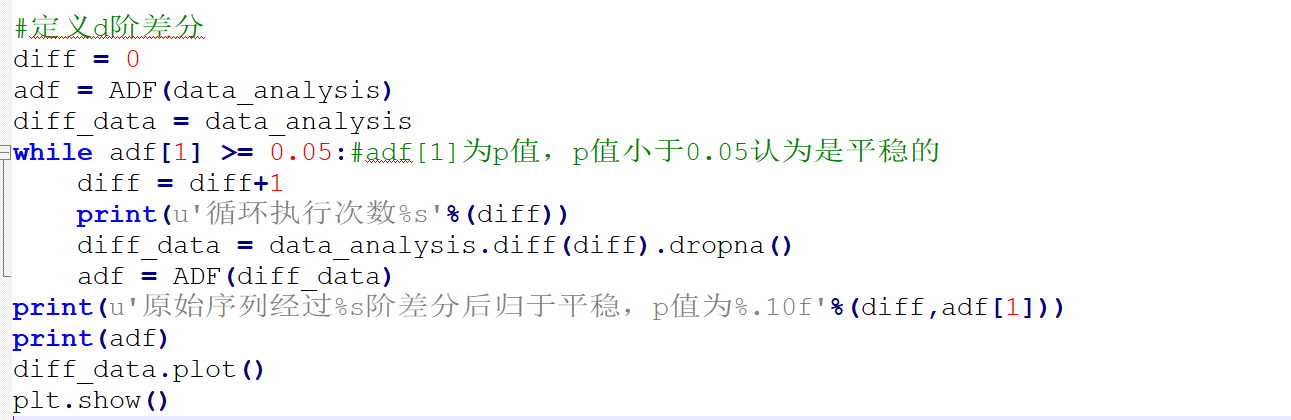
# Arima算法使用步骤：

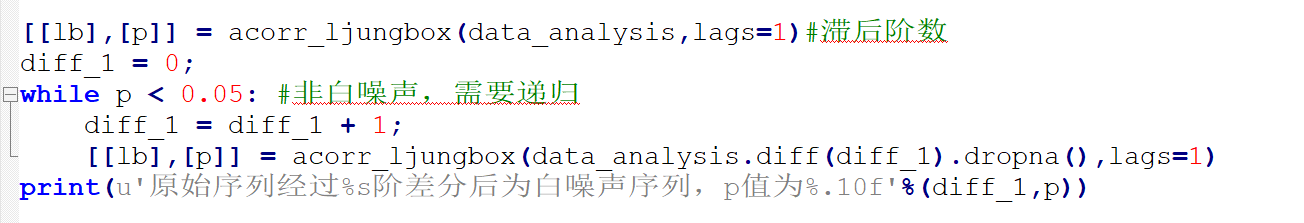
1. 清理数据：从es中导出历史数据，然后进行处理；按照天统计数据，计算实际剩余资源，作为预测数据。



1. 数据平稳性分析：对数据绘图，观测是否为平稳时间序列；对于非平稳时间序列要先进行d阶差分运算，化为平稳时间序列；判断标准（ADF检验可以得到单位根检验统计量对应的p值，若此值显著大于0.05，则该序列非平稳，否则认为是平稳数据），参考算法：



1. 白噪声处理：为了验证平稳序列中有用信息是否已被提取完，需要对序列进行白噪声检验，如果序列是白噪声的说明序列中的有用信息已经被提取完毕了，剩下的全是随机扰动，无法进行预测和使用。判断标准（检验时间序列是否为白噪声序列，一般如果统计量的P值小于0.05时，则可以拒绝原假设，认为该序列为非白噪声序列，否则为白噪声序列），需要循环处理平稳性和白噪声处理，保障序列的平稳性和白噪声序列。



1. 模型识别：采用极大似然比方法进行模型的参数估计，估计各个参数值，然后针对各个不同模型，采用BIC信息准则（贝叶斯信息准则）对模型进行定阶，确定p,q参数，从而选择最优模型。

确定最佳p、d、q值,d来源于之前的阶数diff字段

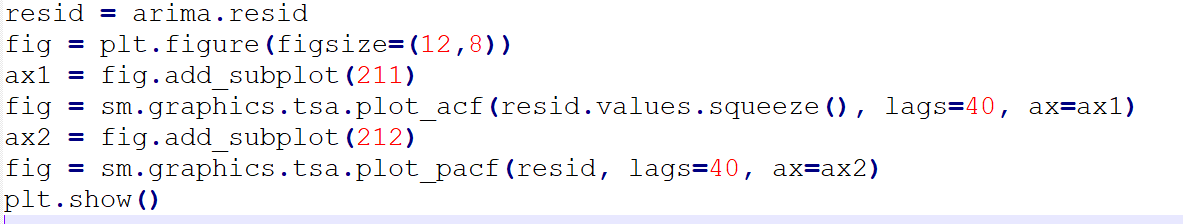
p和q的值可以通过对自相关系数图和偏自相关系数图人为识别来确定，

确定方法是根据自相关系数图和偏自相关系数图的拖尾和截尾的性质来确定的，同样我们也可以通过BIC矩阵中的最小BIC信息量的位置来决定.



1. 模型检测：在指数平滑模型下，观察ARIMA模型的残差是否是平均值为0且方差为常数的正态分布（服从零均值、方差不变的正态分布），同时也要观察连续残差是否（自）相关。

5.1检验残差序列



5.2做D-W检验 德宾-沃森（Durbin-Watson）检验

德宾-沃森检验,简称D-W检验，是目前检验自相关性最常用的方法，但它只使用于检验一阶自相关性。

因为自相关系数ρ的值介于-1和1之间，所以 0≤DW≤４。并且DW＝O＝＞ρ＝１　　 即存在正自相关性

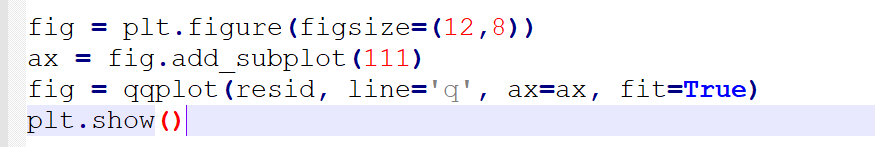
DW＝４＜＝＞ρ＝－１　即存在负自相关性

DW＝２＜＝＞ρ＝０　　即不存在（一阶）自相关性

因此，当DW值显著的接近于O或４时，则存在自相关性，而接近于２时，则不存在（一阶）自相关性。这样只要知道ＤＷ统计量的概率分布，在给定的显著水平下，根据临界值的位置就可以对原假设Ｈ０进行检验。

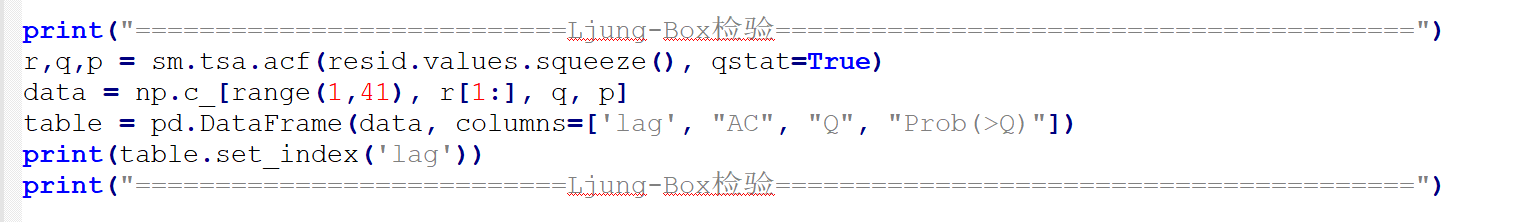
5.3观察是否符合正态分布

这里使用QQ图，它用于直观验证一组数据是否来自某个分布，或者验证某两组数据是否来自同一（族）分布。



5.4Ljung-Box检验

Ljung-Box test是对randomness的检验,或者说是对时间序列是否存在滞后相关的一种统计检验。对于滞后相关的检验，我们常常采用的方法还包括计算ACF和PCAF并观察其图像，但是无论是ACF还是PACF都仅仅考虑是否存在某一特定滞后阶数的相关。LB检验则是基于一系列滞后阶数，判断序列总体的相关性或者说随机性是否存在。时间序列中一个最基本的模型就是高斯白噪声序列。而对于ARIMA模型，其残差被假定为高斯白噪声序列，所以当我们用ARIMA模型去拟合数据时，拟合后我们要对残差的估计序列进行LB检验，判断其是否是高斯白噪声，如果不是，那么就说明ARIMA模型也许并不是一个适合样本的模型。检验的结果就是看最后一列前十二行的检验概率（一般观察滞后1~12阶），如果检验概率小于给定的显著性水平，比如0.05、0.10等就拒绝原假设，其原假设是相关系数为零。就结果来看，如果取显著性水平为0.05，那么相关系数与零没有显著差异，即为白噪声序列。



1. 数据预测
2. 数据校对
3. 模型评价

# 调研计划

后续需要完成数据预测、数据校对、模型评价等工作。根据历史数据建立数据模型完成模型的多次训练，修改模型确定算法。

工作量评估：

数据预测：9人天，根据历史数据计算未来数据。

数据校对：15人天，通过未来数据与现有数据进行比对，通过数据的校对调整arima的参数。

模型评价：15人天，通过平均绝对值、均方根误差、均方根误差和平均绝对误差，这三个指标从不同侧面反映了算法的预测精度。

完成以上工作确定该算法的可行性。

以下时间计划按照3个人投入计算（目前只有孙辽东），具体的工作计划

|  |  |  |
| --- | --- | --- |
| 功能 | 开始时间 | 截至时间 |
| 数据预测 | 2018-01-30 | 2018-02-01 |
| 数据校对 | 2018-02-02 | 2018-02-08 |
| 模型评价 | 2018-02-09 | 2018-02-22 |

# 算法实现

1. 使用java定时导出es的集群的历史数据存档（放入数据库中），可以按照增量的方式导出数据。--后台开发9人天
2. 使用java定时清理数据，以excel的方式导出数据（或者存放到数据库中）--后台开发9人天

2.1计算实际的剩余资源

2.2计算由于增加主机造成的资源影响

2.3按照天统计数据（按照天平均数据）

2.4数据补齐（如果哪天数据为空，需要获取之前的数据作为当天的数据）

1. 使用python定时完成数据的预测，输出预测数据到数据库中

预测数据偏差阈值设置：可容忍的数据误差度--前后台共12人天

在预测过程中完成数据模型的选择和评估。--后台开发9人天

1. 使用java在云平台中完成数据的展现--前后台开发15人天

时间计划按照3个人投入计算，具体工作计划

|  |  |  |
| --- | --- | --- |
| 功能 | 开始时间 | 截至时间 |
| 历史数据导入 | 2018-01-30 | 2018-02-01 |
| 数据清理 | 2018-02-02 | 2018-02-04 |
| 阈值设置（前端） | 2018-01-30 | 2018-02-02 |
| 阈值设置（后端） | 2018-02-05 | 2018-02-09 |
| 模型选择开发 | 2018-02-10 | 2018-02-14 |
| 数据展现（前端） | 2018-02-22 | 2018-02-28 |
| 数据展现（后端） | 2018-02-22 | 2018-03-05 |