2. 淘宝上有几亿的买家和千万级的商家，我们需要在每个月初给头部的商家（已知商家的名单）推荐一定数量的用户，便于商家有针对性的进行营销活动。

要求：a. 请给出具体的方案和实现方法，包括所用的模型、数据、训练和预测的方法。

b. 已有的数据是过去半年所有买家的行为数据，即过去半年的每一天里，单个买家在每个商品上的浏览、收藏、推荐，和购买数量。

问题回答：

首先考虑利用过去半年买家的行为数据。对于用户的行为，可以设置加权方法，模拟用户对该商品的打分。比如浏览为2分，收藏为4分，推荐为6分，购买为10分，再对每个买家的历史打分对时间求平均值。在预处理之后，我们得到一个大的矩阵，其中每一行是单个买家对各个商品的喜好程度（根据分值高低）。该矩阵是一个大的稀疏矩阵，记为M。对于矩阵M，可以利用矩阵分解的办法分析数据，将矩阵M分解为矩阵P和矩阵Q的乘积，即M=PQ，矩阵Q，P中的每一个元素q_{ij}和p_{ij}为模型所期望训练得到的参数。因为M是一个稀疏矩阵，考虑M中的每一个非零元素m_{ij},模型的损失函数为

$$L = \sum_{i,j}(m_{ij} - q_j^T p_i)^2 \tag{1}$$

利用梯度下降法对p_i和q_j迭代求解就可以训练得到最终的参数。

做预测时，将PQ矩阵相乘，即可得到新的矩阵M'，此时M'为稠密的矩阵，每一行代表了单个用户对不同商品的喜好程度。

最初的M是一个非常巨大的矩阵，矩阵分解的开销过大。因此可以从两方面对原数据进行分割。首先绝大多数商家会有主营产品，并且大部分产品属于同一类别，比如图书，家电等等，可以根据商品的不同类别对矩阵的列进行提取和切割。另外，我们可以将数亿个用户当做独立同分布的个体，将数亿个用户拆分成数十甚至数百份，对每一份的数据单独进行矩阵分解，每一个拆分后的矩阵都是独立的，并且拆分后的矩阵参数数量大大减少，可以通过多台机器并行计算提高效率。

最后，对于已知的商家，因为商家经营商品的范围有限，所以可以寻找该商家对应大类的那些矩阵，抽出代表商家所经营的商品对应的列，对每行进行求和，可近似看作该客户对于该商家的喜好程度，对所有矩阵里的喜好程度进行排序，选出最前面的TOP K个客户推荐给商家。也可以先

选定一个代表客户，通过计算向量相似度的方式，选择出相似度最大的K个客户，推荐给商家。