数据集说明

课程项目采用MIRFlickr25000数据集，该数据集是2007年3月21日至2008年1月30期间从社交媒体分享网站Flickr从抓取的25000张图片以及用户的标注数据。

在本项目中，我们只针对用户标注数据进行分析。由于原数据中部分样本的用户标注数据缺失，因此我们对该数据集进行处理，去除没有标注数据的样本。此外，为了适应单分类问题，我们保留原17个类别中的10个类别 (tree, night, clouds, flower, food, dog, car, bird, baby, lake)，并去除同时属于多个类别的样本，最终，我们获得7425个样本。

我们将所有样本划分为训练集、验证集和测试集，具体划分如下：

训练集合：

x\_train ： 5425 个训练样本的列表，每个元素表示一个用户标注文本

y\_train ： 5425 个训练样本对应的类别 （共10个类别）

验证集合：

x\_valid ： 1000 个验证样本的列表

y\_valid ： 1000个验证样本对应的类别

测试集合：

x\_test ： 1000 个测试样本的列表

y\_test ： 1000个测试样本对应的类别

说明：

1. 所有数据采用pickle技术存储在文件data\_student.pkl 中（学生在训练阶段只获得x\_train,y\_train,x\_valid,y\_valid，测试样本将会留作最终的算法性能评估）
2. 如何从'data\_student.pkl'读入数据？

利用pickle自带的解析器，从data\_student.pkl读取相应的数据，读取的语句如下：

import pickle

x\_train,y\_train,x\_valid,y\_valid = pickle.load(open('data\_student.pkl','rb'))