专业扩展探索（1）期末作业要求

**第一部分 数据预处理**

在本文档所在的压缩文件中包括上学期期末实习对小白鼠视频处理后的文件，要求对这些文件中的内容用pandas进行数据处理，然后使用百度开源的可视化工具Echart对处理结果进行可视化展示。

1、数据说明

要处理的数据包含在TextResult1.zip和TextResult2.zip中。解压缩后的文件名类似ch01\_20180116002121(1-2).txt的格式，每个对应的文本文件中的内容如下图。



文件的内容解释如下：

变化点：表示把视频每10帧截取一个画面，比较前后两个画面的像素点的变化。视频一段时间内前后截图的画面像素点变化的总和就是文件中的像素点。

帧数：表示视频总共截取了多少帧的画面。

时间：表示视频总共有多长时间，单位为秒。

文件名中包含了要对数据进行处理的信息。文件名的结构类似ch01\_20180116002121(1-2).txt，需要提取的信息如下。（备注：压缩文件中的文件名没有ch01\_这样的文件名我们不做处理，忽略掉）

ch01：可以抽象为“ch0?”，“？”通常表示一个任意字符，在我们这个实验里表示一个数字。具体的含义表示录制小白鼠视频的一个摄像头的编号。

20180116：这是一个日期，格式是“YYYYMMDD”，四位年度，两位月份，两位天数，表示2018年01月16日。是录制小白鼠视频的日期。

002121：这是时间，格式是“HHMMSS”，格式是小时、分钟、秒，表示录制小白鼠视频的时间。

(1-2)：小白鼠笼子的编号，通常是(1-1)、（1-2）、（1-3）、（1-4）这样的。没有笼号的文件名，在下面的处理中，默认笼号填写为字符“1”。

2、数据处理的第一步

要求从文件名中提取四位的摄像头编号，格式为字符串；八位的日期，格式为字符串；6位的时间，格式为字符串；3位的笼子编号，把括号去掉，数据的格式是字符串，类似这样“1-2”。

从对应文件名的内容中提取变化点数、帧数、时间。这三个我们作为数值型数据看待。

把上面提取到的六个数据项写入到一个新的文本文件中，这个文本文件命名为：“学号+姓名.txt”。这个文件的内容看起来是这样的（有一个表头），数据之间使用空格分割，不要求表头和数据在格式上对齐。

摄像头编号 摄像日期 摄像时间 笼号 变化点数 帧数 视频时间长度

ch01 20180116 002121 1-2 212573 33890 1412.083

把所有上学期实习计算出来的小白鼠的视频处理数据（文件名为：ch0?\_YYYYMMDDHHMMSS(?\_?).txt 的文件，两个压缩包中的文件）都按照附加到这个文件中，每个文件的内容处理为这里的一行数据。

3、数据处理的第二步

重新建立一个新的文本文件，命名为“学号+姓名+结果1.txt”。

这个文件的主要目的是把小白鼠的活动变化点数按照小时作为分割来做数据清洗，每天24小时，从00-23把它们分别处理在不同的时间单元内（一小时看做一个时间单元），这里处理的基本方法如下。

视频录制的时间我们提取出来是HHMMSS的格式，我们的视频大部分在23分钟（1412秒），规则如下：

（1）从HHMMSS中提取分钟，如果MM>=40，则把这段视频计入下一个小时，即HH记为HH+1；如果HH=23，则日期要加1，HH记为下一天的00时段。

（2）HHMMSS中提取分钟，如果MM<40，则把这段视频计入本小时，即HH。

（3）如果HH

“学号+姓名+结果1.txt”的文件内容如下：

摄像头编号 摄像日期 摄像时间 笼号 变化点数 帧数 视频时间长度 时段

ch01 20180116 002121 1-2 212573 33890 1412.083 00

这个文件的结果就是用来按照时段的不同，我们来对比不同笼子里面的小白鼠的活跃程度。

**第二部分 数据分析和统计**

问题描述：每个笼子里面的小白鼠用不同浓度的普洱茶进行喂养。我们的目标是对比不同笼子里面的小白鼠用不同浓度的普洱茶喂养之后他们的活跃程度。

以下每个分析统计的结果要求保存在一个文本文件中，按照顺序分别命名为“学号+姓名+统计结果（序号）.txt”。

（1）按照摄像头编号、笼子编号统计拥有的视频录像的天数

（2）对2018年1月16日到2018年2月4日之间不同摄像头编号，不同笼子编号，不同小时（每天24小时，从00-23）的变化点数进行分组统计。统计结果的格式如下：

摄像头编号 笼子编号 时段 变化点数

（3）把2018年1月16日到2018年2月4日之间共20天分为三个日期段，第一个日期段7天，第二个日期段7天，第三个日期段6天（分别编号为日期段1、日期段2、日期段3）。对不同摄像头编号，不同笼子编号，不同小时（每天24小时，从00-23）的变化点数进行分组统计。统计结果的格式如下：

摄像头编号 笼子编号 日期段 时段 变化点数

（4）对比分析(2)中不同摄像头、不同笼子编号，在相同时段间变化点数的差异是否显著？（共计24个时段分别对比）

（5）对比分析(3)中不同摄像头、不同笼子编号，不同日期段，在相同时段间变化点数的差异是否显著？（共计24个时段分别对比）

**第三部分 数据展示**

对第二部分（1）-（5）的分析结果使用Echart中合适的图标进行展示和对比，使得结果尽可能直观和明显。

可以同不同的图标展示同一种统计结果，以便看到不同统计图的统计效果。

作业提交要求：

（1）把三个部分的数据处理步骤截图写在一个Word文档中，并做必要说明。

（2）在每一部分的处理代码单独保存为一个文件，并用明确清晰的文件名命名。推荐使用Jupyter NoteBook来完成。

（3）作业通过TimeIT服务器提交。

（4）抄袭作业一律零分处理。