1.题目：文本分析练习。

2.数据集：从 UCI Machine Learning Repository、Kaggle、Git Hub、Gitee 等数据源，自行选择一个英文数据集、一个中文数据集。

比如 20 news groups 英文数据集（该数据集为 Multiclass Text Classification 数据集）、餐馆/酒店点评中文数据集。

[1] 中文酒店评论数据集. https://zhuanlan.zhihu.com/p/506382415

3. 任务描述：

准备英文数据集；

准备中文数据集；

针对上述 2 个数据集，完成如下任务。

（1）数据集装载、显示；

（2）分词；

（3）全集、分类别子集的词云可视化；

（4）分类模型（多种分类算法，比如朴素贝叶斯分类、SVM 分类、决策树、KNN 等）；

（5）分类效果评价（accuracy, precision, recall）；

（6）利用模型，对单一文档进行分类。

4. 成果提交：请于截止时间前，把数据与源代码（*.ipynb）一起压缩为 zip 文件，提交到 obe.ruc.edu.cn 网站。

备注：notebook 的每个代码 Cell，要求加上必要的注释。

5. 参考文献：无

6.评价方法：

（1）本练习为必选练习，根据上述任务完成度评分。

（2）部分完成、晚交适当扣分。

（3）自行完成，不要抄袭，抄袭 0 分。