# 2017 机器学习第三次作业要求

## 1 作业

**数据集** DBLP 数据库中的 IJCAI, AAAI, COLT, CVPR, NIPS, KR, SIGIR, SIGKDD 八个会议中**从 2007 年至今**的所有数据。也可在 DBLP 中适当扩展数据集，但这八个会议必须包含在内。

**任务 1**

- 每一个会议都有各自的支持者，现在请你将每个会议各自的研究者寻找出来，并且根据时间信息，看看哪些人依然活跃，哪些人不再活跃。
- 在找到各自的研究者群体后，我们希望找到经常性在一起合作的学者，将之称为'团队'。请你根据研究者合作发表论文次数为根据进行频繁模式挖掘，找出三个人以上的'团队'。

**任务 2**

- 每一篇论文都会涉及到一个或多个主题，请你先定出主题词，然后根据每个'团队'发表的论文的情况，提炼出这个团队最常涉猎的主题。
- 团队和主题多是会随着时间而动态变化。请你根据自己所定的时间段（五年，三年，两年或是一年）描述团队的构成状况以及其研究主题的变化情况。

## 2 作业要求

- 完成上面两个任务，并将分析过程和结果写成报告与程序打包提交，在文中写清楚每个人的分工以及每个人的姓名学号等信息，命名格式为组号 _ 作业次数.zip，如本次 3 号组的命名为 3_3.zip

- 提交的截止日期为 1 月 3 日晚 23 点 59 分，迟交者本次作业记 0 分，请大家在截止日期之前将作业发送到ML_PKU_2017@163.com

# 3 关于本次作业数据集的说明

本次作业的数据集我们提供两个，大家可以从中二选一使用。

第一个数据集是 DBLP 的原始数据集,地址为http://dblp.uni-trier.de/xml 中，数据格式为 xml 格式。其中包含很多会议，杂志中论文的情况。

第二个数据集是我们对第一个数据集进行了筛选后的数据集，其中仅包含 2007 年以来上述八个会议的数据。数据示例格式如下

```
########
Author name: Shan-Hung Wu
Author name: Keng-Pei Lin
Paper title: Efficient Processing of Metric Skyline Queries.
IEEE Trans. Knowl. Data Eng.
2009
```

其中 # 代表起始行，之后是作者行，行数与作者人数相等，接着是论文标题行，再是会议（期刊）行，最后是年份行。

两份数据我们之后均会上传至网盘。

# 4 网盘地址

教材，讲义，作业要求和每个队伍的队号信息将放在网盘中共享。

链接: https://pan.baidu.com/s/1pLdrjth 密码: uiv4

# 5 助教邮箱

有问题请及时联系助教。

李媛：ylmath1993@163.com

王恒亮：wanghl@pku.edu.cn