开源项目聚类与学习：

1. 数据建模，特征提取。（\*\*\*）

可用的特征：

连续特征：repo\_name, repo\_description, issue\_body等

离散特征：repo\_language等

如何辨别有效的信息？

1. 距离定义，聚类方法：（\*\*）

连续特征：欧式距离 ?

离散特征： ?

混合距离： ?

1. 学习（\*）

何种模型？混合特征还能用MLP or textCNN吗?