大数据原理与技术作业四

电影评论情感分类

【学号】22336259　　【姓名】谢宇桐　　【专业】计算机科学与技术

【实验要求】

- 使用IMDB影评数据集（或其他数据）
- 使用TF-IDF进行特征提取并用逻辑回归/SVM/RNN实现分类，分析准确率差异

【实验内容】

- 由于本地IDE环境配置太过麻烦，我使用kaggle进行实验，数据即为imdb-dataset-of-50k-movie-reviews。格式如下：

```
review    sentiment
One...    positive
There..   negative
...       ...
```

- TF-IDF（Term Frequency-Inverse Document Frequency，词频-逆文档频率）是一种用于信息检索与文本挖掘的常用加权技术。TF-IDF是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

```
# 初始化TF-IDF向量化器
tfidf_vectorizer = TfidfVectorizer(max_features=5000, stop_words='english')

# 拟合并转换训练数据
X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)

# 仅转换测试数据
X_test_tfidf = tfidf_vectorizer.transform(X_test)
```

逻辑回归（Logistic Regression）是一种广泛使用的分类算法，它的主要思想是将输入变量的线性组合映射到0到1之间的概率，用于预测二元输出变量的概率。

```
# 初始化逻辑回归模型
logistic_model = LogisticRegression()

# 训练模型
logistic_model.fit(X_train_tfidf, y_train)

# 预测测试集
y_pred_logistic = logistic_model.predict(X_test_tfidf)

# 计算准确率
accuracy_logistic = accuracy_score(y_test, y_pred_logistic)
print(f"逻辑回归模型的准确率: {accuracy_logistic:.4f}")
```

SVM：支持向量机（support vector machines），它将实例的特征向量映射为空间中的一些点，SVM 的目的就是想要画出一条线，以 “最好地” 区分这两类支持与反对的点。

```
# 初始化SVM模型
svm_model = SVC(kernel='linear')

# 训练模型
svm_model.fit(X_train_tfidf, y_train)

# 预测测试集
y_pred_svm = svm_model.predict(X_test_tfidf)

# 计算准确率
accuracy_svm = accuracy_score(y_test, y_pred_svm)
print(f"SVM模型的准确率: {accuracy_svm:.4f}")
```

运行结果：

```
逻辑回归模型的准确率: 0.8892
SVM模型的准确率: 0.8840
逻辑回归模型的准确率高于SVM模型。
```

可以看到这两个模型准确率相近，逻辑回归略高于SVM，且逻辑回归速度非常快，因为逻辑回归通常在高维稀疏数据（如TF-IDF向量化的文本数据）上表现良好。