# 基于K近邻的车牌号识别

## 一、案例简介

## 每年国际上召开的大大小小学术会议不计其数，发表了非常多的论文。在计算机领域的一些大型学术会议上，一次就可以发表涉及各个方向的几百篇论文。按论文的主题、内容进行聚类，有助于人们高效地查找和获得所需要的论文。本案例数据来源于AAAI 2014上发表的约400篇文章，由

## [UCI](https://archive.ics.uci.edu/ml/datasets/AAAI+2014+Accepted+Papers!)公开提供，提供包括标题、作者、关键词、摘要在内的信息，希望大家能根据这些信息，合理地构造特征向量来表示这些论文，并设计实现或调用聚类算法对论文进行聚类。最后也可以对聚类结果进行观察，看每一类都是什么样的论文，是否有一些主题。

## 注：数据中的group和topic也不能完全算是标签，因为

## 1. 有些文章作者投稿时可能会选择某个group/topic但实际和另外group/topic也相关甚至更相关；

## 2. 一篇文章可能有多个group和topic，作为标签会出现有的文章同属多个类别，这里暂不考虑这样的聚类；

### 二、作业说明：

1. 将文本转化为向量，实现或调用无监督聚类算法，对论文聚类，例如10类（可使用已有工具包例如sklearn）；

2. 并将聚类结果可视化成散点图

3. 高维向量的降维旨在去除一些高相关性的特征维度，尝试使用PCA进行降维后，再聚类

4.尝试使用一些评价指标，对聚类后的结果进行评价

## 三、数据预处理与可视化

## **四、**模型构建与训练

## 五、结果

**六、讨论**

## 七、结论

## 八、参考资料

## 九、遇到问题以及解决方法