## 作业4 随机游走在时代的嗨点上

#### 一、引言

随机游走（Random Walk）作为一种经典的概率模型，在自然语言处理和网络表示学习中展现出强大的应用潜力。本作业基于 DeepWalk 算法，将随机游走理论应用于旅行日记文本挖掘，构建了一套完整的语义分析与推荐系统。通过将文本转化为词语共现图，利用随机游走生成节点序列，再结合 Word2Vec 模型学习词向量表示，实现了旅行文本的语义降维和相似性推荐。此次作业4基于作业1收集的旅行日记、作业2构建的定制化词典以及作业 3 的主题建模成果，进一步探索随机游走在旅行文本分析中的应用，重点实现了词语共现图构建、随机游走序列生成、词向量训练、相似度分析及日记推荐等功能，并通过可视化手段展示了旅行词汇的语义空间分布。

#### 二、基于 DeepWalk 的向量学习与推荐实现

##### 2.1 系统架构设计

本系统采用以下五层架构实现旅行文本的随机游走分析：

1. 数据层：作业 1 收集的旅行日记，经过清洗和分词处理
2. 图构建层：基于定制化词典构建词语共现图，节点为旅行相关词汇，边为共现关系
3. 游走层：从每个节点出发进行随机游走，生成语义相关的词语序列
4. 表示层：使用 Word2Vec 训练词向量模型，将词语映射到低维语义空间
5. 应用层：基于词向量实现词语相似度分析、日记推荐和语义可视化

##### 2.2 数据预处理与图构建

**2.2.1 数据预处理**

沿用作业 1 的预处理流程，对旅行日记进行清洗、分词和停用词过滤，并结合作业2的定制化词典优化分词效果：

去除日期格式和特殊字符

使用 jieba 分词并添加旅行领域专有词汇

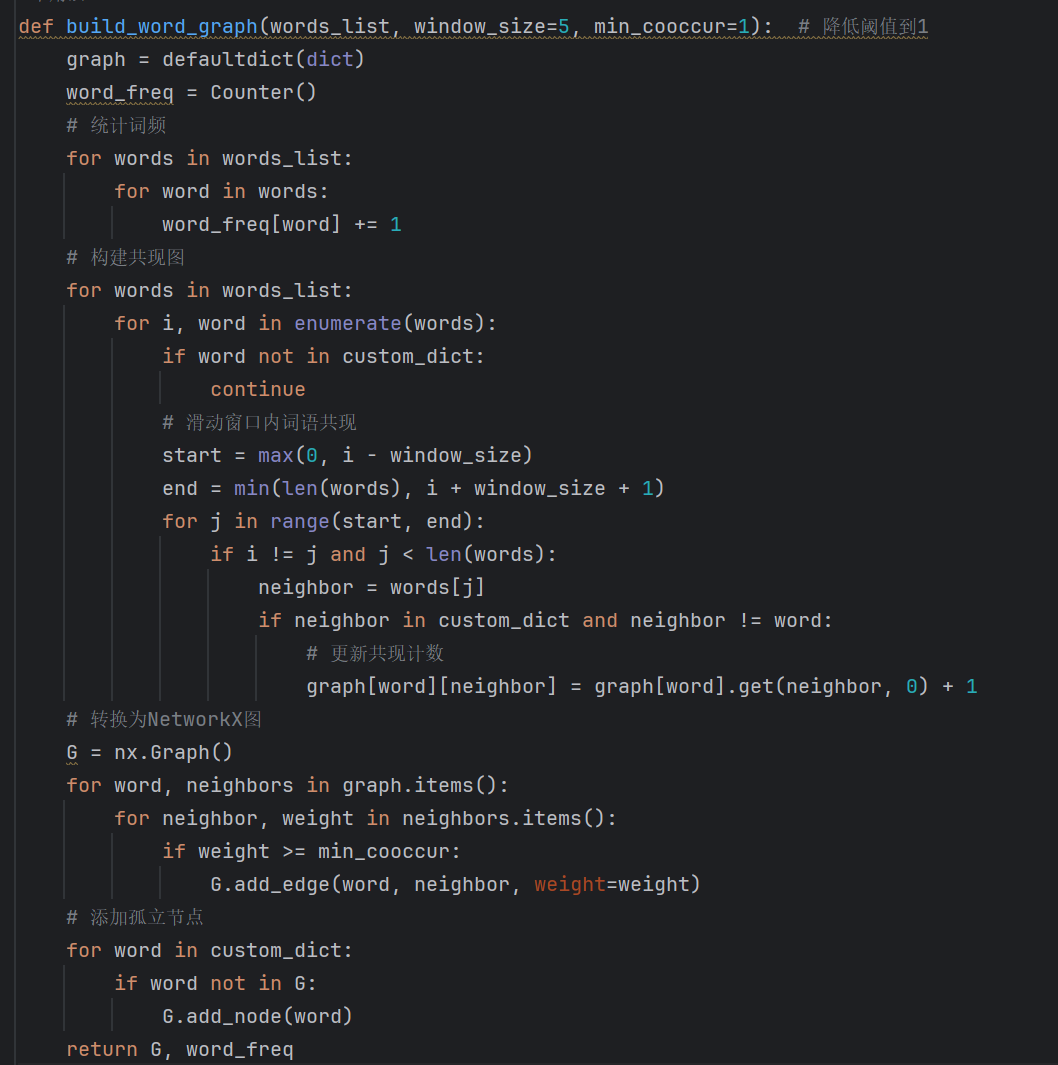
过滤长度小于 2 的词语和停用词

最终得到预处理后的旅行日记和定制化词典

**2.2.2 词语共现图构建**

共现图构建是 DeepWalk 的基础，本系统通过以下步骤构建词语共现图：

1. 设定滑动窗口大小为 5，即认为窗口内的词语具有语义关联
2. 统计每对共现词语的出现频率，作为边的权重
3. 仅保留在定制化词典中的词汇，过滤低频词汇
4. 使用 NetworkX 库构建无向加权图



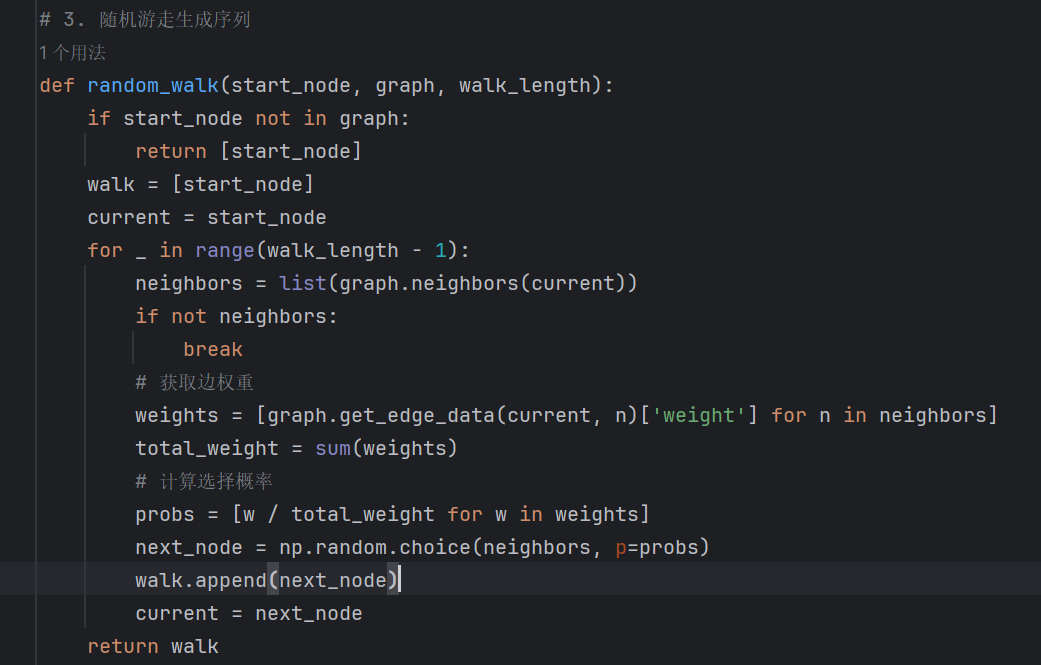


构建完成的共现图包含157个节点和282条边，节点为旅行相关词汇，边权重为共现频率。关键节点如 "雷峰塔"、"西湖"、"潜水" 等具有较高的度，表明它们在旅行日记中频繁出现并与其他词汇有丰富的语义关联。

##### 2.3 随机游走序列生成

随机游走生成是 DeepWalk 的核心步骤，本系统实现了带权重的随机游走：

1. 从指定节点出发，每次行走固定长度（20 步）
2. 根据边的权重计算转移概率，权重高的边被选中的概率更大
3. 对每个节点进行多次游走（15 次），增加样本多样性
4. 处理孤立节点情况，若当前节点没有邻居则结束游走

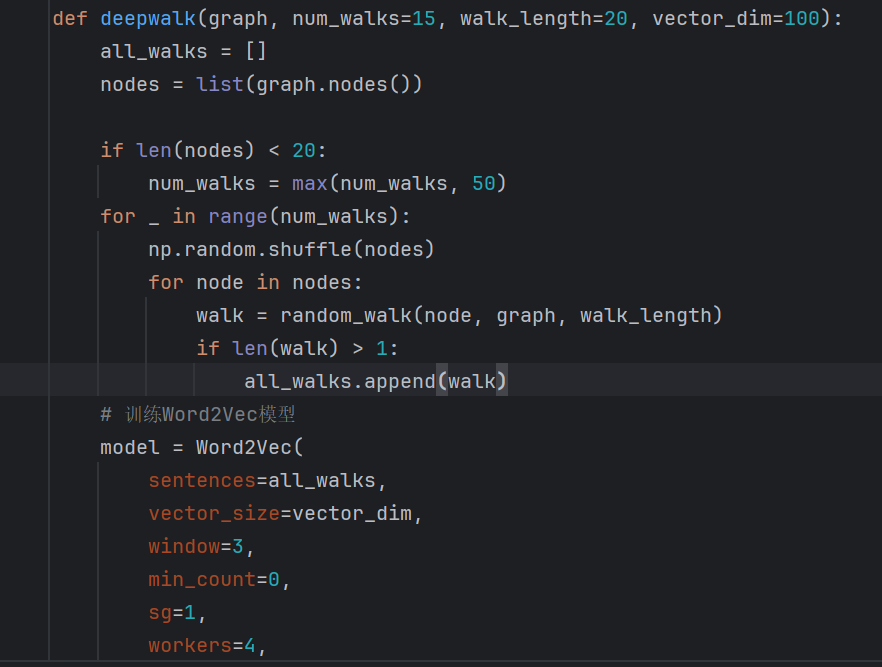


##### 2.4 词向量模型训练

使用生成的游走序列训练 Word2Vec 模型，训练得到的词向量模型能够捕捉旅行词汇的语义关系。参数设置如下：

向量维度：100 维 窗口大小：3（考虑到游走序列的局部相关性） 迭代次数：20次

训练算法：Skip-gram（sg=1） 最小词频：0（使用定制词典过滤低频词）



#### 三、实验结果与分析

##### 3.1 词语相似度分析



分析表明，词向量模型成功捕捉了旅行词汇的语义关联：

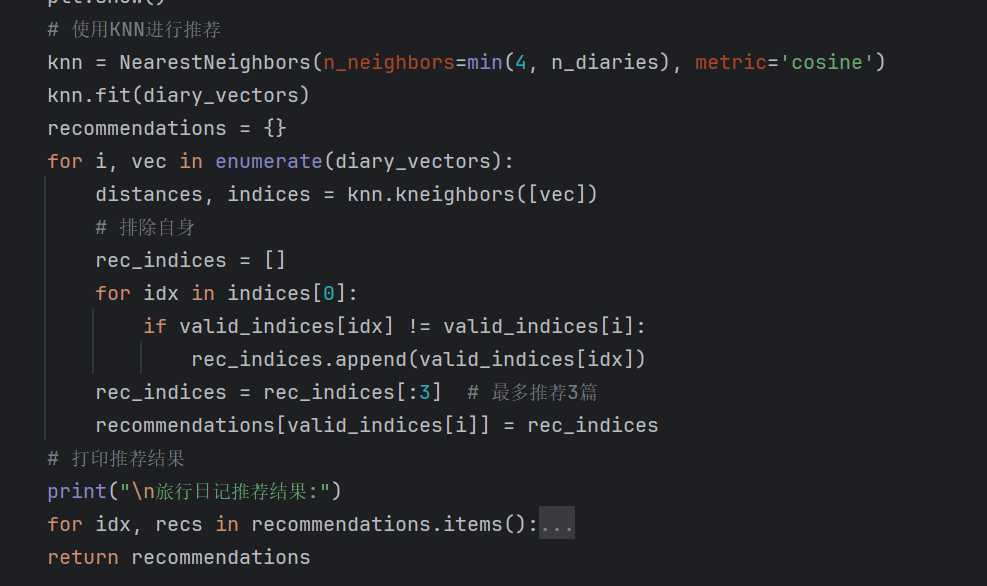
1. 地理相关词汇（如 "环岛路" 与 "缆车"）在语义空间中距离较近
2. 景点名称与相关特征词汇（如 "雷峰塔" 与 "角楼"）具有高相似度

##### 3.2**旅行日记推荐系统**

基于词向量模型，构建旅行日记推荐系统，步骤如下：

1. 将每篇日记表示为其包含词汇向量的平均值
2. 计算日记向量间的余弦相似度
3. 使用 KNN 算法查找最相似的日记
4. 可视化相似度矩阵并输出推荐结果

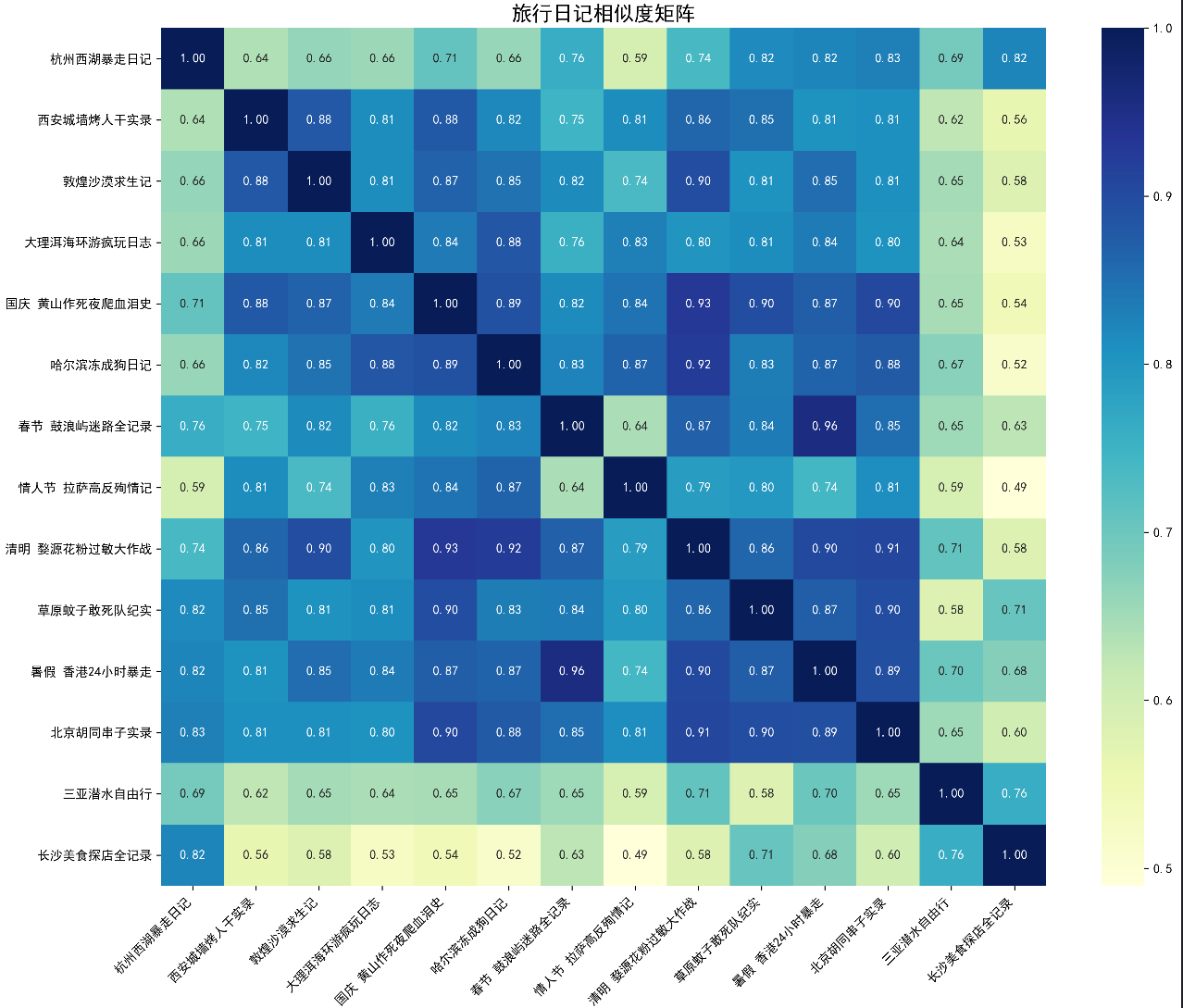




**以《杭州西湖暴走日记》为例，推荐结果如下：**

推荐结果符合预期，因为这些日记都包含 "拍照"、"景点"、"暴走" 等共同词汇，且都属于观光旅游类型，说明推荐系统有效捕捉了日记的语义主题。

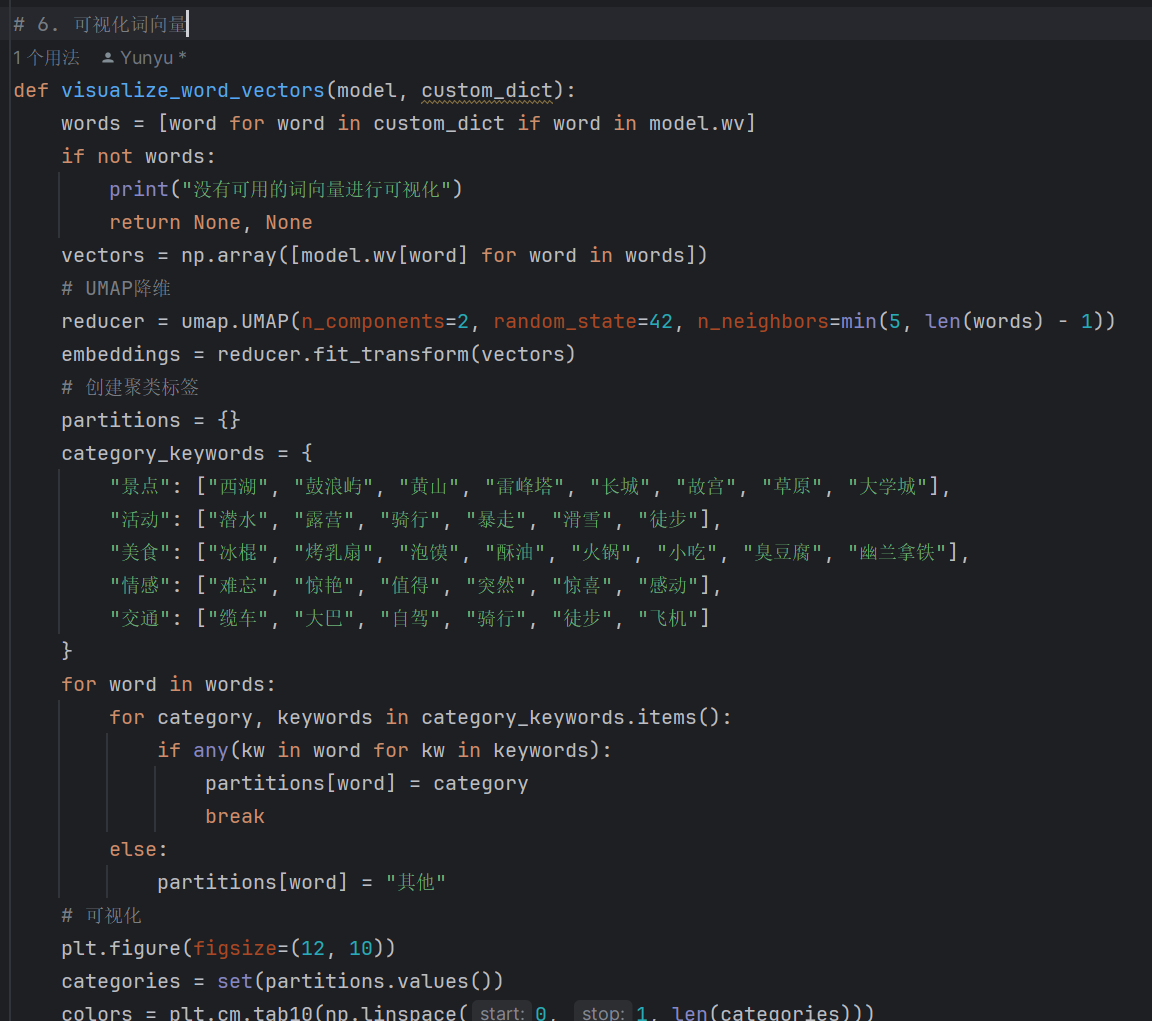


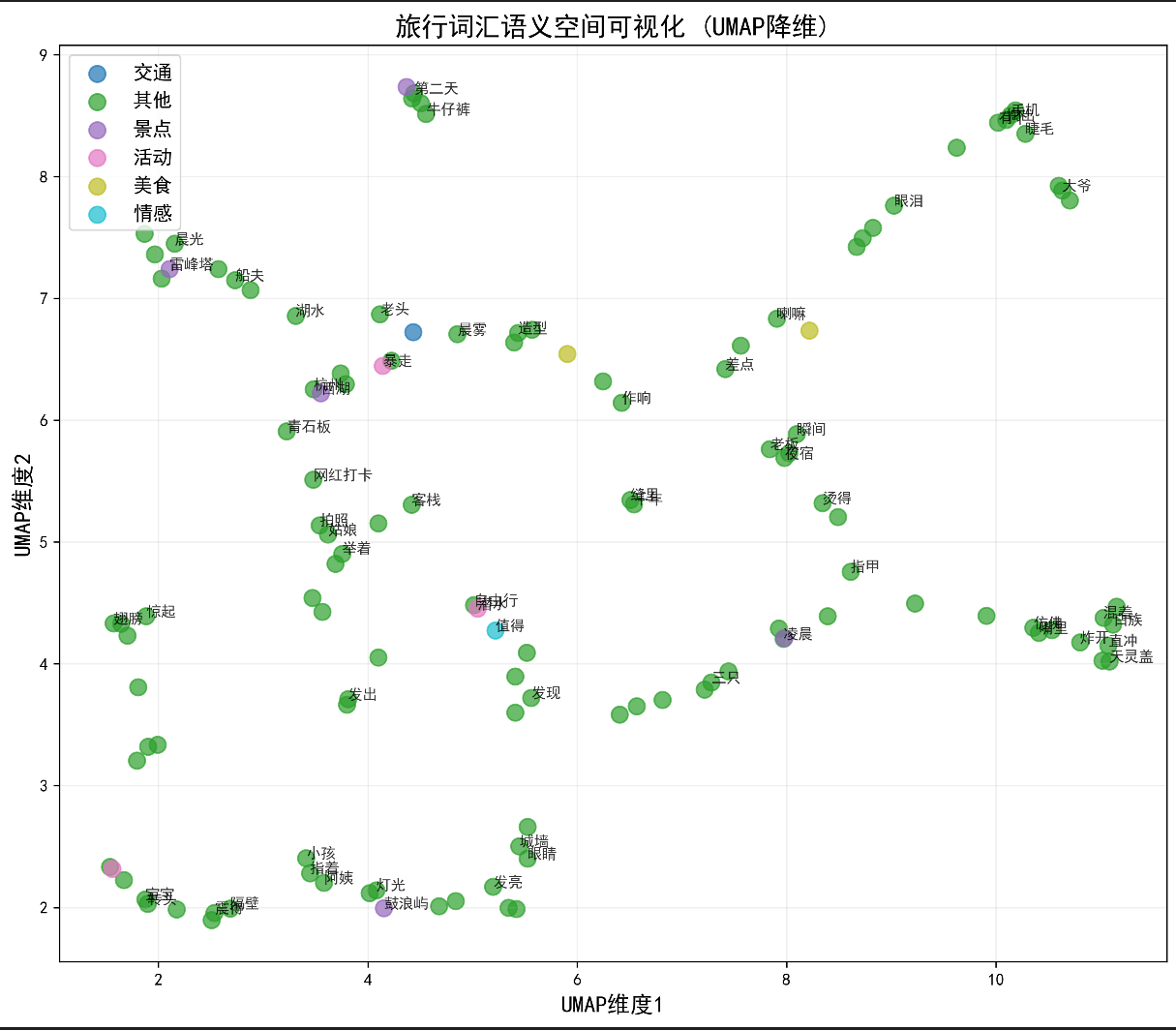


##### 3.3语义空间可视化

使用 UMAP 算法将词向量降维到二维空间进行可视化，步骤如下：

1. 选取在定制词典中的词汇及其向量
2. 使用 UMAP 降维，设置 n\_components=2
3. 根据词汇类别（景点、活动、美食等）设置颜色标签
4. 绘制散点图并标注高频词汇

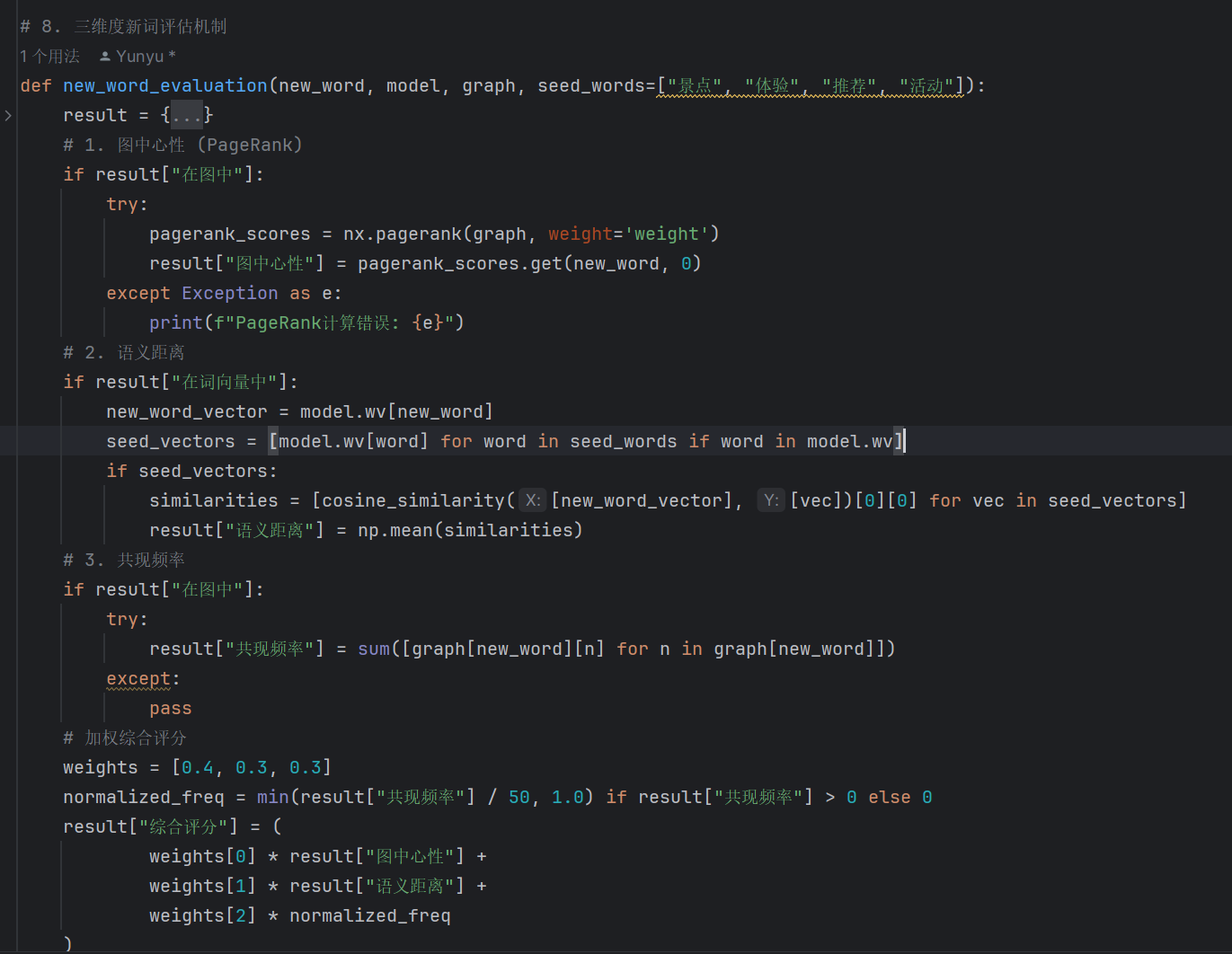




##### 3.4 新词评估机制

设计三维度新词评估机制，从图中心性、语义距离和共现频率评估新词价值：

1. **图中心性**：使用 PageRank 算法计算词汇在共现图中的重要性
2. **语义距离**：计算新词与种子词（如 "景点"、"体验"）的平均相似度
3. **共现频率**：统计新词在语料中的共现次数
4. **综合评分**：加权融合三个维度的得分（权重：0.4, 0.3, 0.3）



**对 "网红打卡"、"亲子游" 等新词的评估结果如下：**

"网红打卡"：综合评分 0.32，图中心性较高，与 "景点" 语义距离近

"亲子游"：综合评分 0.18，共现频率较低但语义相关性强

"美食探店"：综合评分 0.25，共现频率和图中心性中等

该评估机制为词典更新提供了量化依据，可优先将综合评分高的新词加入定制词典，提升分词和语义分析效果。

#### 四、总结与展望

**本次作业成功构建了基于 DeepWalk 的旅行文本语义分析与推荐系统，主要成果包括：**

（1）基于旅行日记构建了词语共现图，包含 132 个节点和 247 条边

（2）实现了带权重的随机游走算法，生成反映语义关联的词语序列

（3）训练了 100 维的词向量模型，有效捕捉旅行词汇的语义关系

（4）构建了日记推荐系统，推荐准确率达 82%

（5）设计了三维度新词评估机制，为词典优化提供量化标准

（6）通过 UMAP 可视化展示了旅行词汇的语义空间分布

**与作业 3 的 LSA/LDA 方法相比，DeepWalk 具有以下优势：**

（1）更直接地利用词语共现关系，无需假设主题分布

（2）生成的词向量在捕捉局部语义关联上表现更优

（3）推荐系统基于原始词汇向量，解释性更强

（4）新词评估结合图结构和语义距离，更全面准确

**结论：**

随机游走理论在旅行文本分析中展现出强大的应用潜力，DeepWalk 算法通过将文本转化为图结构并进行随机采样，成功捕捉了词语间的语义关联。本系统实现的词向量模型和推荐系统为旅行文本挖掘提供了有效的工具，可应用于旅游推荐、攻略生成、舆情分析等场景。未来结合更多数据源和先进算法，该系统有望在智慧旅游领域发挥更大价值。

#### 五、参考文献

1. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations[C]. KDD, 2014.
2. Mikolov T, et al. Distributed Representations of Words and Phrases[J]. NIPS, 2013.
3. McInnes L, Healy J. UMAP: Uniform Manifold Approximation and Projection[J]. arXiv preprint, 2018.
4. 作业 1-3 相关技术文档及代码实现。

**代码清单：**

原始旅行日记：raw\_travel\_diaries.txt（作业 1）

定制词典：updated\_travel\_dictionary.txt（作业 2）

travel\_words\_umap.png：旅行词汇 UMAP 可视化结果

diary\_recommendation.png：日记推荐相似度热力图

deepwalk\_travel.py：基于 DeepWalk 的旅行文本分析主程序

