# 文本分类实验考核要求

**1、概述**

**文本分类**是自然语言处理（NLP）中的一种核心任务，它的目标是根据文本的内容或特征，将文本分配到预定义的类别中。文本分类广泛应用于各种实际场景，例如垃圾邮件检测、情感分析、新闻分类等。

实验要求：学生能完成一个文本分类任务。具体分类类别不限制，如新闻分类、垃圾邮件检测、产品评论分类等等均可以。

实验目的：掌握数据预处理、文本建模、熟悉分类算法、对分类器（模型）性能评估以及可视化分析的能力。

**2、实验步骤**

1. **自行寻找和准备适当的训练集和测试集：**如从公开数据集获取或自建数据集。
2. **对文本数据进行预处理：**可能包括分词、去停用词、构建词典表示（如词袋模型或TF-IDF），并针对数据特点进行必要的特殊处理（如统一文本格式、处理缺失值或特殊符号）。
3. **任意选择一种分类器进行分类：**机器学习（例如朴素贝叶斯、SVM、决策树）等或者深度神经网络完成模型训练
4. **用测试集评估模型性能：**计算准确率、精确率、召回率和F1值。
5. **撰写实验报告：**需描述你完成整个实验的详细步骤，具体可能包括数据来源、预处理步骤、模型选择过程、调整参数以及性能评价等等，同时需要提出实验中遇到的问题或者可能的优化方案以及自己的总结反思。

**3、考核方式**

（1）请于12.8日24点之前，提交源代码及实验报告合并为压缩包，命名为姓名+学号，将压缩包发送至助教邮箱。如超过此时间，将没有实验报告成绩。具体划分范围：

1-9班同学提交至助教1邮箱：my\_yang@bupt.edu.cn

10-14班或其他情况的同学提交至助教2邮箱：jaydenchan@bupt.edu.cn

1. 源代码：包括训练集、测试集、算法模型等全部相关源代码。
2. 实验报告。

上述源代码和实验报告请打包合并为压缩包，提交到指定邮箱！！！

（2）12.9上课时检查每位同学实验运行情况。要求：

1. 每位同学尽可能携带电脑，方便进行现场代码演示（若不方便携带电脑需提前和老师沟通）
2. 每位同学需对实验设计、数据处理流程、模型选择、性能评估和结果分析进行简要汇报,并回答检查过程中提出的相关问题。

# 实验评分标准

**1、报告部分评分（0-70分）**

1. **训练集和测试集选取（0-10分）**

* 训练集和测试集的大小是否合适，是否符合实验要求；
* 训练集和测试集的比例是否合理；
* 数据集是否涵盖两个或者更多类别，且类别分布合理；

1. **文本预处理（0-20分）**

* 预处理基本步骤是否合理，包括分词、去停用词和文本向量化等；
* 是否根据数据特点进行特殊处理，如处理特殊符号、统一格式。

1. **分类器选择（0-25分）**

* 分类算法选择是否合理。
* 是否选用合适的核函数或模型调优。
* 是否设计创新的方法或改进点？

1. **性能评估（0-15分）**

* 使用测试集评估模型性能，计算准确率、精确率、召回率和F1值，结果合理且完整
* 可视化图表表达；
* 实验过程的详细完整记录和分析。
* 实验结果的展示和讨论

1. **实验运行情况课堂检查（0-30分）**

**(1)助教邮箱所接收的源代码是否完整（0-5分）：**

* 包括训练集、测试集、分类器代码及相关依赖文件；
* 是否缺少关键模块或文件，导致实验无法正常运行。

**(2)实验过程是否正确完成（0-15分）：**

* 数据预处理步骤是否完整且合理；
* 模型训练是否能够正常运行；
* 参数调整过程是否体现出对模型优化的思考；
* 是否能够合理解释运行过程中可能出现的异常。

**(3)实验结果是否符合要求（0-10分）：**

* 模型评估指标（如准确率、精确率、召回率和F1值）是否合理且与报告一致；
* 可视化结果（如混淆矩阵、分类效果图）是否具备且准确；
* 实验结果是否体现出对分类问题的深入理解。