这次美团点评杯，我们的数据挖掘算法使用场景为**用户选定房间入住之后**，我们的客户端会为用户**推荐**定制化的信息，类似于推荐系统。

考虑到用户的酒店信息属于个人隐私，我们没有找实际数据去做参考。因此我们假定把用户出行目的分为两种类型，**一是出差，二是游玩**。然后我们又把用户的消费水平分为两种类型，**一是消费能力强，二是消费能力弱**。我们这样划分的依据是现实生活中出去订酒店的经验。而这样划分的目的是我们认为出差的人更关心的信息是交通信息或者天气信息，而出来游玩的人更关心的是吃喝玩乐、景点门票的信息；又可以根据消费能力为用户推荐相应消费水平的信息，通过这些达到优化用户体验的效果。

关于**预测用户出行**目的，我们使用的算法为借鉴**kmeans**的聚类算法。之所以使用聚类算法是因为虽然我们假设用户信息中包含了两种类型模式，但是我们并不知道每个用户的具体类别。因此无法使用基于监督学习的分类算法，而使用聚类算法后，用户会依照我们制定的规则自动分为两类，我们不必了解每一类具体是什么。只要聚类完成后我们取出每一类的中心即可。

首先，我们为每一个用户构建一条**出行轨迹特征向量**，该特征向量是依据由用户信息中的**地理信息**抽取而成。使用地理信息构成特征向量的依据是我们认为出行目的为差旅的用户出行目的地（也就是所订酒店的地理位置）是相对固定的，而出行目的是游玩是用户其目的地是分散的。

然后我们实现了度量两条出行特征向量距离的方法：类比于自然语言处理中的文本向量相似度，我们将它改到我们的算法中，其中的核心算法类似于**文本中公共子串的长度度量**。

最后，我们以**用户出行轨迹向量**为输入，加上自己实现的距离度量函数完成聚类算法。具体步骤为：

1.从原始用户行为轨迹中随机抽取两条作为中心向量。

2.遍历所有轨迹，依照这两条中心向量进行聚类。

3.在生成的两类特征向量中重新求中心向量。

4.重复1-3直到一定的迭代次数或者两个中心向量最具有代表性。

完成上面的聚类之后，既可以得到两个最具有代表性的**中心向量**。这时如果有一个用户完成了订单，我们可以构成该用户的新的特征向量。将这个新的用户向量与聚类生成的两个特征向量比较，可以得到该用户的所属类别。依照该类别即可给用户推荐相应的界面。