基于Word2Vec的短新闻文本分类

一、实验目的

本实验旨在评估不同词向量维度、文档向量表示方法和分类器组合对中文短新闻分类任务性能的影响。实验目标是：

- 探索词向量维度（100维 vs 300维）对模型表现的影响；
- 比较文档向量表示方法（平均词向量 vs TF-IDF加权词向量）；
- 评估多种分类器（SVM、逻辑回归、随机森林）在文本表示下的效果。

二、实验设置

1. 数据预处理

- 本实验使用的数据集为短新闻分类数据集，该数据集包含真新闻、虚假新闻两个类别，二分类任务较为简单，适合使用SVM、逻辑回归、随机森林等传统机器学习方法。
- 使用结巴分词对文本进行分词；
- 去除停用词；
- 构建100、300维词向量（Word2Vec）；
- 分别使用两种方法将词向量转换为文档向量：
  - **mean**：取句中所有词向量的均值；
  - **tfidf**：使用TF-IDF作为词权重加权平均。

2. 模型配置

- **词向量维度**：100维和300维；
- **分类器**：
  - 支持向量机（SVM）
  - 逻辑回归（Logistic Regression）
  - 随机森林（Random Forest）

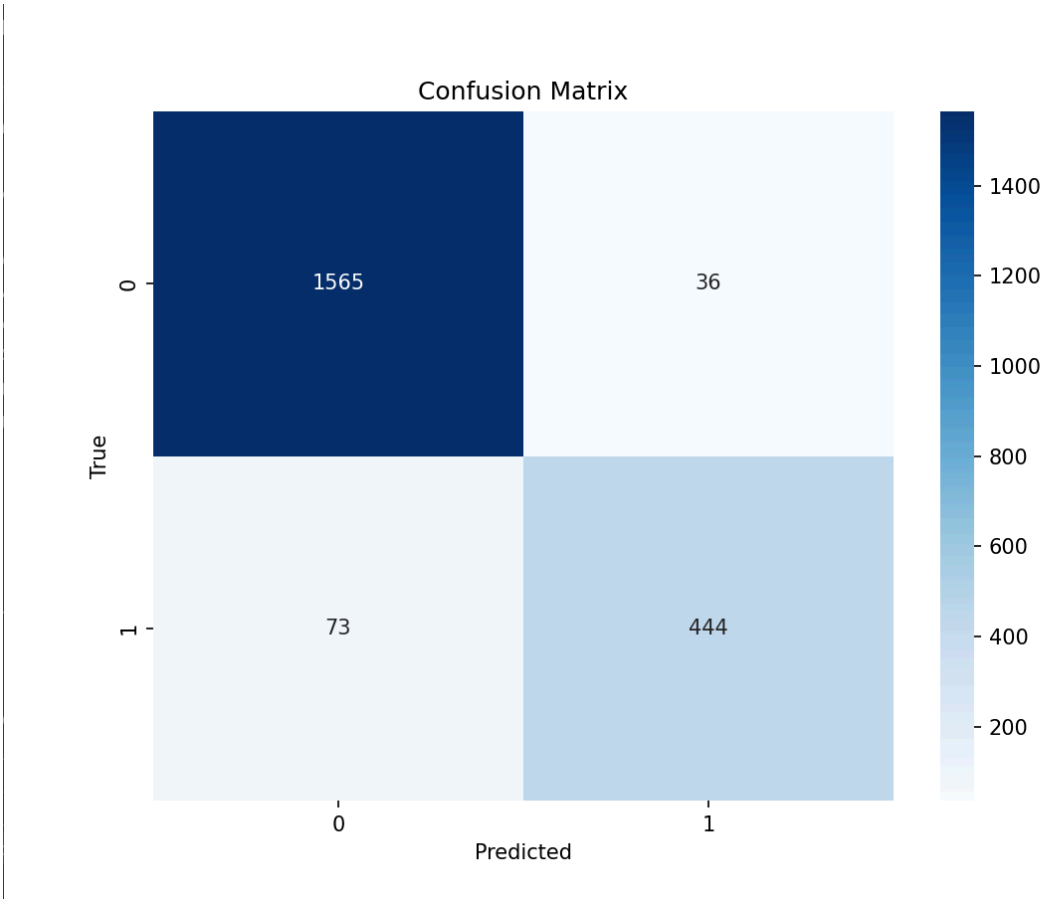
三、实验结果分析

下表为实验中的所有模型配置及其准确率（Accuracy）与加权F1得分（F1-Weighted）：

Rank	维度	表示方法	分类器	准确率 (Acc)	F1 分数
1	300	tfidf	RF	**0.9485**	**0.9479**
2	100	mean	RF	0.9448	0.9443
3	300	mean	RF	0.9443	0.9437
4	100	tfidf	RF	0.9438	0.9431
5	300	tfidf	SVM	0.8905	0.8860

Rank	维度	表示方法	分类器	准确率 (Acc)	F1 分数
6	300	tfidf	LogReg	0.8890	0.8853
7	100	tfidf	LogReg	0.8890	0.8849
8	100	tfidf	SVM	0.8890	0.8842
9	100	mean	SVM	0.8876	0.8828
10	100	mean	LogReg	0.8857	0.8815
11	300	mean	LogReg	0.8857	0.8813
12	300	mean	SVM	0.8853	0.8800

下图为混淆矩阵，在最优配置下，只有极少数文本分类错误。在分类错误的文本中，把假新闻预测成真新闻的较多。



关键观察与结论

1. **最佳配置**为：
    - 维度：300
    - 表示方法：TF-IDF 加权
    - 分类器：随机森林（Random Forest）
    - **准确率：94.85%，F1 分数：94.79%**
2. **随机森林优势明显**：

- 随机森林（RF）：在所有维度和表示方法下，随机森林始终优于其他分类器，尤其在 TF-IDF 表示下的表现显著高于 SVM 和逻辑回归。无论词向量维度或文档向量方法如何变化，RF的准确率始终高于94.3%，显著优于其他分类器。
  例如：dim=100时，mean方法下RF准确率为94.48%，而SVM仅88.76%。
- 使用随机森林时，最高准确率94.85%（dim=300, method=tfidf）。这是因为RF能自动处理高维稀疏特征，且对噪声和过拟合具有较强鲁棒性，适合短文本的复杂语义建模。
- SVM与逻辑回归：准确率均低于90%，与RF差距显著（约5-6%）。这是因为线性分类器（如SVM、逻辑回归）对特征分布的敏感度较高，可能难以捕捉短文本中的非线性关系。

3. **TF-IDF 表示略优于平均值表示**：
   - 在相同维度下，使用 TF-IDF 表示通常能获得更好的效果。这是因为 TF-IDF 可以强调关键信息词，而平均值方法对所有词一视同仁。
   - 当词向量维度为300时，TF-IDF加权使RF的准确率比简单平均提升0.42%（94.85% vs. 94.43%），而100维时仅提升0.57%（94.38% vs. 93.88%），表明高维词向量更需要加权策略。
4. **300维比100维略有优势**：
   - 在相同分类器和文档向量方法下，300维模型显著优于100维。更高维度的词向量能编码更丰富的语义信息，尤其对短文本的细粒度分类任务有益。高维词向量能更充分表达语义信息，但也对模型有更高计算要求。
5. **SVM 与逻辑回归性能相近，但整体弱于 RF**：
   - 在 TF-IDF 和mean表示下，两者效果接近。