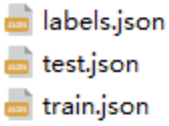
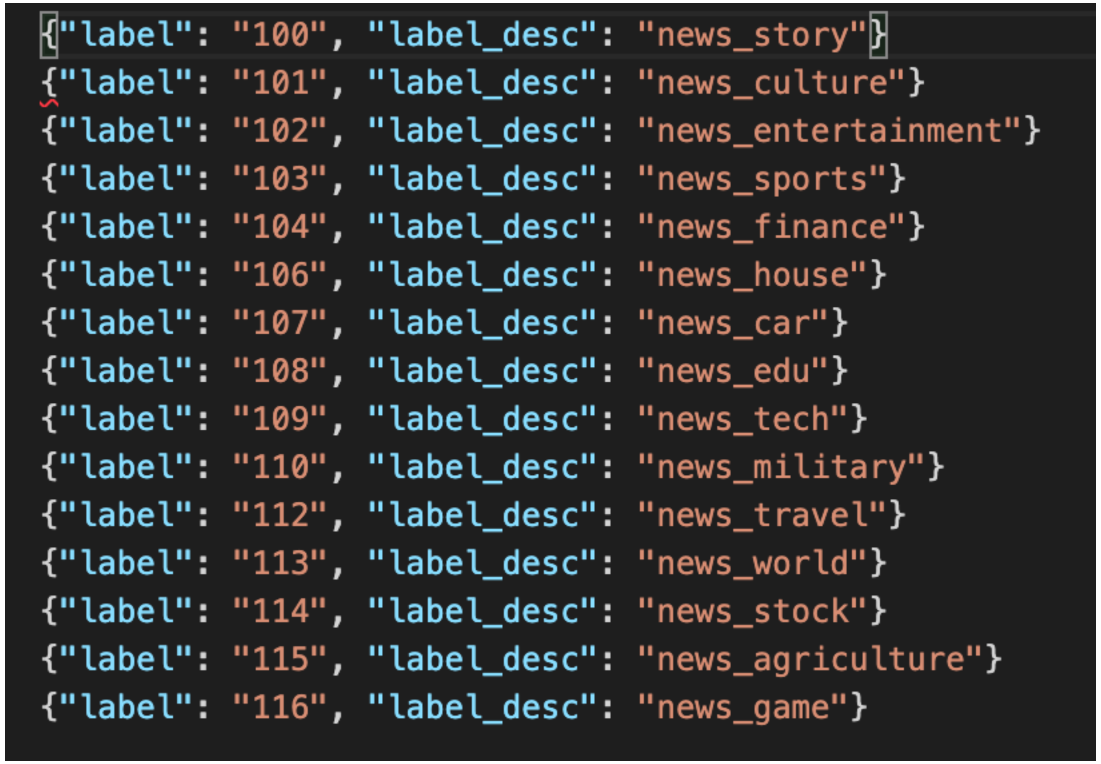
实验2
Experiment 2

文本数据挖掘

实验数据



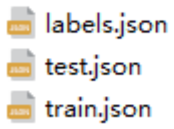
labels.json
test.json
train.json



```
{"label": "100", "label_desc": "news_story"}
{"label": "101", "label_desc": "news_culture"}
{"label": "102", "label_desc": "news_entertainment"}
{"label": "103", "label_desc": "news_sports"}
{"label": "104", "label_desc": "news_finance"}
{"label": "106", "label_desc": "news_house"}
{"label": "107", "label_desc": "news_car"}
{"label": "108", "label_desc": "news_edu"}
{"label": "109", "label_desc": "news_tech"}
{"label": "110", "label_desc": "news_military"}
{"label": "112", "label_desc": "news_travel"}
{"label": "113", "label_desc": "news_world"}
{"label": "114", "label_desc": "news_stock"}
{"label": "115", "label_desc": "news_agriculture"}
{"label": "116", "label_desc": "news_game"}
```

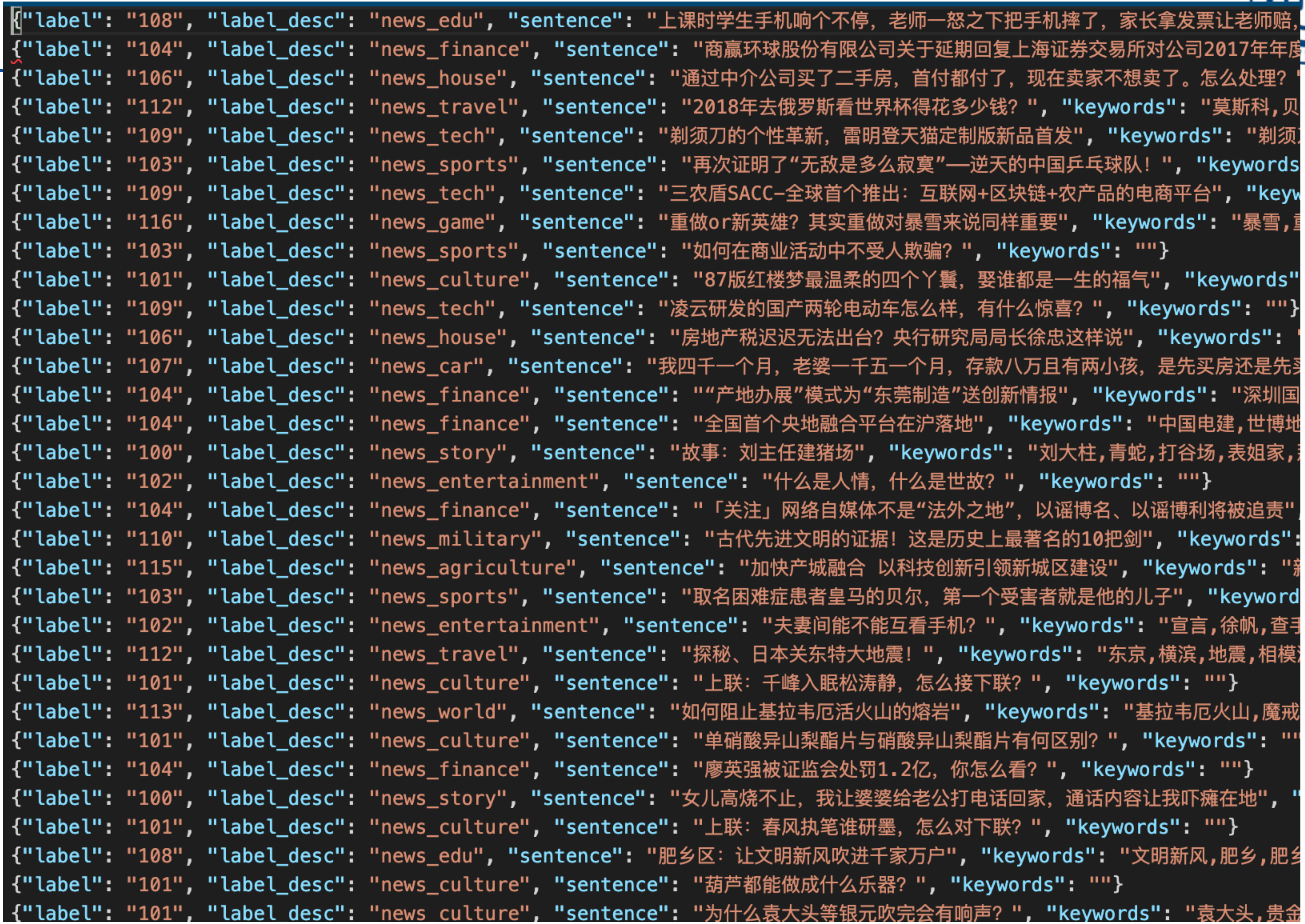
labels.json存放的是类别序号以及类别的描述信息

实验数据



```
labels.json
test.json
train.json
```

train.json和test.json给出了所属的类别以及sentence文本数据和keywords关键词



```
{"label": "108", "label_desc": "news_edu", "sentence": "上课时学生手机响个不停，老师一怒之下把手机摔了，家长拿发票让老师赔，
{"label": "104", "label_desc": "news_finance", "sentence": "商赢环球股份有限公司关于延期回复上海证券交易所对公司2017年年
{"label": "106", "label_desc": "news_house", "sentence": "通过中介公司买了二手房，首付都付了，现在卖家不想卖了。怎么处理？"
{"label": "112", "label_desc": "news_travel", "sentence": "2018年去俄罗斯看世界杯得花多少钱？", "keywords": "莫斯科,贝
{"label": "109", "label_desc": "news_tech", "sentence": "剃须刀的个性革新，雷明登天猫定制版新品首发", "keywords": "剃须
{"label": "103", "label_desc": "news_sports", "sentence": "再次证明了“无敌是多么寂寞”——逆天的中国乒乓球队！", "keywords
{"label": "109", "label_desc": "news_tech", "sentence": "三农盾SACC-全球首个推出：互联网+区块链+农产品的电商平台", "keyw
{"label": "116", "label_desc": "news_game", "sentence": "重做or新英雄？其实重做对暴雪来说同样重要", "keywords": "暴雪,
{"label": "103", "label_desc": "news_sports", "sentence": "如何在商业活动中不受人欺骗？", "keywords": ""}
{"label": "101", "label_desc": "news_culture", "sentence": "87版红楼梦最温柔的四个丫鬟，娶谁都是一生的福气", "keywords"
{"label": "109", "label_desc": "news_tech", "sentence": "凌云研发的国产两轮电动车怎么样，有什么惊喜？", "keywords": ""}
{"label": "106", "label_desc": "news_house", "sentence": "房地产税迟迟无法出台？央行研究局局长徐忠这样说", "keywords": "
{"label": "107", "label_desc": "news_car", "sentence": "我四千一个月，老婆一千五一个月，存款八万且有两小孩，是先买房还是先
{"label": "104", "label_desc": "news_finance", "sentence": "“产地办展”模式为“东莞制造”送创新情报", "keywords": "深圳国
{"label": "104", "label_desc": "news_finance", "sentence": "全国首个央地融合平台在沪落地", "keywords": "中国电建,世博地
{"label": "100", "label_desc": "news_story", "sentence": "故事：刘主任建猪场", "keywords": "刘大柱,青蛇,打谷场,表姐家,
{"label": "102", "label_desc": "news_entertainment", "sentence": "什么是人情，什么是世故？", "keywords": ""}
{"label": "104", "label_desc": "news_finance", "sentence": "「关注」网络自媒体不是“法外之地”，以谣博名、以谣博利将被追责",
{"label": "110", "label_desc": "news_military", "sentence": "古代先进文明的证据！这是历史上最著名的10把剑", "keywords":
{"label": "115", "label_desc": "news_agriculture", "sentence": "加快产城融合 以科技创新引领新城区建设", "keywords": "
{"label": "103", "label_desc": "news_sports", "sentence": "取名困难症患者皇马的贝尔，第一个受害者就是他的儿子", "keyword
{"label": "102", "label_desc": "news_entertainment", "sentence": "夫妻间能不能互看手机？", "keywords": "宣言,徐帆,查
{"label": "112", "label_desc": "news_travel", "sentence": "探秘、日本关东特大地震！", "keywords": "东京,横滨,地震,相模
{"label": "101", "label_desc": "news_culture", "sentence": "上联：千峰入眠松涛静，怎么接下联？", "keywords": ""}
{"label": "113", "label_desc": "news_world", "sentence": "如何阻止基拉韦厄活火山的熔岩", "keywords": "基拉韦厄火山,魔戒
{"label": "101", "label_desc": "news_culture", "sentence": "单硝酸异山梨酯片与硝酸异山梨酯片有何区别？", "keywords": ""
{"label": "104", "label_desc": "news_finance", "sentence": "廖英强被证监会处罚1.2亿，你怎么看？", "keywords": ""}
{"label": "100", "label_desc": "news_story", "sentence": "女儿高烧不止，我让婆婆给老公打电话回家，通话内容让我吓瘫在地", "
{"label": "101", "label_desc": "news_culture", "sentence": "上联：春风执笔谁研墨，怎么对下联？", "keywords": ""}
{"label": "108", "label_desc": "news_edu", "sentence": "肥乡区：让文明新风吹进千家万户", "keywords": "文明新风,肥乡,肥
{"label": "101", "label_desc": "news_culture", "sentence": "葫芦都能做成什么乐器？", "keywords": ""}
{"label": "101", "label_desc": "news_culture", "sentence": "为什么袁大头等银元吹完会有响声？", "keywords": "袁大头,贵
```

实验要求

实验一：在给定的数据集上，将文本数据分类成不同类别。

- 数据使用：只能使用文本数据（使用sentence，不用keywords）；
- 模型选择：只能使用CNN变种，比如TextCNN；
- 数据要求：数据处理时可以使用相应的工具，如jieba，可以使用数据增强方法，但不能引入新数据，也不能使用其他的数据集；

实验要求

实验二：同时使用文本数据和关键词的文本分类

- 数据使用：同时使用sentence和keywords；
- 模型选择：只能使用CNN变种，比如TextCNN；
- 数据要求：数据处理时可以使用相应的工具，如jieba，可以使用数据增强方法，但不能引入新数据，也不能使用其他的数据集；

开发环境

主要语言为python3
开发环境pycharm、vscode、jupyter notebook都可以

参考资料

TextCNN：
(2014 EMNLP) Convolutional Neural Networks for Sentence Classification
https://arxiv.org/abs/1408.5882

文本分类的一些算法：
https://zhuanlan.zhihu.com/p/76003775