**《大数据智能》**

**（ 2020学年 秋季 学期）**

**作**

**业**

**报**

**告**

**学 号：**

**姓 名：**

**班 级：**

**任课教师： 朱聪慧**

作业报告

实验名称： 构建中文文本分类系统

成绩：

实验类别：验证/综合型实验 实验要求：1人1组 时间：

## ￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣￣

1. 实验目的

熟悉词袋模型，使用sklearn框架中的贝叶斯或者SVM分类器，训练分类模型（在训练数据上），测试分类性能（在测试数据上）。

1. 实验内容

* 文件格式处理：

1. 解压文本分类数据包“文本分类.zip”。

C000008 文件夹 --财经 类别

C000024 文件夹 --军事 类别

每个文件是一个句子，已经分词，去掉停用词

文件号0~200，用作训练集

文件号201~220，用作测试集

1. 在所有文件，统计词频（可以合并，打开总的，也可以用程序遍历文件夹，打开每一个文件）
2. 将输入文件表示为标准词袋模型（分别生成训练、测试文件）
3. 训练模型（svm或者贝叶斯）
4. 测试原始词袋模型性能（使用精确率指标）
5. 加入TF-IDF特征（重新生成训练、测试文件）
6. 训练模型（svm或者贝叶斯）
7. 测试融合TF-IDF词袋模型性能（使用精确率指标）

三、使用的算法描述

1. 程序源码（拷贝至此处，或者作为附件和报告一起提交均可）

五、程序运行结果（将程序运行结果的截图，拷贝至此处）

六、心得体会