**第四周作业：对活动进行聚类**

第四周和第五周的作业均数据来源于Kaggle竞赛：Event Recommendation Engine Challenge，根据

events they’ve responded to in the past

user demographic information

what events they’ve seen and clicked on in our app

用户对某个事件是否感兴趣

竞赛官网：

https://www.kaggle.com/c/event-recommendation-engine-challenge/data

活动描述信息在events.csv文件：共110维特征

前9列：event\_id, user\_id, start\_time, city, state, zip, country, lat, and lng.

event\_id：活动的id,

user\_id：创建活动的用户的id .

city, state, zip, and country： 活动地点 (如果知道的话).

lat and lng： floats（活动地点的经度和纬度）

start\_time： 字符串，ISO-8601 UTC time，表示活动开始时间

后101列为词频：count\_1, count\_2, ..., count\_100，count\_other

count\_N：活动描述出现第N个词的次数

count\_other：除了最常用的100个词之外的其余词出现的次数

**作业要求：**

根据活动的关键词（count\_1, count\_2, ..., count\_100，count\_other属性）做聚类，可采用KMeans聚类

尝试K=10，20，30，..., 100, 并计算各自CH\_scores。

提示：由于样本数目较多，建议使用MiniBatchKMeans。

文件说明：

1. 可以先运行0. EDA.ipynb，看一下竞赛所有数据的情况；
2. 总体活动的数目太多（300w+记录），可以只需对训练集train.csv和测试集test.cv出现的活动（13418条记录）举行聚类即可。运行1. Users\_Events.ipynb可得到只在训练集train.csv和测试集test.cv出现的活动，可自己修改代码存为csv格式，在进行聚类。