Homework10数据清洗说明文档

我的总体思路分为以下5部分：

1. 我写了一个爬虫程序将所有文件爬取到本地，然后第一步就是过滤到空文件,将文件大小小于2KB 的全部删掉。
2. 之后考虑每个文件是否有缺失值，如有缺失值，则用该维度（因为这些文件都有3个维度，需要分别考虑）的均值进行填充。
3. 之后考虑将采集时间过长或者过短的文件过滤掉，由于采集频率是5HZ，所以只需要用数据条数除以5即可得到时间，然后我的标准是保留15分钟到60分钟的文件。
4. 经过前三个步骤之后，再考虑每个文件内的数据，通过画出线图可以初步断定，有些文件数据波动是不大的，这就意味着这些文件不合格，所以应该删掉，由于有3个维度，我分别计算3个维度的方差，然后将其可视化，可以发现，三个维度波动情况是差不多的，所以我以第一个维度为标准，过滤掉方差小于1的文件。
5. 最后一步，将前四步得到的文件列表进行分类，按文件名可分为6类（一共三个大类，每个大类分为健康和焦虑）