浙江工业大学

文本分析与挖掘

上机实验

计算机科学与技术学院

新闻数据分类

一、实验目的

1. 对新闻数据进行适当预处理和表示；
2. 通过降维和可视化对新闻数据的分布进行初步探索；
3. 对新闻数据进行分类，对比不同分类算法、不同表示模型对分类结果的有效性。

二、实验内容

1.新闻数据预处理和词袋表示

a. 加载 20newsgroups 数据集，并打印查看其中一个文档（如果太长可以打印部分）。

b. 针对以上数据集特点，适当调整和完善预处理函数 EngPreprocess(),对 20newsgroups 原始数据进行恰当的预处理，并将预处理后的文档保存到文件；请对每个预处理步骤进行具体描述。

c. 给出预处理之后的数据统计信息，包括文档数目、词的数目、文档的平均长度等。

d. 读入预处理完的文档数据，进行词袋表示，得到基于词频的和 TF-IDF 的两种数据矩阵。

2.降维和可视化

a.对基于 TF-IDF 的数据矩阵进行 PCA 降维，得到二维表示。

b.打印二维表示散点图（用不同颜色代表真实类别标签），观察并讨论该新闻数据的分布情况。

注意：如果整个数据集太大，可以选择一个子集（比如选择 5 个类别，或者选择更多类别但是每个类别包含更少文档）进行降维和可视化操作。

3.用朴素贝叶斯算法对新闻数据进行分类

e. 用朴素贝叶斯算法对基于 Part1 中得到的词频表示的新闻数据进行分类，得到 5 折交叉验证的准确率（Accuracy）。

f. 改用基于 TF-IDF 表示的新闻数据，重复 a 中实验，对比准确

率。

g. 对原始数据（不进行预处理）进行 TF-IDF 表示，重复 a 中实验，对比步骤 b 的准确率，讨论预处理的作用。

h. 改变 min_df 和 max_df 这两个参数（具体含义）的值得到 TF-IDF 表示，重复 a 中实验，对比步骤 b 的准确率并讨论。

4.不同算法有效性和时间的对比

a. 基于预处理后的 TF-IDF 表示，用支持向量机对其进行分类，给出 5 折交叉验证的准确率。

b. 对比朴素贝叶斯和支持向量机的训练和测试时间。综合准确率和两个时间讨论：实际应用中如何对算法进行选择。