比赛数据：

Train\_set 2014-11-01 至 2015-03-31 151天 110541018 条微博

Test\_set 2015-04-01 至 2015-04-15 15天 9644805 条微博

线下数据集划分：

136<time<=151 offline\_ test

121<time<=136 offline\_val

106<time<=121 offline\_ train

数据预处理：

特征工程：

1. 用户特征（画像）：

1.1用户活跃程度

用户总发微博数；用户近30天发微博数；用户近7天发微博数；用户近3天发微博数；总活跃天数；近30天活跃天数；近7天活跃天数；总微博数/总活跃天数；近30天微博数/近30天活跃天数；近7天微博数/近7天活跃天数；活跃天数的标准差；上一次活跃距离预测日的时间（小时）；每天最大发微博数，以及这一天总行为数；每小时最大发微博数，以及这一小时总行为数。

1.2用户是否是大V

用户历史总行为数；用户历史的中位数；用户历史的平均数；用户历史的标准差；用户总行为/总微博数 ；用户总行为/总活跃天数；单条最大总行为数；

1.3用户使用微博的习惯 ps：需要统计发微博的时间，才能确定时间如何分段 一般上班高峰，吃饭前后，睡觉前后

发微博在一天中的时间，

1.4粉丝数据 ps:确定粉丝用户 fans\_id和 和uid 的重合程度，如果重合度高，就可以使用多级结构

用户粉丝的数量，用户粉丝的质量（活跃度和影响力），用户粉丝中大V的数量。

用来构建社区网络结构，为后面做铺垫。

1. 文本特征：

2.1 LDA 主题模型

2.2 关键词，词频统计

1. 网络分析： 如何构建社区网络？

3.1首先利用**KCore算法**剔除社区网络中影响较小的用户节点，保留主要的关键节点，再利用PageRank找到最具影响力的关键节点，然后利用**单源最短路径SSSP**找到目标用户到指定节点的距离，并且把距离作为特征。

3.2 直接利用**标签传播聚类**算法，利用带有权重的节点和边，对图进行聚类，把聚类结果作为标签。

3.3 给筛选出具有代表性的关键节点打标签，然后利用**标签传播分类**算法分类，找到重要的相似节点，作为特征。