文本分析与数据挖掘期末作业

期末作业总体要求

40个人分为10组，每组4人，每组需完成3个任务，前两个为必完成，第三个为可选。

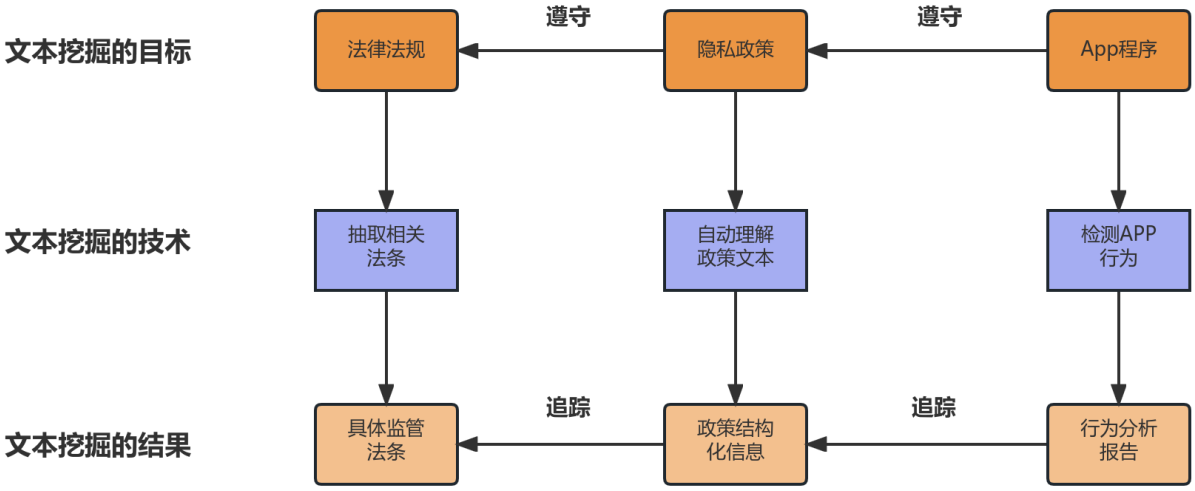
分组代码如下

```
with open('./stu_names.txt') as f:
    stu_name_str = f.read()

stu_name_list = stu_name_str.strip().split()

assert len(set(stu_name_list)) == 40

import random

random.seed('text_mining_2023')

random.shuffle(stu_name_list)

groups = [stu_name_list[n:n+4] for n in range(0,
len(set(stu_name_list)), 4)]

for i, group in enumerate(groups, start=1):
    print('第{}组: {}'.format(i, group))
```

分组结果见课程钉钉群

期末文本挖掘项目整体框架如下所示

不忘专业的初心（管理与应用），牢记课程的使命（文本分析与数据挖掘）



期末作业框架图

每个小组的具体任务

1. 必做，从下面法律、法规中抽取监管App隐私数据实践的 具体法条

- 《民法典》
- 《个人信息保护法》
- 《数据安全法》
- 《网络安全法》
- 《电信和互联网用户个人信息保护规定》

  提交 隐私法规条款.xlsx 文件，最后一列请填写一条句子，如果有多条句子，那么就需要填写多行结果。

2. 必做-1-无监督文本挖掘，对隐私政策进行聚类或者主题分析，对于主题或类簇的命名和解释可参考（但不限于）第一步的结果。

- 提交 无监督文本挖掘.ipynb 文件，注意，文件中需要包含源代码和对结果的文字解释。

2. 必做-2-有监督文本挖掘，结合聚类或者主题分析的结果，以句子为单位标注隐私政策文本的类别标签（收集和使用、共享、保存、其他），为每个类别标注200条数据，共计800条数据，从800条数据中随机采样600条数据作为训练集，剩下200条作为测试集，确保训练集和测试集中均包含5个类别的数据，训练文本分类模型，并在测试集上测试模型的性能，对测试结果进行分析。

- 提交 有监督文本挖掘.ipynb 文件，文件中需要包含源代码和对结果的文字解释；
- 提交 隐私政策语句分类标注.xlsx 文件，最后一列请填写一条句子，如果有多条句子，那么就需要填写多行结果。

3. 选做-分析App行为与隐私政策声明的一致性，选取本小组获得的App样本，至少选取三个，从豌豆荚（https://www.wandoujia.com/ ） 下载其安卓APK文件，借助在线静态分析工具（https://mobsf.live/ ） ， 进行软件包静态分析，依据静态分析结果报告，识别App在个人信息 收集和使用 行为方面与隐私政策声明存在的不一致。比如，政策未声明收集和使用某些信息，但实际上App收集和使用了这些信息；政策只是宽泛地声明收集和使用个人信息，但App实际上收集和使用了许多具体的个人信息；政策声明App不会收集和使用某些信息，但实际上App实际上收集和使用了这些信息，不限于上述三种情况。

- 提交 App个人信息收集和使用行为不一致分析.xlsx ；
- App的静态分析报告PDF版本。

提交目录清单

(1) 小组成员分工.txt

(2) 隐私法规条款.xlsx

(3) 无监督文本挖掘.ipynb

(4) 有监督文本挖掘.ipynb

(5) 隐私政策语句分类标注.xlsx

(6) App个人信息收集和使用行为不一致分析.xlsx （可选）

(7) App的静态分析报告PDF版本（可选）

主要评分标准

(1) 文本预处理是否完善；

(2) 无监督文本挖掘算法应用和结果呈现、分析；

(3) 隐私法规条款列举的完整情况、隐私政策文本数据标注质量；

(4) 有监督文本挖掘算法应用和结果呈现、分析。