**上车位置挖掘**

更新日期：2016/12/13

# 1、数据预处理

## 1-1数据源获取

Hive表导出：service\_order\_ext、service\_order\_ext

时间范围：最近三个月，若数据稀疏再补充采样

地点范围：北京

## 1-2坐标按区域映射

## 方法一：采用Geohash对以下坐标进行映射

|  |  |  |
| --- | --- | --- |
| 序号 | 数据名 | 映射精度 |
| 1 | 创建订单坐标 | 6位，±0.61km  7位，±0.076 km |
| 2 | 司机到达坐标 | 8位，±0.01911km  9位，±0.00478km |
| 3 | 开始服务坐标 |
| 4 | 服务结束坐标 |

## 方法二（备选）：采用逆地址编码

|  |  |  |
| --- | --- | --- |
| 序号 | 数据名 | 备注 |
| 1 | 创建订单坐标 | 地图接口获取拟地址，做murmurhash3 |
| 2 | 司机到达坐标 |
| 3 | 开始服务坐标 | 订单的address+position，做murmurhash3 |
| 4 | 服务结束坐标 |

# 2、频繁项集挖掘

问题定义：

主要步骤：

1. 统计Geohash的频次，以的频次作为支持度，设置合理的截断阈值，排除干扰数据

优化点：根据数据是否稀疏，考虑增加统计数据集或缩小Geohash的精度

1. 统计Geohash的 的组合频次，设置合理的阈值截断
2. 根据步骤二的结果，输出Geohash的关联关系：，推荐的Geohash根据组合出现的频次排序，并采用MinMaxScale作为权重值

# 3、映射坐标逆解析

## 3-1 方法一（平均法）：

假设上车地点的坐标比较准，且Geohash取了比较高的精度，则直接求区域坐标的均值或中位数，可以尝试的方法如下：

1. 对Geohash内的推荐坐标集做异常值处理，假设其满足高斯分布，则采用1-2个标准差的方式进行数据过滤，然后求坐标均值
2. 分别对Geohash内的推荐坐标集的lat, lng求中位数，以中位数作为推荐点坐标

上述数据生成后，先投到地图上看一看效果，若不行再采用后续方法修正

## 3-2 方法二（逆地址修正）：

若方法一的求出的推荐坐标不准，则采用该坐标通过地图服务获取创建订单点的逆地址数据，通过东西南北的街道坐标进行修正，可以尝试如下方法：

1. 采用权重大于一定阈值的推荐坐标，求与其距离最近的街道坐标，直接替换为街道坐标信息
2. 按照街道坐标的方位，分别对推荐坐标的精度或纬度进行修正。例如：用南北的街道修正纬度，东西的街道修正精度

## 3-3 方法三（导航规划）：

根据方法一求出的推荐坐标，在东西南北4个方向生成偏移量坐标，调用地图服务做路径规划，取最近规划点的坐标