import matplotlib.pyplot as plt  
from wordcloud import WordCloud, ImageColorGenerator  
import numpy as np  
import PIL.Image as Image  
coloring = np.array(Image.open("/Users/apple/Desktop/wechat.jpg"))  
my\_wordcloud = WordCloud(background\_color="white", max\_words=2000,  
                    mask=coloring, max\_font\_size=60, random\_state=42, scale=2,  
                    font\_path="/Library/Fonts/Microsoft/SimHei.ttf").generate(word\_space\_split)  
image\_colors = ImageColorGenerator(coloring)  
plt.imshow(my\_wordcloud.recolor(color\_func=image\_colors))  
plt.imshow(my\_wordcloud)  
plt.axis("off")  
plt.show()

运行上面的代码，得到下面的图：



看来，在我的微信好友的个人签名里，有人善于聆听，有人强调善良，有人重视灵魂，有人凝视人生。一直在等待，不放弃寻找。历尽曲折，不畏虚伪。真心如一，不能辜负。一步一生多努力，一起一切one more dream（真编不下去了）。

**4. 最后**

以上是不是很有趣呢？是不是又打开了一扇新世界的大门呢？

我微信好友也不是太多，如果微信好友有几千个，可以得到几千条数据，分析一下还是很有价值的。

当然，itchat包还有很多其他的功能还有待发掘，包括自动回复微信信息、自动添加好友、管理微信群等，有时间再慢慢摸索吧。