**说明文当**

1. **范围**

本需求分析指导本次python脚本程序，该脚本程序用于处理扫描的数据

1. **总体要求**

1）需要计算出每次任务中每秒爬取的页面数PagePerSecond，并附加到附件txt中的最后一列。

2）根据每次任务计算出的PagePerSecond，去除附件中这一组数据中PagePer-

Second值偏离比较大的任务。

1. **需求分析**
2. 根据每次任务中的total\_pages即任务爬虫爬取总页面数以及time\_used即任务扫描完成所用的时间（单位为s），利用公式total\_pages/time\_used计算每秒爬取的页面数PagePerSecond，并附加到txt中的最后一列
3. 根据每次任务计算出的PagePerSecond，利用正态分布概率表，过滤数据中PagePerSecond偏离值比较大的任务
4. 根据任务爬虫爬取总页面数total\_pages，利用正态分布概率表，过滤数据中页面总数偏离数比较大的任务
5. **可行性分析**

根据峰度计算公式： （其中， 为的平均值，S为Y的方差），可以算出，任务数据的峰度值为3，符合正态分布，因此，可以根据正态分布来对数据进行处理。

1. **设计**

查阅正态分布概率表知(μ±δ)的概率为68.268949%，(μ±1.96δ)的概率为95.449974%，(μ±2.58δ)的概率为99.730020%。由于给出数据较少，为看出效果，我们选用的范围为(μ±δ)对数据进行处理。从keypage-test-2014-07-10-14-30-01.983972.txt文件中提取数据，计算出每秒爬取的页面数PagePerSecond，对数据进行处理后写到save1.txt文件中。

1. **解决方案**

本次采用python语言，

1. 从keypage-test-2014-07-10-14-30-01.983972.txt文件中读取数据存入flist列表中，并用字符串分割获取爬虫爬取总页面数、任务执行的时间
2. 计算每秒爬取的页面数：pagepersecond
3. 计算pagepersecond的平均值和方差：other
4. 根据other求得的平均值和方差判断数据是否属于正态分布范围判断：Gaussian
5. 计算爬虫爬取总页面数的平均值和方差：filterByTotalpages
6. 根据filterByTotalpages求得的平均值和方差判断数据是否属于正态分布范围判断：Gaussianforpages
7. 运用两个循环分别对每秒爬取的页面数和总页面数进行过滤，对flist进行修改，得出最终数据
8. 将该数据存入新的文件夹save1.txt中
9. **问题**

1）如果使用正态分布对总页数进行过滤，那么爬虫爬取较大的页面数据也将被过滤，是否可取？

2）为了在数据中有效果，在用爬取总页数进行数据过滤时，正态分布选用为：

(μ±1.96δ)，与用每秒爬虫过滤不一致。