

# 说明文档

该过程大致分两步进行，第一步爬取数据，第二步进行数据处理。

## 一、爬取数据：

根据给出的 url 获取网页源码，以 ' \n' 分开。

```
url='http://yang.lzu.edu.cn/data/index.txt'
r=requests.get(url)
s=str(r.text)
r0=s.split('\n')
r0
['./accelerometer/anxiety/female',
 './accelerometer/anxiety/female/20191107191800_308_accelerometer.json',
 '\n',
```




再将以某些固定前缀和后缀的 json 文件的数字编号记录下来。

```
for line in r0:
    if line.startswith('./accelerometer/anxiety/female') and line.endswith('.json'):
        a = re.findall(r"female/(.*)_accelerometer", line)
        aaf.append(str(a))
```

最后将这些文件写入本地。

```
for i in aaf:
    url1='http://yang.lzu.edu.cn/data/accelerometer/anxiety/female/'+i.strip()[2:-2]+'_accelerometer.json'
    r1=requests.get(url1)
    fpl=open(r'D:\data\accelerometer\anxiety\female\{}_accelerometer.json'.format(i.strip()[2:-2]),mode='w')
    fpl.write(r1.text)
    fpl.close()
```

检查文件存在。

OVERY (D:) > data > accelerometer > anxiety > female			搜索"fer"
名称	修改日期	类型	
 20191107191800_308_acceleromete...	2020/4/11 23:11	JSON File	
 20191108111045_558_acceleromete...	2020/4/11 23:11	JSON File	
 20191108162812_694_acceleromete...	2020/4/11 23:11	JSON File	

## 二、数据预处理：

已知这些数据是参与者回答问卷时通过手机传感器记录的数值，有两个加速度 accelerometer 和 device\_motion，还有一个螺旋仪 gyroscope, 且记录频率为 5Hz.

首先观察下载好的数据，有些数据文件大小为 1kb, 打开文件发现内部只有中括号，需要删除大小小于 1kb 的文件。

```
import os

def get_path(file_path):
    for root, dirs, files in os.walk(file_path):
        for file in files:
            filename = os.path.join(root, file)
            size = os.path.getsize(filename)
            if size <= 1024:
                os.remove(filename)
                print("remove", filename)

if __name__ == "__main__":
    file_path = 'D:\data'
    get_path(file_path)
```

```
remove D:\data\gyroscope\anxiety\female\20191114161025_2943_gyroscope.json
remove D:\data\gyroscope\anxiety\female\20191114171226_3006_gyroscope.json
remove D:\data\gyroscope\anxiety\female\20191116154726_3651_gyroscope.json
remove D:\data\gyroscope\anxiety\female\20191118173012_3937_gyroscope.json
remove D:\data\gyroscope\health\female\20191111111559_1665_gyroscope.json
```

其次，记录的频率为 5Hz，即每秒 5 次，将获取到的文件的行数除以 5 再除以 60 就可以得到单位为分钟的回答问卷的时间。

```
import os
import json
import pandas as pd

def get_f(path):
    for root, dirs, files in os.walk(path):
        for file in files:
            filename = os.path.join(root, file)
            fpl=pd.read_json(filename)
            time=fpl.shape[0]/300
            print(time)

if __name__ == "__main__":
    path='D:\data'
    get_f(path)
```

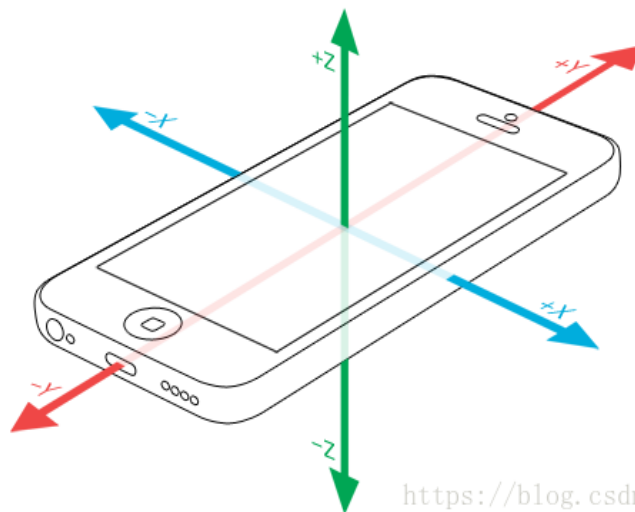
```
59.38333333333333
16.01
21.93
12.853333333333333
9.63
12.516666666666667
109.76
```

因为答题时间大约在十几分钟到三四十分钟正常，上图可以看到甚至有答题时间为 109 分钟的，所以需要将这段时间特别长或者特别短的文件删除，我在这里设定 10 分钟到一小时为正常时间。

```
if(time<10.0 or time>60.0):  
    os.remove(filename)  
    print('remove',filename)
```

```
remove D:\data\accelerometer\anxiety\female\20191108191818_762_accelerometer.json  
remove D:\data\accelerometer\anxiety\female\20191109151608_988_accelerometer.json  
remove D:\data\accelerometer\anxiety\female\20191112101903_1991_accelerometer.json  
remove D:\data\accelerometer\anxiety\female\20191119174618_4121_accelerometer.json  
remove D:\data\accelerometer\health\female\20191108151946_633_accelerometer.json  
remove D:\data\accelerometer\health\female\20191115152332_3370_accelerometer.json
```

除此之外，有些答题者答题时可能将手机放在桌上，导致幅度过小。根据图片，可以推断当手机被放在桌上时，z 轴变化不大，device\_motion 对应 alpha.



<https://blog.csdn.net/wangmx1993328>

属性	类型	说明
alpha	number	当手机坐标 X/Y 和地球 X/Y 重合时，绕着 Z 轴转动的夹角为 alpha，范围值为 [0, 2*Pi)。逆时针转动为正。
beta	number	当手机坐标 Y/Z 和地球 Y/Z 重合时，绕着 X 轴转动的夹角为 beta。范围值为 [-1*Pi, Pi)。顶部朝着地球表面转动为正。也有可能朝着用户为正。
gamma	number	当手机 X/Z 和地球 X/Z 重合时，绕着 Y 轴转动的夹角为 gamma。范围值为 [-1*Pi/2, Pi/2)。右边朝着地球表面转动为正。

所以计算出 z 轴和 alpha 所对应的方差，若方差过小则说明手机被放置在桌面上，数据的参考价值不大，将这些文件删除。

```
def get_fn(path):
    for root, dirs, files in os.walk(path):
        for file in files:
            filename = os.path.join(root, file)
            fp1=pd.read_json(filename)
            z=fp1.var()['z']
            if(z<0.01):
                os.remove(filename)
                print('remove', filename)

if __name__ == "__main__":
    path='D:\data\gyroscope'
    get_fn(path)
```

```
remove D:\data\gyroscope\anxiety\female\20191108162812_694_gyroscope.json
remove D:\data\gyroscope\anxiety\female\20191109150824_974_gyroscope.json
remove D:\data\gyroscope\anxiety\female\20191109195733_1159_gyroscope.json
remove D:\data\gyroscope\anxiety\female\20191114114301_2856_gyroscope.json
remove D:\data\gyroscope\anxiety\female\20191114191036_3074_gyroscope.json
remove D:\data\gyroscope\health\female\20191108110732_549_gyroscope.json
```

```
def get_fn(path):
    for root, dirs, files in os.walk(path):
        for file in files:
            filename = os.path.join(root, file)
            fp1=pd.read_json(filename)
            z=fp1.var()['gamma']
            if(z<10):
                os.remove(filename)
                print('remove', filename)

if __name__ == "__main__":
    path='D:\data\device_motion'
    get_fn(path)
```

```
remove D:\data\device_motion\anxiety\female\20191109150824_974_device_motion.json
remove D:\data\device_motion\anxiety\female\20191109195733_1159_device_motion.json
remove D:\data\device_motion\anxiety\female\20191110171332_1435_device_motion.json
```

至此，完成了大致的数据预处理。