数据挖掘Project 2

作业要求：

一、数据预处理与特征提取

samples_50000/路径下提供了50000篇英文新闻文档，格式说明见Project 1。

1. 抽取新闻类别作为分类的目标。每个新闻可能有多个类别。可以对每个类别进行二分类，即判断一篇新闻属于或不属于该类别。也可以直接作为多分类预测任务（注意有的分类工具是不支持这种任务的，此时建议使用二分类）。
2. 使用Project 1中得到的新闻全文的tf-idf向量作为新闻的特征。
3. 对于分类任务，将所有新闻按照9:1的比例随机分成训练集和验证集，其中训练集用于训练分类模型，验证集用于评测分类效果。

二、基本分类器运用与比较

至少使用以下几种分类算法在训练集上进行训练（鼓励同学们探索更多更新的分类算法），对于得到的每一个分类器，在验证集上计算Precision、Recall和F1-measure三个指标，并根据这些指标和训练速度比较不同分类器的效果。

- ✓ Logistic Regression
- ✓ Naive Bayes
- ✓ SVM(Support Vector Machine)
- ✓ Decision Tree
- ✓ MLP(Multi-Layer Perceptron)

三、ensemble 算法运用与比较

至少使用以下几种ensemble算法在训练集上进行训练（鼓励探索），根据Precision、Recall和F1-measure三个指标在验证集上比较各自的分类效果，并比较各种方法的训练速度。

- ✓ Bootstrap
- ✓ AdaBoost
- ✓ Random Forest
- ✓ Gradient Boost(建议使用xgboost工具包)

四、聚类算法运用与比较

至少使用以下几种聚类算法对整个数据集进行聚类（鼓励探索，鼓励降维后对聚类结果进行可视化），计算NMI(Normalized Mutual Information)、AMI(Adjusted Mutual Information)两个指标，并根据这些指标和运行速度比较不同聚类算法的效果。

- ✓ K-means
- ✓ DBSCAN

提示：

1. 有的模型在运行时需要将全部训练集加载到内存中，一定要以稀疏矩阵的形式来表示tf-idf矩阵。如果仍然有内存不够的问题，可以根据自己的实际内存情况减小数据集规模，如从50000篇中随机选取30000篇作为数据集。
2. 部分文档中有可能存在乱码，建议使用utf-8编码方式进行读取。如果乱码影响到文件读取或程序正常运行，建议在做预处理时先写一个脚本将乱码去除。
3. 如果有的类别新闻数量过少，比如只有几篇或者十几篇新闻，可以将该类别剔除。

作业说明：

以小组为单位完成。
可以使用各种语言，建议使用PYTHON或R。
可以使用各种工具包。
作业要求中所有鼓励均为加分项。
提交源代码、报告。报告中要求说明：

- **使用的工具包**
- **各种算法的分类效果与运行速度**
- **组内分工情况**
- **其他任何你认为有趣的结论**