exp 3-NLP在金融场景下的应用

文件预处理

先将合并的连接词进行处理，除去前后的空格：

```python
data = pd.read_csv("nlp_dataset/business_scope.csv")
label = pd.read_csv("nlp_dataset/index_member.csv")
for tc in data.ts_code:
    tc = tc.strip()
for cc in label.con_code:
    cc = cc.strip()
```

将两个文件按“ts_code”进行连接：

```python
label = label.rename(columns={"con_code":"ts_code"})
da = pd.merge(data, label, on="ts_code")
```

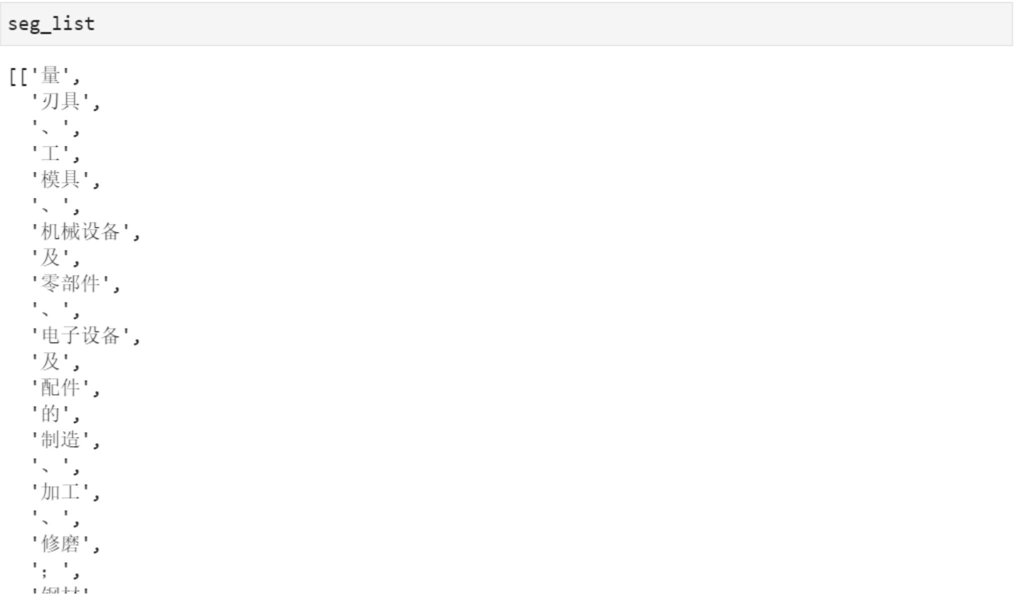
数据预处理

分词

将市场范围文本进行逐行分词，内容存储在一个list中：

```python
N = len(da.business_scope)
seg_list = []
for i in range(N):
    seg_list.append((jieba.lcut(da.business_scope[i])))
```

内容如下：



```
seg_list
```

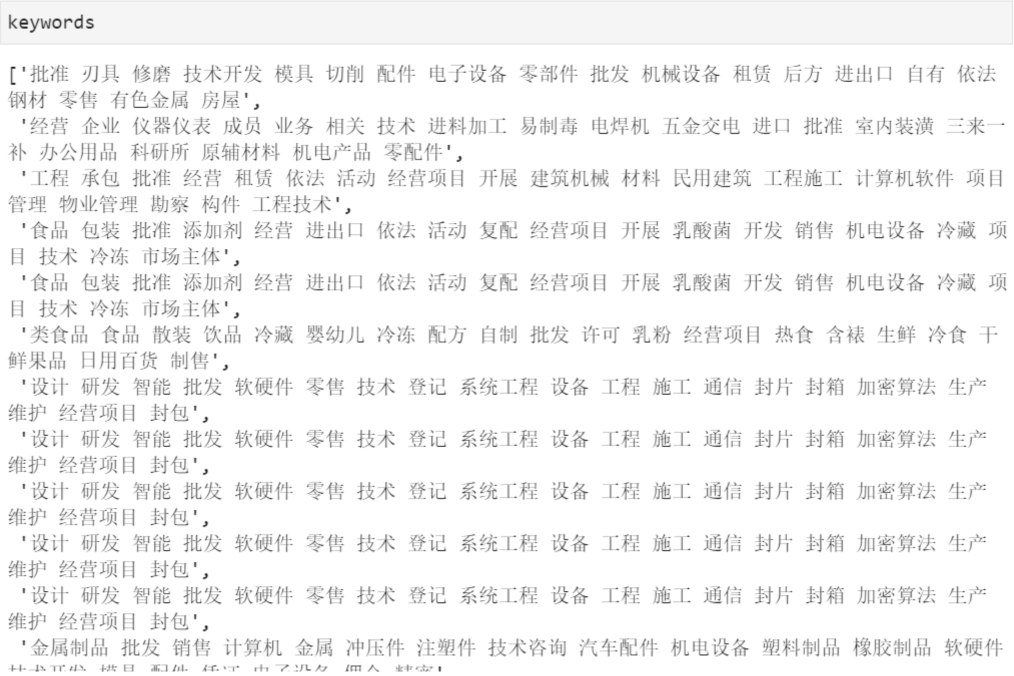
```
[['量',
  '刃具',
  '、',
  '工',
  '模具',
  '、',
  '机械设备',
  '及',
  '零部件',
  '、',
  '电子设备',
  '及',
  '配件',
  '的',
  '制造',
  '、',
  '加工',
  '、',
  '修磨',
  '；',
```

关键词提取

将市场范围文本进行逐行提取关键词，内容存在keywords的list中：

```python
keywords = []
for i in range(N):
    keywords.append(analyse.extract_tags(da.business_scope[i]))
```

内容如下：



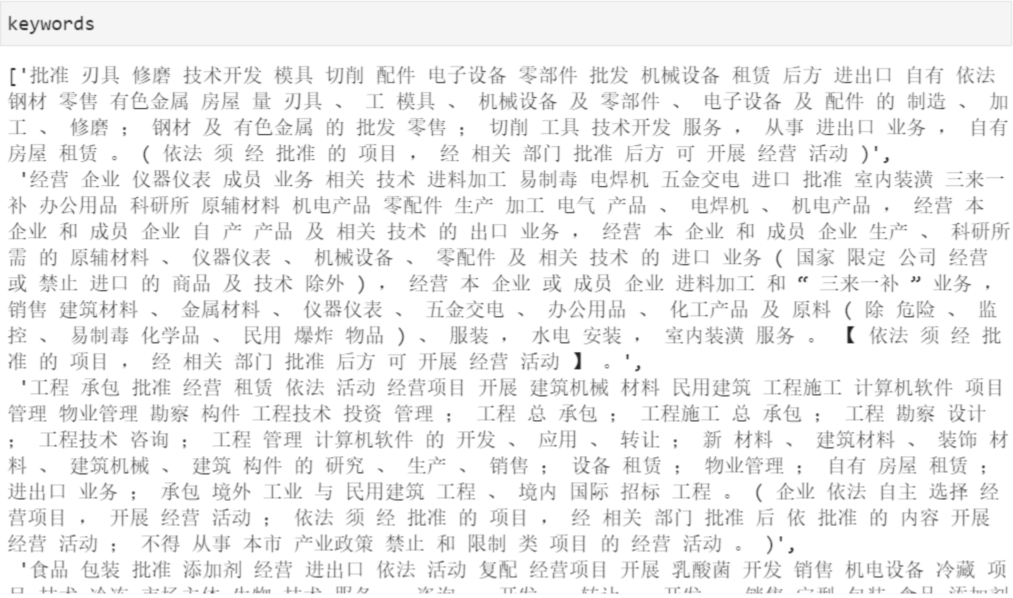
```
keywords
```

```
['批准 刃具 修磨 技术开发 模具 切削 配件 电子设备 零部件 批发 机械设备 租赁 后方 进出口 自有 依法
钢材 零售 有色金属 房屋',
 '经营 企业 仪器仪表 成员 业务 相关 技术 进料加工 易制毒 电焊机 五金交电 进口 批准 室内装潢 三来一
补 办公用品 科研所 原辅材料 机电产品 零配件',
 '工程 承包 批准 经营 租赁 依法 活动 经营项目 开展 建筑机械 材料 民用建筑 工程施工 计算机软件 项目
管理 物业管理 勘察 构件 工程技术',
 '食品 包装 批准 添加剂 经营 进出口 依法 活动 复配 经营项目 开展 乳酸菌 开发 销售 机电设备 冷藏 项
目 技术 冷冻 市场主体',
 '食品 包装 批准 添加剂 经营 进出口 依法 活动 复配 经营项目 开展 乳酸菌 开发 销售 机电设备 冷藏 项
目 技术 冷冻 市场主体',
 '类食品 食品 散装 饮品 冷藏 婴幼儿 冷冻 配方 自制 批发 许可 乳粉 经营项目 热食 含裱 生鲜 冷食 干
鲜果品 日用百货 制售',
 '设计 研发 智能 批发 软硬件 零售 技术 登记 系统工程 设备 工程 施工 通信 封片 封箱 加密算法 生产
维护 经营项目 封包',
 '设计 研发 智能 批发 软硬件 零售 技术 登记 系统工程 设备 工程 施工 通信 封片 封箱 加密算法 生产
维护 经营项目 封包',
 '设计 研发 智能 批发 软硬件 零售 技术 登记 系统工程 设备 工程 施工 通信 封片 封箱 加密算法 生产
维护 经营项目 封包',
 '设计 研发 智能 批发 软硬件 零售 技术 登记 系统工程 设备 工程 施工 通信 封片 封箱 加密算法 生产
维护 经营项目 封包',
 '设计 研发 智能 批发 软硬件 零售 技术 登记 系统工程 设备 工程 施工 通信 封片 封箱 加密算法 生产
维护 经营项目 封包',
 '金属制品 批发 销售 计算机 金属 冲压件 注塑件 技术咨询 汽车配件 机电设备 塑料制品 橡胶制品 软硬件
```

将分词结果和关键词串联

```python
for i in range(N):
    keywords[i] = keywords[i]+seg_list[i]
```

内容如下：



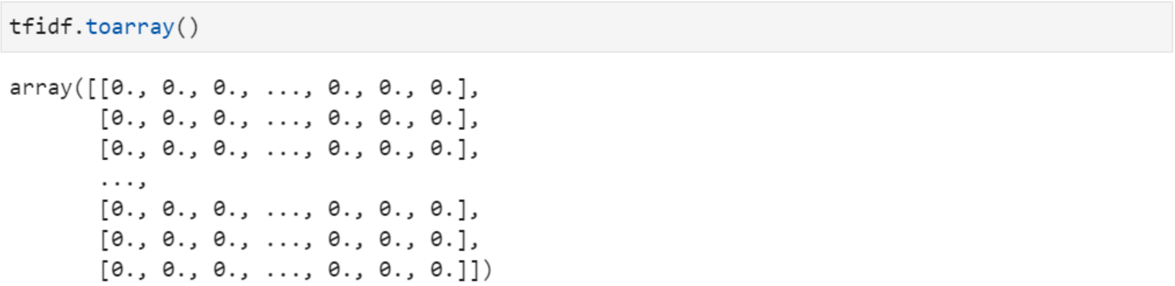
```
keywords
```

```
['批准 刃具 修磨 技术开发 模具 切削 配件 电子设备 零部件 批发 机械设备 租赁 后方 进出口 自有 依法
钢材 零售 有色金属 房屋 量 刃具 、 工 模具 、 机械设备 及 零部件 、 电子设备 及 配件 的 制造 、 加
工 、 修磨 ； 钢材 及 有色金属 的 批发 零售 ； 切削 工具 技术开发 服务 ， 从事 进出口 业务 ， 自有
房屋 租赁 。 ( 依法 须 经 批准 的 项目 ， 经 相关 部门 批准 后方 可 开展 经营 活动 )',
 '经营 企业 仪器仪表 成员 业务 相关 技术 进料加工 易制毒 电焊机 五金交电 进口 批准 室内装潢 三来一
补 办公用品 科研所 原辅材料 机电产品 零配件 生产 加工 电气 产品 、 电焊机 、 机电产品 ， 经营 本
企业 和 成员 企业 自 产 产品 及 相关 技术 的 出口 业务 ， 经营 本 企业 和 成员 企业 生产 、 科研所
需 的 原辅材料 、 仪器仪表 、 机械设备 、 零配件 及 相关 技术 的 进口 业务 ( 国家 限定 公司 经营
或 禁止 进口 的 商品 及 技术 除外 ) ， 经营 本 企业 或 成员 企业 进料加工 和 “ 三来一补 ” 业务 ，
销售 建筑材料 、 金属材料 、 仪器仪表 、 五金交电 、 办公用品 、 化工产品 及 原料 ( 除 危险 、 监
控 、 易制毒 化学品 、 民用 爆炸 物品 ) 、 服装 ， 水电 安装 ， 室内装潢 服务 。 【 依法 须 经 批
准 的 项目 ， 经 相关 部门 批准 后方 可 开展 经营 活动 】 。',
 '工程 承包 批准 经营 租赁 依法 活动 经营项目 开展 建筑机械 材料 民用建筑 工程施工 计算机软件 项目
管理 物业管理 勘察 构件 工程技术 投资 管理 ； 工程 总 承包 ； 工程施工 总 承包 ； 工程 勘察 设计
； 工程技术 咨询 ； 工程 管理 计算机软件 的 开发 、 应用 、 转让 ； 新 材料 、 建筑材料 、 装饰 材
料 、 建筑机械 、 建筑 构件 的 研究 、 生产 、 销售 ； 设备 租赁 ； 物业管理 ； 自有 房屋 租赁 ；
进出口 业务 ； 承包 境外 工业 与 民用建筑 工程 、 境内 国际 招标 工程 。 ( 企业 依法 自主 选择 经
营项目 ， 开展 经营 活动 ； 依法 须 经 批准 的 项目 ， 经 相关 部门 批准 后 依 批准 的 内容 开展
经营 活动 ； 不得 从事 本市 产业政策 禁止 和 限制 类 项目 的 经营 活动 。 )',
 '食品 包装 批准 添加剂 经营 进出口 依法 活动 复配 经营项目 开展 乳酸菌 开发 销售 机电设备 冷藏 项
```

词频向量化，TF-IDF处理

```
vectorizer = CountVectorizer()
vctf = TfidfVectorizer()
vc_fit = vectorizer.fit_transform(keywords)
tfidf = vctf.fit_transform(keywords)
```

结果如下：



```
tfidf.toarray()
```

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

分类器学习

训练集，测试集划分

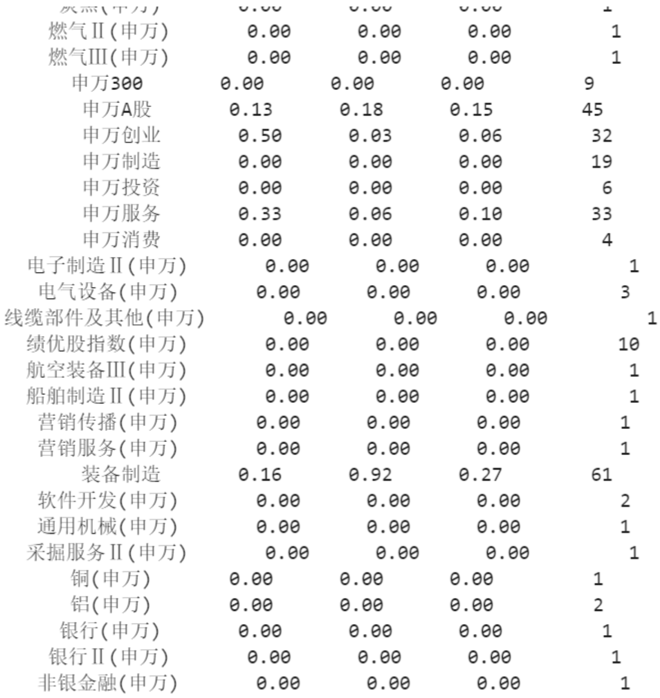
利用KFold将数据集进行10次交叉验证划分，随机打乱。

```
kf = KFold(n_splits=10, shuffle=True, random_state=0)
```

调用朴素贝叶斯分类器进行分类，标签为index_name中的数据。调用classification_report对分类器效果进行评估。

```
clf = MultinomialNB()
label = da.index_name
for train_index, test_index in kf.split(tfidf):
    clf = MultinomialNB().fit(tfidf[train_index], label[train_index])
    y_pred = clf.predict(tfidf[test_index])
    print(classification_report(label[test_index], y_pred))
```

结果如下：



```
      炭黑(申万)       0.00      0.00      0.00         1
     燃气Ⅱ(申万)       0.00      0.00      0.00         1
     燃气Ⅲ(申万)       0.00      0.00      0.00         1
        申万300       0.00      0.00      0.00         9
        申万A股       0.13      0.18      0.15        45
        申万创业       0.50      0.03      0.06        32
        申万制造       0.00      0.00      0.00        19
        申万投资       0.00      0.00      0.00         6
        申万服务       0.33      0.06      0.10        33
        申万消费       0.00      0.00      0.00         4
  电子制造Ⅱ(申万)       0.00      0.00      0.00         1
    电气设备(申万)       0.00      0.00      0.00         3
线缆部件及其他(申万)       0.00      0.00      0.00         1
  绩优股指数(申万)       0.00      0.00      0.00        10
  航空装备Ⅲ(申万)       0.00      0.00      0.00         1
  船舶制造Ⅱ(申万)       0.00      0.00      0.00         1
    营销传播(申万)       0.00      0.00      0.00         1
    营销服务(申万)       0.00      0.00      0.00         1
        装备制造       0.16      0.92      0.27        61
    软件开发(申万)       0.00      0.00      0.00         2
    通用机械(申万)       0.00      0.00      0.00         1
  采掘服务Ⅱ(申万)       0.00      0.00      0.00         1
        铜(申万)       0.00      0.00      0.00         1
        铝(申万)       0.00      0.00      0.00         2
      银行(申万)       0.00      0.00      0.00         1
     银行Ⅱ(申万)       0.00      0.00      0.00         1
    非银金融(申万)       0.00      0.00      0.00         1
```

如图所示。

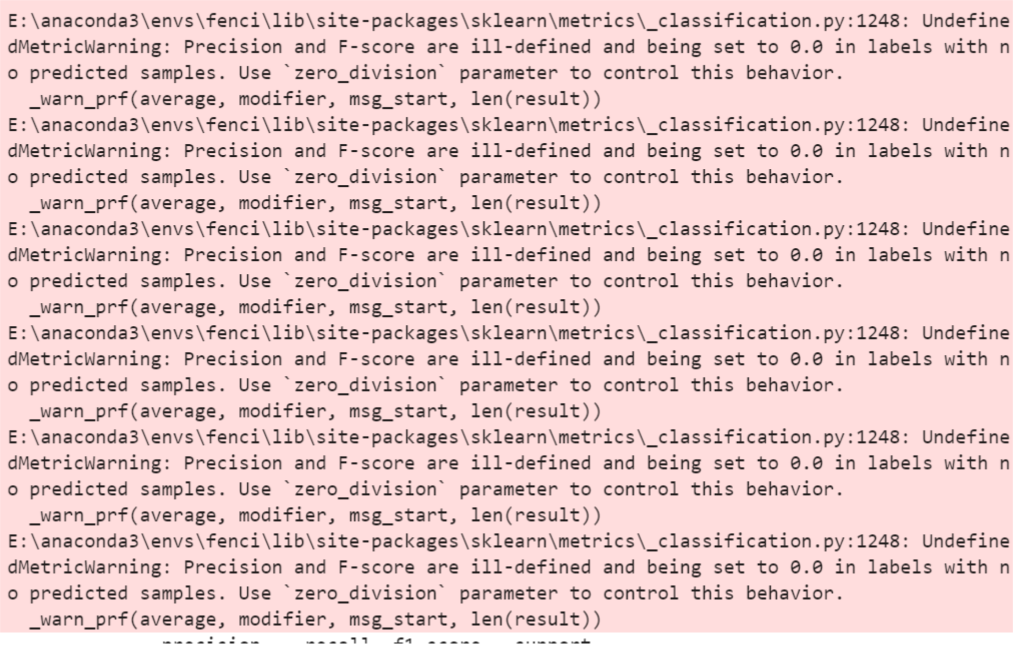
但是测试得到的各项数据几乎都为0，只有少数行有非0数据显示，原因不明。可能是“申万A股”，“申万创业”，“申万服务”，“装备制造”的数据量大，学习效果好，其余数据量太小(只有1,2,10,...)。也可能是实验中某些步骤出现错误。但是由于调用库函数，同时也对库函数的使用以及nlp数据处理的不熟悉，调试很久也只能得到类似以上的结果。

结果分析

如图所示，总量范围大于30的类均有数据显示，第一列为精准度，第二列为召回率，第三列为F1-score，分类结果中，精准度均不高，而召回率在数量大于60时得到了比较好的结果，而这三项指标在30~60数据量的范围内均表现出随着数据量增大而越大的趋势，故在该实验中改进方法有：1. 将数据有选择性地训练，抽取数据量大的类别进行分类，数据量小的分类效果不明显。2.加大样本数量，在实验中采用了10次交叉验证，但是效果很差，可能是数据量不够大。

这里列出实验过程中的几个疑问：

1. 分词所用函数应该是jieba.cut()还是jieba.lcut().
2. 分词后的词向量与关键词如何“串联”.
3. 分类器出现了如下warning，原因未明



```
E:\anaconda3\envs\fenci\lib\site-packages\sklearn\metrics\_classification.py:1248: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with n
o predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
E:\anaconda3\envs\fenci\lib\site-packages\sklearn\metrics\_classification.py:1248: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with n
o predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
E:\anaconda3\envs\fenci\lib\site-packages\sklearn\metrics\_classification.py:1248: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with n
o predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
E:\anaconda3\envs\fenci\lib\site-packages\sklearn\metrics\_classification.py:1248: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with n
o predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
E:\anaconda3\envs\fenci\lib\site-packages\sklearn\metrics\_classification.py:1248: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with n
o predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
E:\anaconda3\envs\fenci\lib\site-packages\sklearn\metrics\_classification.py:1248: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with n
o predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
              precision    recall  f1-score   support
```