# 文本分类任务

## 任务简介

文本分类是自然语言处理（NLP）的核心任务之一，其目标是将文本数据自动归类到预定义的类别中。例如，判断一条评论是“正面”还是“负面”情感，或将新闻文章分类为“体育”“科技”等主题。

## 典型应用场景

* 情感分析：分析用户评论、社交媒体内容的情感倾向（积极/消极/中性）。

“这款耳机音质很好，续航时间长，佩戴舒适，性价比很高。”→ 积极

“充电器插头接触不良，用了两天就坏了，客服处理速度慢。”→ 消极

“手机壳颜色和图片一致，材质普通，价格适中。”→ 中性

* 主题分类：为新闻、论文或网页内容打标签（如政治、金融、娱乐）。

《霸道总裁爱上我》→ 言情

《如何在股市当霸道总裁》→ 财经

* 垃圾内容识别：检测垃圾邮件、广告或违规文本。

"【全网最低价】代写毕业论文，硕博团队操作，保过查重！电话138XXXXXXX" → 违禁服务

"独家对话圣诺医药新CEO潘洪辉：重整小核酸药物第一股" → 正经推送

* 诈骗识别：检测文本是否为钓鱼邮件、虚假信息或诱导转账的信息

【XX银行】您的账户存在风险，请登录APP修改密码" → 合法通知

【XX银行】账户被冻结！立即点击http://假网站.com解冻" → 诈骗（附赠木马大礼包）

* 意图识别：在对话系统中理解用户请求（如“订票”“咨询”“投诉”）。

"这款手机的电池容量是多少？" → 意图：咨询参数

"支持支付宝付款吗？" → 意图：咨询支付方式

"收到的衣服有破损，要求换货" → 意图：投诉+售后

"买的充电器充不进电，疑似假货" → 意图：投诉+质检

# 任务描述

## 报告撰写要求

1. 完成报告中第二章的第2部分“TF-IDF关键词提取算法”（第1部分是上个任务的“TextRank关键词提取算法”）
2. 完成报告中第三章“文本分类”，描述GitHub仓库<https://github.com/nlpresearchers/nlpbook/tree/master/chapter07/bayes-mails-classify-master-main> 完成文本分类任务所使用的算法、原理和流程，需包含以下内容：

* 算法基础：说明该仓库采用多项式朴素贝叶斯分类器，解析其基于条件概率的特征独立性假设，阐述贝叶斯定理在邮件分类中的具体应用形式
* 数据处理流程：描述分词处理、停用词过滤等预处理步骤的实现逻辑
* 特征构建过程：对比分析高频词特征选择与TF-IDF特征加权的差异，说明两种方法的数学表达形式及实现差异

## 2、代码部署

基于GitHub仓库<https://github.com/nlpresearchers/nlpbook/tree/master/chapter07/bayes-mails-classify-master-main> 部署代码。其中classify.py为朴素贝叶斯文本分类实现，“邮件\_files.zip”为原始数据。如“邮件\_files.zip”下载困难可至“学习通-资料-数据”获取。解压邮件\_files.zip至项目指定路径，并在本地成功运行代码。

提交方式二选一：

1. 保留classify.py文件，将运行结果截图放至README.md中。
2. 新建classify.ipynb文件（需合理划分代码单元格），移植classify.py的内容，提交时结果应显示在单元格下方。无需在README.md中放置截图。

## 3、优化特征选择方法

目标：实现特征选择方法的参数化切换机制，允许通过传入参数在以下两种特征提取方式间灵活选择：

* 高频词特征（即原代码的特征选择方式）
* TF-IDF加权特征

Hint：采用sklearn.feature\_extraction.text.TfidfVectorizer实现TF-IDF值计算。

## 4、书写项目文档

在项目根目录新建README.md文件，需包含：

* 代码核心功能说明
* 高频词/TF-IDF两种特征模式的切换方法

将完整项目推送至新建GitHub仓库（注：因GitHub单仓库README渲染限制，请勿复用既往作业仓库）。将仓库地址提交至作业提交文档。

注：README中出现的所有代码需要用行间代码/代码块格式，代码格式任务二已实践过；出现的所有公式需要使用LaTeX公式。一般AI复制过来的公式已自带LaTeX格式，只需要在前后加上$（行内公式）或$$（块级公式），参考：<https://zhuanlan.zhihu.com/p/672248115>

## 5、样本平衡处理（选做，可获额外加分）

目标：缓解垃圾邮件（127条）与普通邮件（24条）的样本量失衡问题

Hint：在模型训练前插入过采样/欠采样代码，建议采用imblearn.over\_sampling.SMOTE实现。

## 6、增加模型评估指标（选做，可获额外加分）

目标：在基础预测功能外，输出包含精度/召回率/F1值的分类评估报告

Hint：通过sklearn.metrics.classification\_report实现多维度模型评估。

如完成了选做的第5点和第6点，请体现在README中。