# Spark MLlib机器学习01作业

Spark MLlib机器学习第一周的作业，主要考查大家对RDD知识的掌握，本次作业将通过真实案例进行练习，希望大家认真作业。

1、数据源：

用户安装列表，数据格式：上报日期、用户ID、安装包名；

解释下：通过第三方工具，收集到每个用户手机中的安装应用包名，收集的时间间隔是每天收集一次，所以本次数据是日期、用户、包名，也就是收集用户每一天的安装列表。

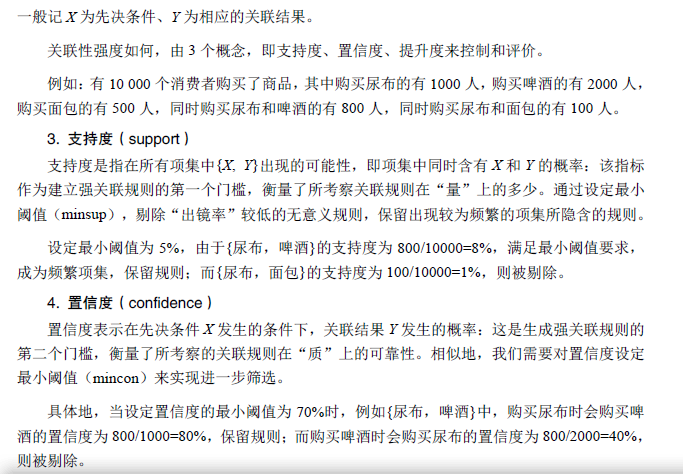
2、作业：

1）通过Spark读取安装列表数据，并且统计数据总行数、用户数量、日期有哪几天；

2）指定任意连续的2天，计算每个用户第2天的新安装包名；

3）指定任意1天的数据，根据每个用户的安装列表，统计每个包名的安装用户数量，由大到小排序，并且取前1000个包名，最后计算这1000个包名之间的支持度和置信度，名词解释如下：





其中设置支持度的阈值是0.1%，置信度阈值30%，大家与可以自己根据实际情况进行设置阈值；

结果需要是：包名A、包名B、支持度值、A->B的置信度、B->A的置信度。