《统计计算》大作业

题目：基于 LDA 的文本分类

1. 数据集

数据集：IFLYTEK' 长文本分类

地址：https://www.cluebenchmarks.com/introduce.html

描述：该数据集共有1.7万多条关于app应用描述的长文本标注数据，包含和日常生活相关的各类应用主题，共119个类别："打车":0,"地图导航":1,"免费WIFI":2,"租车":3,…,"女性":115,"经营":116,"收款":117,"其他":118(分别用0-118表示)。每一条数据有三个属性，从前往后分别是 类别ID，类别名称，文本内容。

> 数据量：训练集(12,133)，验证集(2,599)，测试集(2,600)
>
> 例子：
>
> {"label": "110", "label_des": "社区超市", "sentence": "朴朴快送超市创立于2016年，专注于打造移动端30分钟即时配送一站式购物平台，商品品类包含水果、蔬菜、肉禽蛋奶、海鲜水产、粮油调味、酒水饮料、休闲食品、日用品、外卖等。朴朴公司希望能以全新的商业模式，更高效快捷的仓储配送模式，致力于成为更快、更好、更多、更省的在线零售平台，带给消费者更好的消费体验，同时推动中国食品安全进程，成为一家让社会尊敬的互联网公司。,朴朴一下，又好又快,1.配送时间提示更加清晰友好2.保障用户隐私的一些优化3.其他提高使用体验的调整4.修复了一些已知bug"}

2. Requirements

从数据集中选取前 10 个类别, 每个类别 100 篇文档, 用 LDA 主题模型进行文本分类;

注: 可使用 gensim, sklearn 等第三方开源库; 工具方法不做限制;

3. 考核形式

3人一组, 于 6 月 11 日 提交分组名单;

提交 Report 和 源码;

4. 评分标准

1. 分类准确率 (70%)
2. Report (30%)
    a. 整体设计 (30%)
    b. 算法说明 (包括步骤或代码注释等等, 30%)
    c. 预处理 (20%)
    d. 结果分析与可视化 (20%)

Dealine: 25th June