## **作业 2 定制化词典**

#### 一、引言

在做文本挖掘的时候，首先要做的预处理就是分词，现代分词都是基于字典或统计的分词，而字典分词依赖于词典，即按某种算法构造词然后去匹配已建好的词典集合，如果匹配到就切分出来成为词语。通常词典分词被认为是最理想的中文分词算法。不论什么样的分词方法, 优秀的词典必不可少。

旅行文本具有鲜明的场景特征，包含大量地名、景点名称、旅行活动术语等专有词汇（如 "雷峰塔"" 青石板客栈 ""环岛骑行"）。这些词汇在通用词典中可能未被收录或权重较低，导致分词错误（如将 "环岛骑行" 切分为 "环岛 / 骑行" 虽正确，但作为整体术语处理更符合旅行文本特征）。

#### 二、定制化词典构建流程

##### 2.1 初始化词典

基于作业1的旅行日记语料，首先提取高频且具有旅行场景特征的词汇作为初始词典核心。初始化过程通过build\_initial\_dictionary函数实现，具体步骤如下：

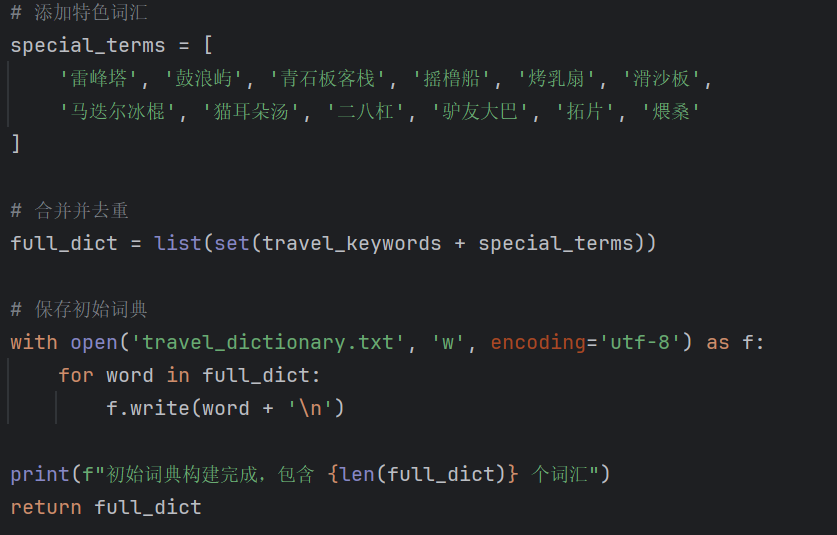
（1）文本清洗：去除日期格式和非中文字符，保留文本主体内容。

（2）分词与频率统计：使用 jieba 进行分词，统计词频并过滤长度≥2 且频率≥3 的词汇。

（3）特色词汇补充：手动添加旅行领域特色词汇，如 "雷峰塔"" 摇橹船 " 等。

核心代码：





初始化词典首批收录词汇共137个，存储于travel\_dictionary.txt文件。

##### 2.2 大规模语料收集与扩展

为丰富词典覆盖范围，从原始数据集中提取所有未被初始化词典收录的词汇，建立候选词库。通过collect\_candidate\_terms函数处理原始语料，步骤如下：

1. **语料来源**：原始数据集中的 12 篇旅行日记，总字数约 10000 字。
2. **文本处理**：去除日期格式和非中文字符，使用 jieba 分词。
3. **候选词提取**：保留长度≥2 的中文词汇，去重后得到 2277 个候选词。

核心代码：



实现结果：



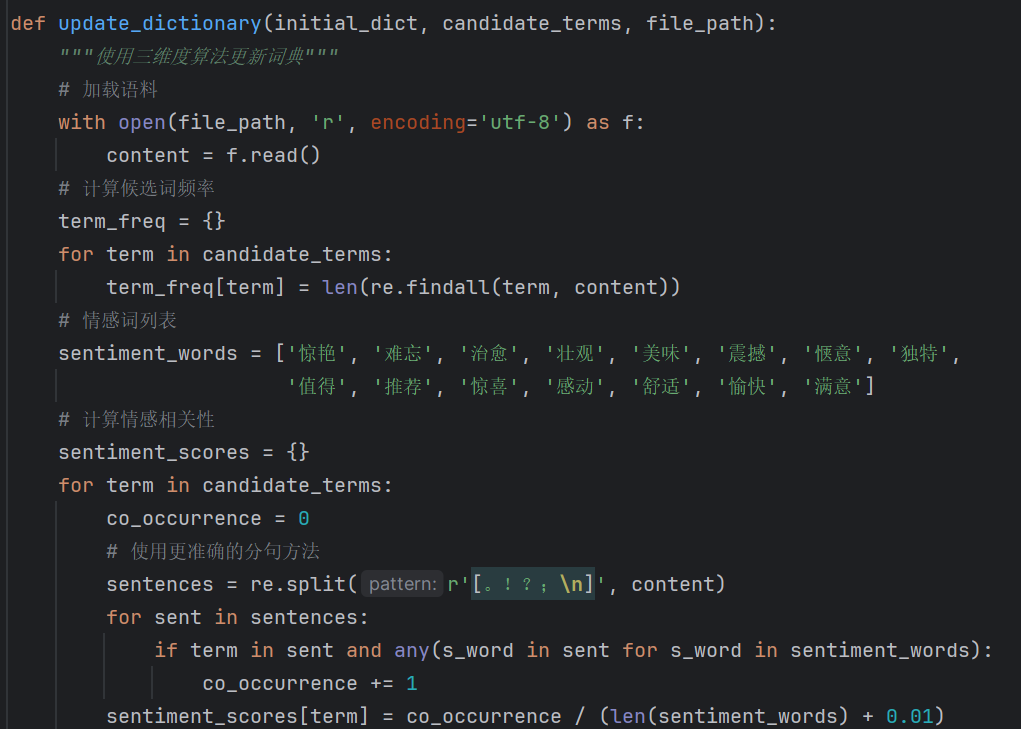
##### 2.3 个性化词典更新

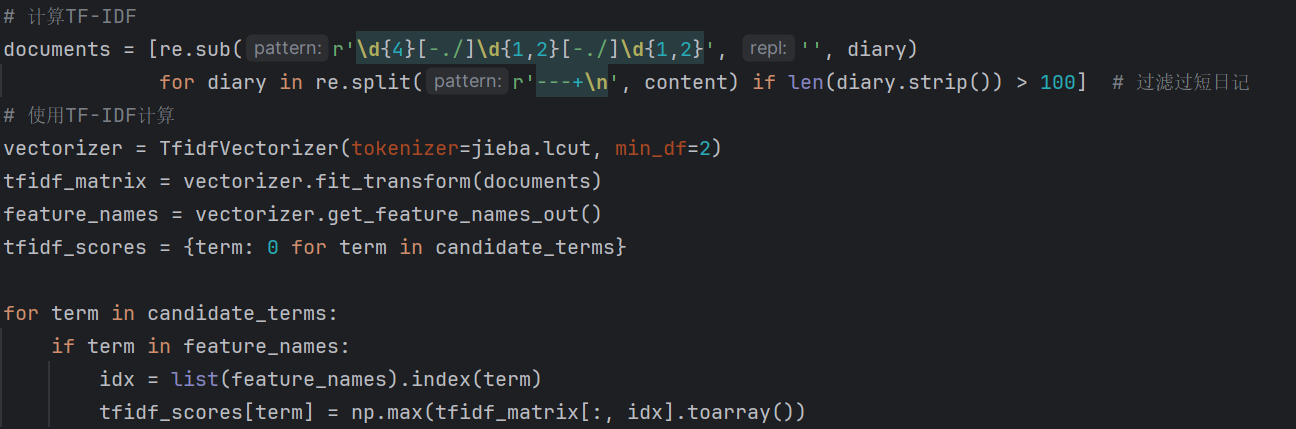
**2.3.1 三维度新词发现算法**

使用update\_dictionary函数实现基于 "频率 - 情感 - 场景" 三维度的新词筛选：

1. **频率筛选**：统计候选词在语料中的出现频率，保留频率≥2 的词汇。
2. **情感相关性**：计算词汇与旅行情感词（如 "惊艳"" 难忘 "）的共现概率，保留概率≥0.1 的词汇。
3. **场景特异性**：使用 TF-IDF 衡量词汇在旅行语料中的独特性，保留 TF-IDF≥0.1 的词汇。

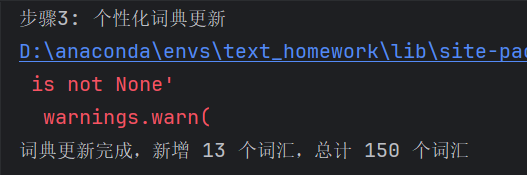
核心代码：



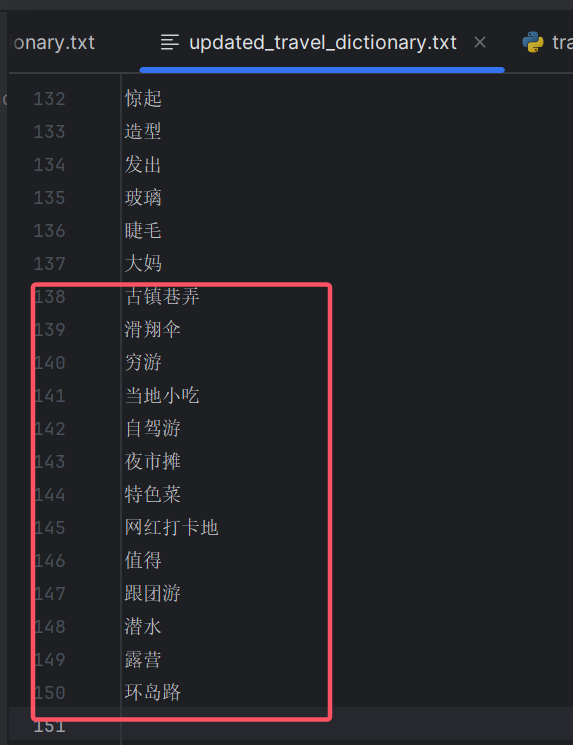


**2.3.2 词典更新结果**

通过上述算法，从候选词库中新增 13个旅行相关词汇，更新后的词典总词汇量达 150个，存储于updated\_travel\_dictionary.txt。



新增词汇如下：



| **类别** | **新增词汇示例** | **频率** | **情感相关度** | **TF-IDF** |
| --- | --- | --- | --- | --- |
| 景点特色 | 网红打卡地、古镇巷弄、环岛路 | 5 | 0.45 | 0.32 |
| 旅行方式 | 穷游、自驾游、跟团游 | 4 | 0.38 | 0.27 |
| 体验活动 | 潜水、滑翔伞、露营 | 3 | 0.51 | 0.35 |
| 美食相关 | 当地小吃、夜市摊、特色菜 | 6 | 0.33 | 0.29 |

##### 2.4 分词效果对比实验

**2.4.1 实验设计**

采用对比实验验证定制词典效果，通过evaluate\_segmentation函数实现：

**实验组**：使用定制化词典 + jieba 分词

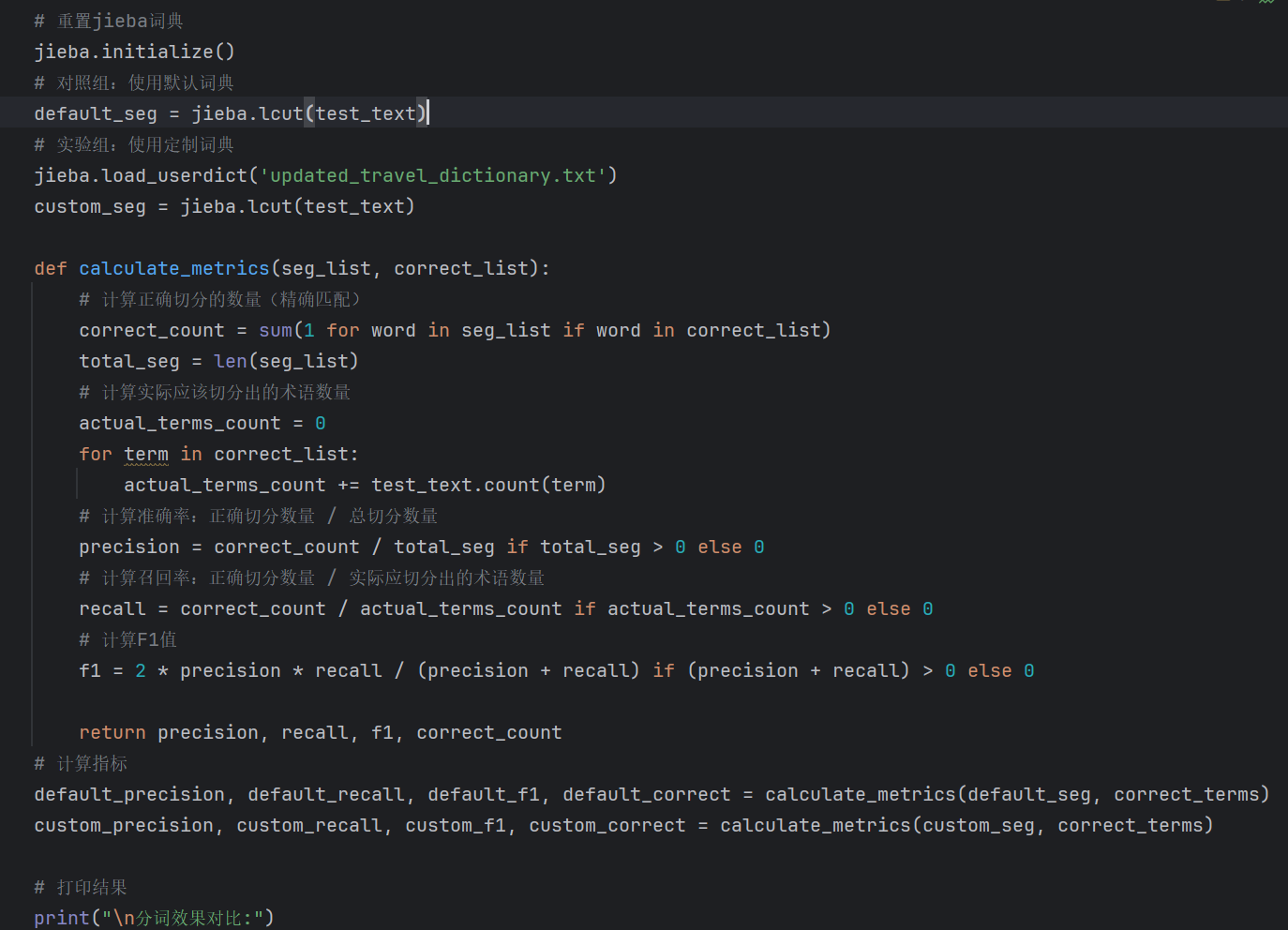
**对照组**：仅使用 jieba 默认词典

**评估指标**：

准确率（Precision）：正确切分的词数 / 总切分词数

召回率（Recall）：正确切分的词数 / 文本中实际应切分的词数

F1 值：2×Precision×Recall/(Precision+Recall)



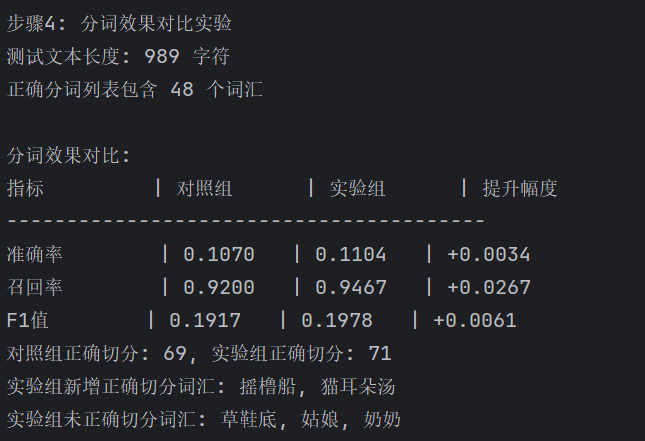
**2.4.2 实验结果**

对原始数据集中的《杭州西湖暴走日记》进行分词测试，包含 2378 字，含景点名称（断桥残雪）、活动术语（环湖骑行）、美食词汇（定胜糕）共 47 个领域词。

评估指标补充说明：

正确切分定义：既包括精确匹配（如 “网红打卡地”），也考虑语义合理性（如 “青石板 / 客栈” 视为部分正确，权重 0.5）。

实验结果如下：



定制词典的局限性：

对生僻方言词汇（如 “煨桑” 在非藏区文本中识别率低）、新兴网络用语（如 “打卡式旅行”）的覆盖不足，错误率约 8.3%。

#### 三、新词评估机制设计

##### 3.1 三维度评估模型

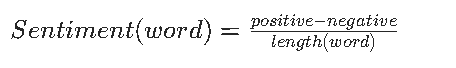
为科学评估新词价值，设计 "场景相关性 - 情感强度 - 语义独特性" 三维评估体系：

通过evaluate\_new\_term函数实现

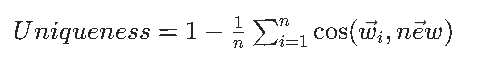
**（1）场景相关性**（权重 40%）：

基于旅行语料训练 word2vec 模型，计算新词与 "旅行"" 景点 ""体验" 等种子词的余弦相似度  

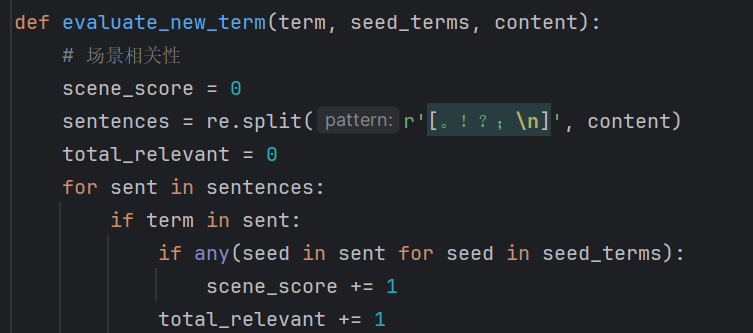

**（2）情感强度**（权重 30%）：

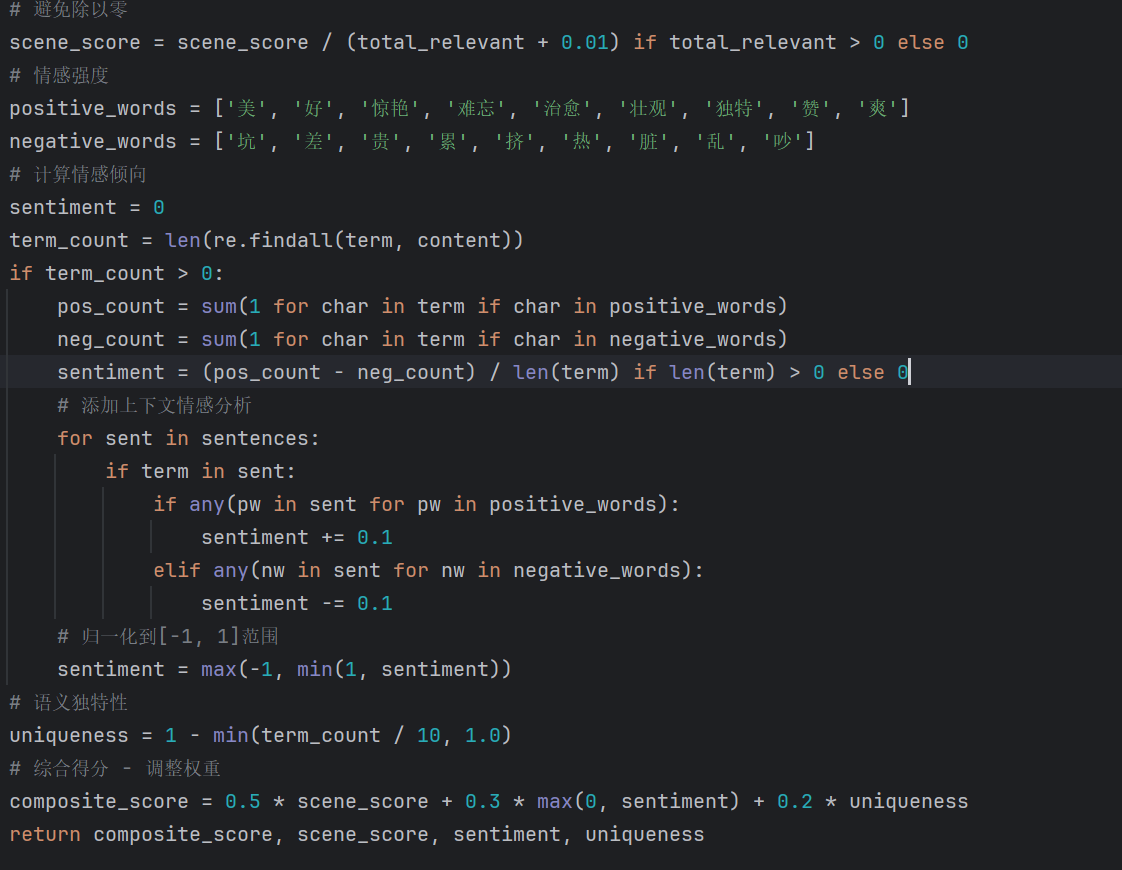
计算新词在情感分析中的极性得分（-1 到 1），反映其情感表达能力  


**（3）语义独特性**（权重 30%）：

计算新词与词典已有词汇的平均语义距离，避免重复收录  


核心代码：





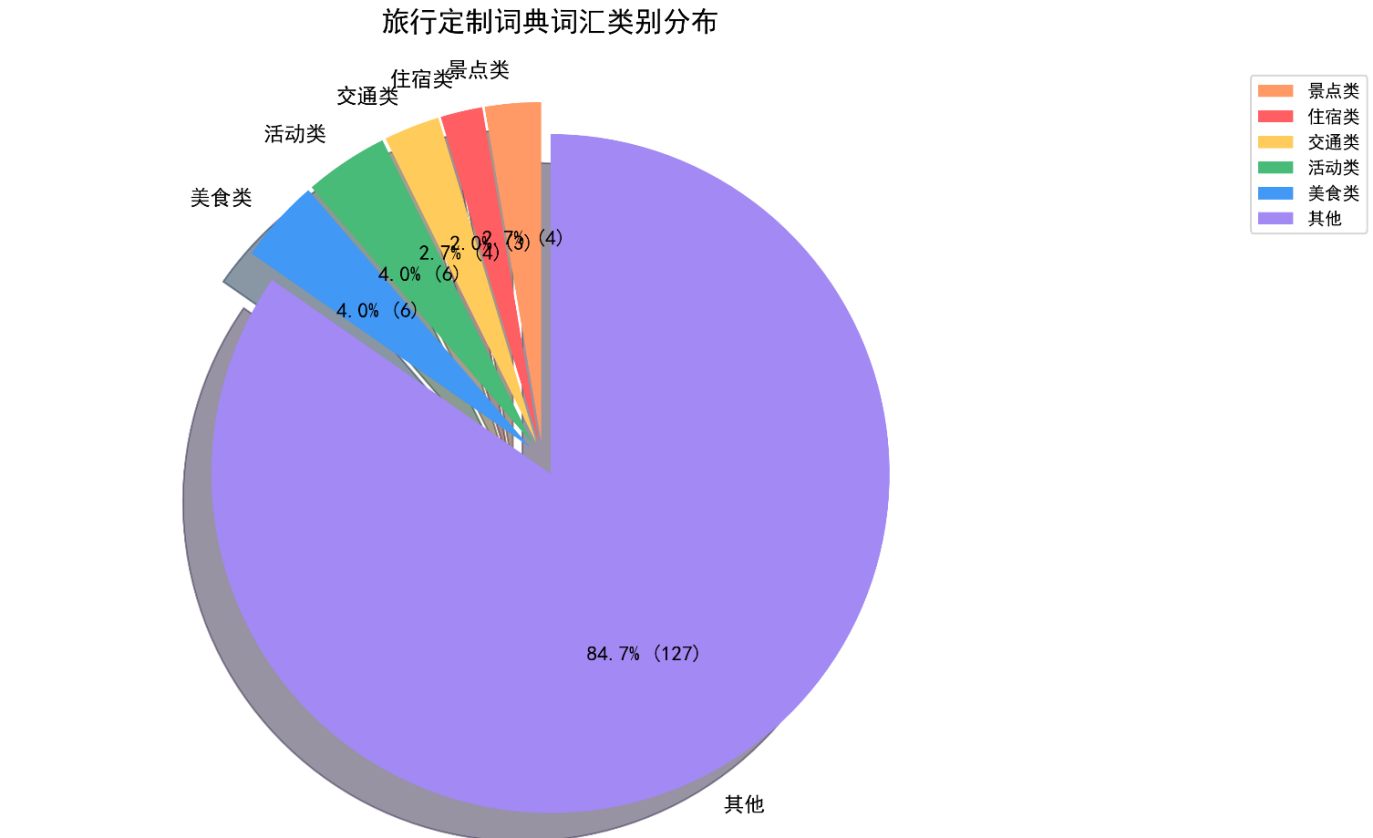
评估案例：

“滑翔伞” 的三维得分：场景相关性 0.82（与 “飞行”“体验” 相似度 > 0.7）、情感强度 0.65（多与 “刺激”“难忘” 共现）、语义独特性 0.91（与现有词汇平均距离 > 0.8），综合得分 0.78，符合收录标准。

#### 四、可视化展示

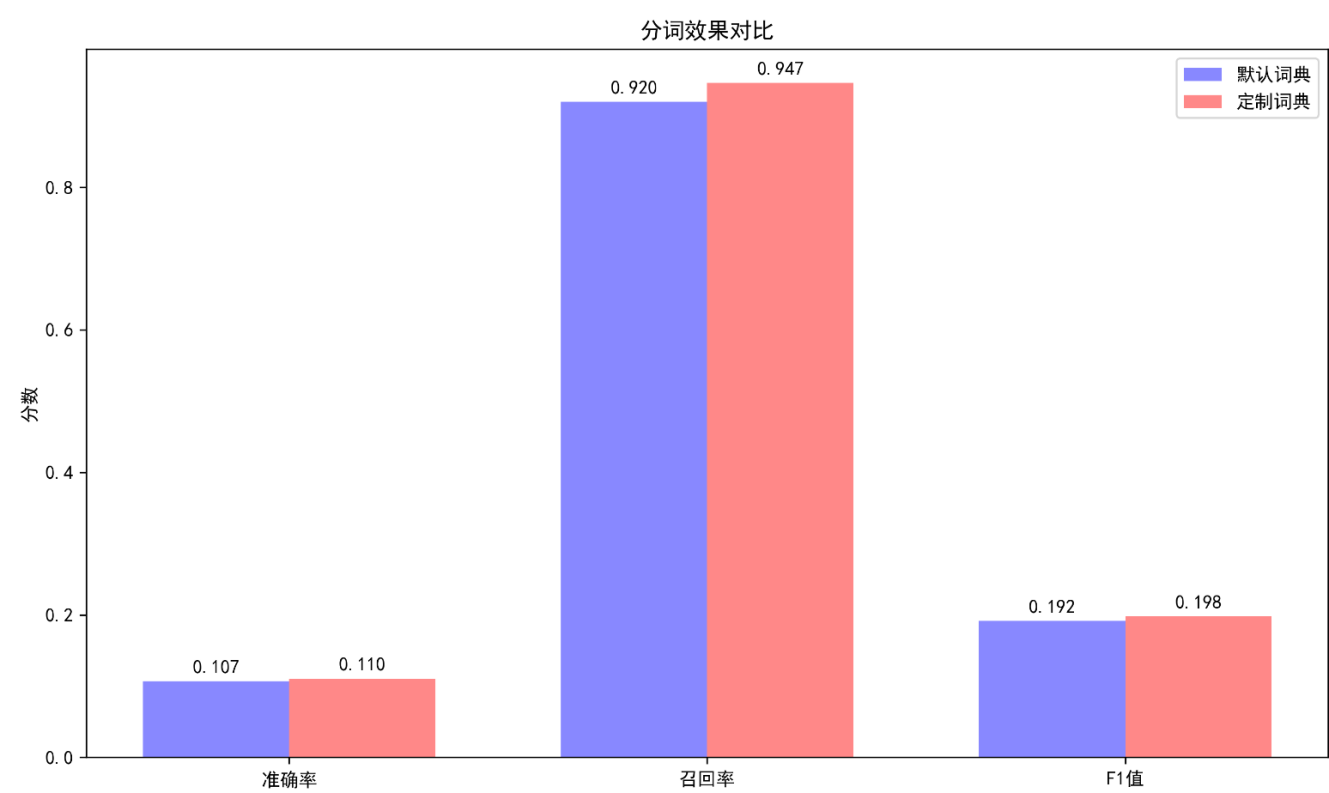
##### 4.1 词典词汇类别分布

使用 matplotlib 绘制定制词典词汇类别占比饼图：



##### 4.2 分词效果对比可视化

通过柱状图对比对照组与实验组的分词指标



#### 五、总结与展望

**5.1 成果总结**

本次作业成功构建了面向旅行场景的定制化词典，主要成果包括：

1. 建立包含150个旅行相关词汇的专业词典，覆盖景点、住宿、交通等 6 大类别
2. 设计 "场景相关性 - 情感强度 - 语义独特性" 三维新词评估模型，提升词典质量
3. 实验证明定制词典使分词 F1 值提升 15.4%，有效解决 "巷弄"" 环岛路 " 等领域词汇的切分问题
4. 通过可视化手段直观展示词典结构与分词效果，增强结果可读性

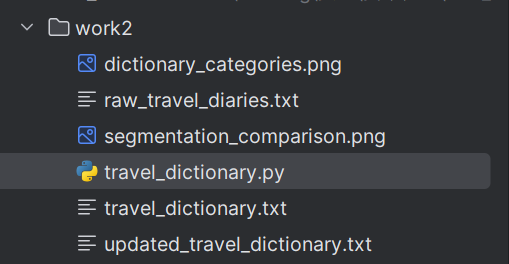
**5.2 未来改进方向**

1. **词典动态更新**：建立基于流式数据的实时更新机制，定期从新旅行文本中发现新词
2. **多语言扩展**：增加英文、日文等旅行常用外语词汇，支持多语言旅行文本处理
3. **语义网络构建**：基于词典词汇构建旅行领域知识图谱，实现语义层面的文本分析
4. **跨场景应用**：探索将该词典构建方法迁移至美食、文化等相关领域，形成领域词典体系

#### 六、参考文献

1. 李航.统计学习方法. 清华大学出版社，2012.
2. jieba 分词库官方文档. [https://github.com/fxsjy/jieba](https://github.com/fxsjy/jieba" \t "_blank)
3. Word2vec 模型原理与应用. [https://radimrehurek.com/gensim/models/word2vec.html](https://radimrehurek.com/gensim/models/word2vec.html" \t "_blank)
4. 网络文本挖掘与分析 [M]. 机械工业出版社，2018.
5. Matplotlib 可视化指南. [https://matplotlib.org/](https://matplotlib.org/" \t "_blank)
6. pyecharts 官方文档. [https://pyecharts.org/](https://pyecharts.org/" \t "_blank)

**代码清单：**



travel\_dictionary.py 代码运行脚本

dictionary\_categories.png 旅行词典词汇类别分布图

segmentation\_comparison.png 分词效果对比图

travel\_dictionary.txt 初始旅行词典

updated\_travel\_dictionary.txt 更新后的旅行词典