实验一　文本数据的分类与分析

【实验目的】

1. 掌握数据预处理的方法，对训练集数据进行预处理；
2. 掌握文本建模的方法，对语料库的文档进行建模；
3. 掌握分类算法的原理，基于有监督的机器学习方法，训练文本分类器；
4. 利用学习的文本分类器，对未知文本进行分类判别；
5. 掌握评价分类器性能的评估方法。

【实验类型】

数据挖掘算法的设计与编程实现。

【实验要求】

1. 文本类别数：>=10类；
2. 训练集文档数：>=500000篇；每类平均50000篇。
3. 测试集文档数：>=500000篇；每类平均50000篇。
4. 分组完成实验，组员数量<=3，个人实现可以获得实验加分。

【实验内容】

利用分类算法实现对文本的数据挖掘，主要包括：

1. 语料库的构建，主要包括利用爬虫收集Web文档等；
2. 语料库的数据预处理，包括文档建模，如去噪，分词，建立数据字典，使用词袋模型或主题模型表达文档等；

   注：使用主题模型，如LDA可以获得实验加分；
3. 选择分类算法（朴素贝叶斯（必做）、SVM/其他等），训练文本分类器，理解所选的分类算法的建模原理、实现过程和相关参数的含义；
4. 对测试集的文本进行分类
5. 对测试集的分类结果利用正确率和召回率进行分析评价：计算每类正确率、召回率，计算总体正确率和召回率，以及F-score。

【实验验收】

1. 编写实验报告，实验报告内容必须包括对每个阶段的过程描述，以及实验结果的截图展示。
2. 以现场方式验收实验代码。
3. 实验完成时间11月24日.