实验 3：NLP 在金融场景下的应用实验指南

一、 实验目的

本实验旨在对金融文本数据进行自然语言处理并对结果进行评估分析。具体实验内容为对 TUSHARE 上市公司基本信息中的股票经营范围文本数据进行自然语言处理，并根据相应的行业分类标签进行文本分类，并评估分类结果。

二、 实验步骤

1. 数据获取

数据来源为 TUSHARE（开源的股票数据接口包）。
TUSHARE 官网：（https://tushare.pro）。
数据授权 token：

```
'a6dae538a760f0b9e39432c1bff5e50a1c462a1a087e994dae18fa04')
```

1. 首先确认 pandas 和 sklearn 是否已安装。如果没有，根据官网说明进行安装，推荐 Anaconda。

Pandas 安装说明：（https://pandas.pydata.org/pandas-docs/stable/install.html）。
sklearn 安装说明：（https://scikit-learn.org/stable/install.html）。

2. 获取经营范围文本数据
   根据经营范围文本数据接口说明与示例（ https://tushare.pro/document/2?doc_id=112 ）获得经营范围文本数据，其中 fields 参数选择股票代码“ ts_code ”和经营范围“business_scope”。
3. 获取行业分类标签数据
   根据行业分类标签数据接口说明与示例（http://tushare.org/classifying.html）获得行业分类标签数据。根据股票代码“code”，可将上一步中获得的股票的经营范围与行业名称相对应。
   1） 得到的数据中可能存在 NA 值，去掉这些值的参考工具如下：
   pandas dropna 参考文档：（https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dropna.html。
   2） 两种数据来源的股票代码名称和数据有差别，需进行修改统一。
   参考工具：
   pandas strip 参考文档：（https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.str.strip.html）；
   pandas rename 参考文档：（https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rename.html）。
   3） 将两种数据来源通过股票代码相结合，参考工具如下：
   pandas merge 参考文档：（https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.merge.html） 。
   需要注意两种数据 merge 操作后可能会产生新的 NA 值。

2. 将文本数据数值化

1. 根据官网说明进行“结巴”中文分词的安装。

“结巴”中文分词官方参考文档：（https://github.com/fxsjy/jieba）。

2. 利用“结巴”中文分词技术对经营范围文本数据进行分词。
3. 利用“结巴”中文分词技术对经营范围文本数据进行关键词提取。
4. 分词结果和关键词串联作为预处理后的文本数据。
5. 对预处理后的文本数据进行词频向量化，并进行 TF-IDF 处理得到文本数据数值化向量。（也可自主实现 TF-IDF 的算法）

参考工具：sklearn CountVectorizer。

CountVectorizer 参考文档：（https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html）。

参考工具：sklearn TfidfTransformer。

TfidfTransformer 参考文档：（https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html）。

3. 基于数值化文本向量进行分类器学习

1. 进行训练集和测试集的划分。

参考工具：sklearn KFold

KFold 参考文档：（https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html）。

2. 构建朴素贝叶斯多项式分类器。由于行业标签数量众多，可筛选出单类数据量大于 80 的类进行学习。

分类器参考工具：sklearn MultinomialNB。

MultinomialNB 参考文档：（https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html）。

3. 对分类器的效果进行评估，评价指标为 precision，recall，F1-score。

分类评价参考工具：sklearn classification_report

参考文档：（https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html）。

三、 提交内容

1. 实验报告：包含上述三项实验内容和结果分析。

2. 实验代码：提 交完整的实验代码。

3. 作业提交：统一提交到“学在浙大”