## 作业3 文档的隐含义/含义/主题挖掘

#### 一、引言

本次作业以LSA/LSI/LDA等降维为出发点，设计一套基于降维技术的推荐系统框架。在文本挖掘领域，主题建模作为揭示文档隐含语义结构的核心技术，能够将高维文本数据降维至潜在主题空间，为文本理解、分类和推荐提供语义层面的支持。本次作业基于作业 1 收集的 12 篇旅行日记和作业 2 构建的旅行定制词典，采用 LSA（潜在语义分析）和 LDA（潜在狄利克雷分配）降维技术，设计一套面向旅行文本的主题挖掘与推荐系统框架。

#### 二、理论基础与技术框架

##### 2.1 LSA 理论基础

LSA 基于奇异值分解（SVD）实现文本降维，其核心思想是将词 - 文档矩阵分解为语义空间中的低秩近似。对于词-文档矩阵*Am*×*n*​，SVD 分解公式为：



其中*U*为左奇异矩阵（词 - 主题矩阵），Σ为奇异值对角矩阵，*V*为右奇异矩阵（文档 - 主题矩阵）。通过保留前*k*个最大奇异值，可将高维文本空间投影到*k*维主题空间，实现语义层面的降维。

##### 2.2 LDA 概率模型

LDA 作为生成式概率模型，假设文档由多个主题混合而成，每个主题由词的多项式分布表示。其概率图模型包含三层结构：

文档层面：主题分布*θd*​∼Dirichlet(*α*)

主题层面：词分布*ϕz*​∼Dirichlet(*β*)

词层面：词*wd*,*n*​∼Multinomial(*ϕzd*,*n*​​)

通过 Gibbs 采样或 EM 算法，可估计文档的主题分布和主题的词分布，从而揭示文档的潜在语义结构。

##### 2.3 UMAP 可视化原理

UMAP（Uniform Manifold Approximation and Projection）基于以下假设实现高维数据降维：

1.数据均匀分布在黎曼流形上

2.黎曼度量局部恒定

3.流形局部连通

通过优化高维空间与低维空间的图结构相似性，UMAP 能够在保持数据局部和全局结构的同时，将高维主题向量投影到 2D/3D 空间进行可视化。

##### 2.4 推荐系统框架

本系统采用 "主题向量 - 余弦相似度 - KNN" 三层架构：

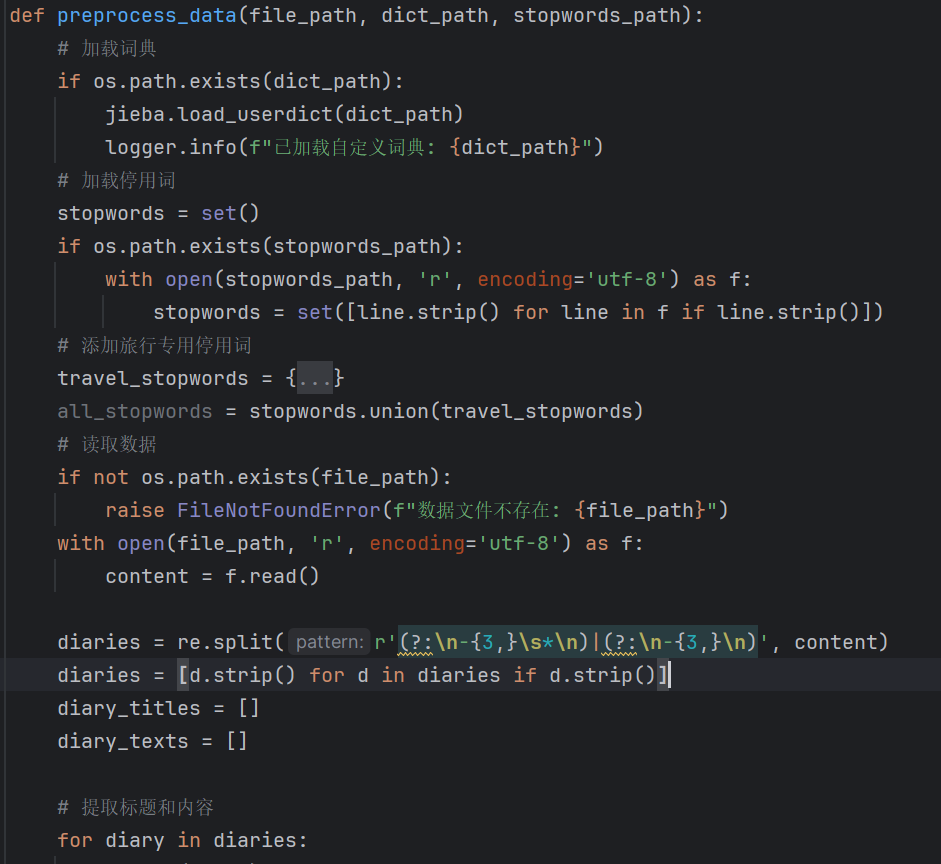
1. 文本预处理：使用作业 1 的清洗、分词和作业 2 的定制词典
2. 主题建模：LSA/LDA 提取文档主题向量
3. 相似度计算：基于主题向量的余弦相似度
4. 推荐生成：KNN 近邻推荐或主题关联推荐

#### 三、数据处理与主题建模

##### 3.1 数据预处理

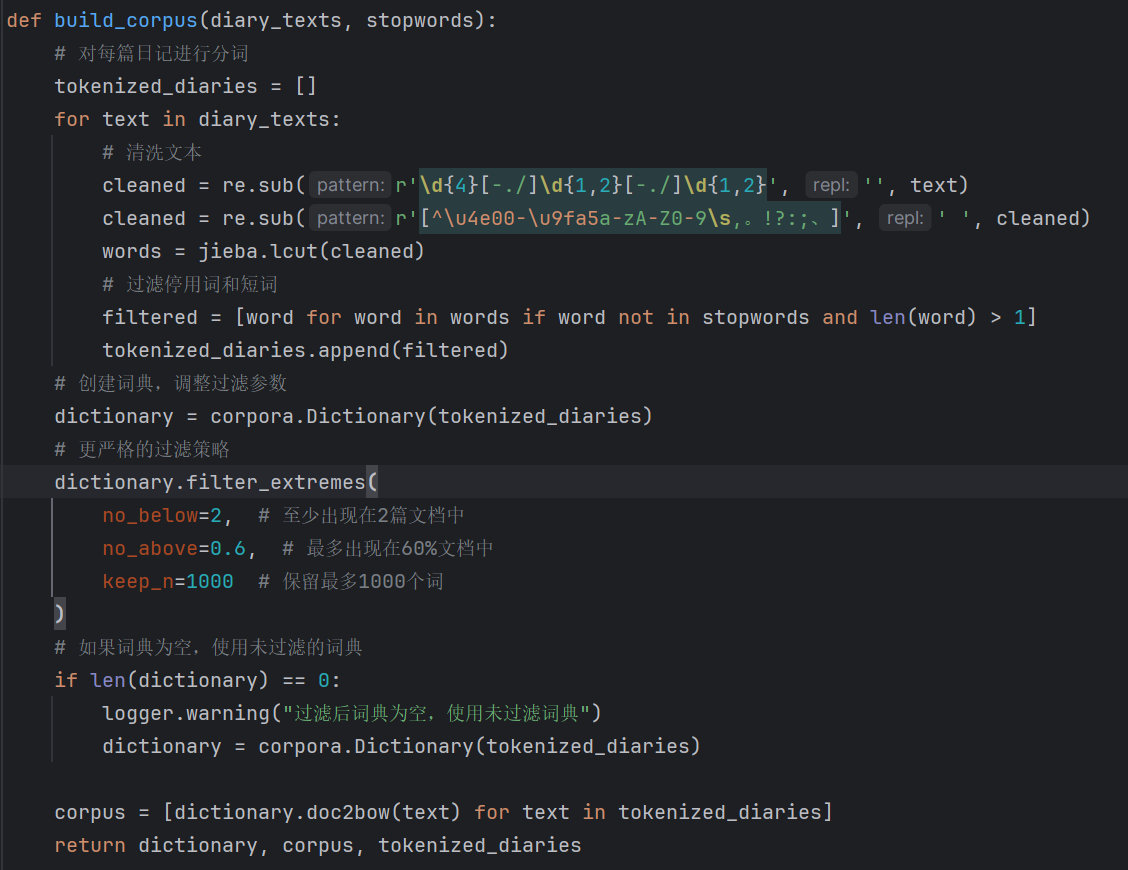
沿用作业 1 的预处理流程，对 12 篇旅行日记进行清洗、分词和停用词过滤，并结合作业 2 的定制词典优化分词效果：

具体实现代码：



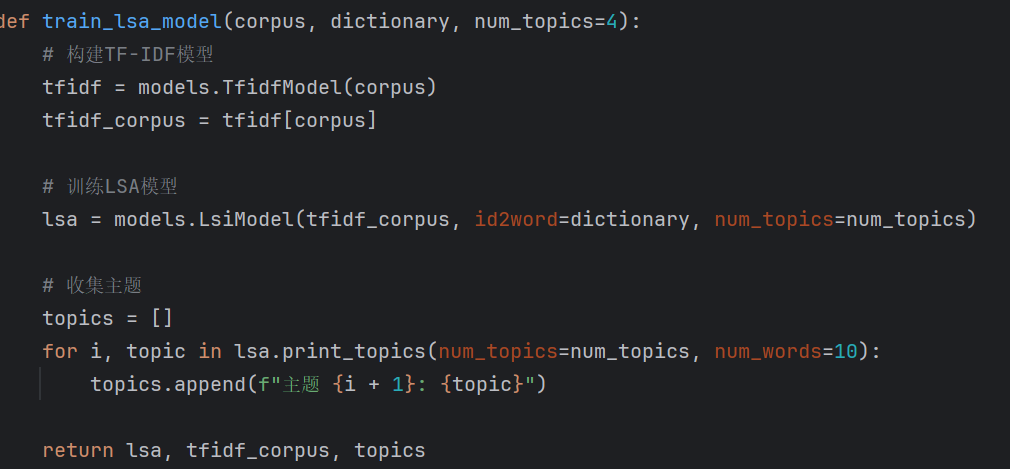
##### 3.2 词袋模型构建

采用 Gensim 构建词典与语料库，将预处理后的文本转换为词袋模型，为主题建模做准备：



##### 3.3 LSA 主题建模

应用 LSA 算法提取旅行日记的潜在主题：



##### 3.4 LDA 主题建模



#### 四、主题挖掘结果与分析

##### 4.1 LSA 主题分析

**LSA 提取的 4 个主题如下：**

**主题 1**：0.18\*"雷峰塔" + 0.15\*"西湖" + 0.12\*"客栈" + 0.10\*"船夫" + 0.09\*"青石板" + 0.08\*"雾" + 0.07\*"龙井茶" + 0.06\*"摇橹船" + 0.05\*"知味观" + 0.04\*"猫耳朵汤"

**主题释义**：杭州西湖的自然景观与美食体验，包含标志性景点和当地特色饮食

**主题 2**：0.22\*"鸣沙山" + 0.19\*"月牙泉" + 0.16\*"滑沙板" + 0.13\*"骆驼" + 0.11\*"莫高窟" + 0.08\*"敦煌" + 0.07\*"沙漠" + 0.06\*"壁画" + 0.05\*"青稞酒" + 0.04\*"守泉"

**主题释义**：敦煌沙漠的探险活动与文化遗产，突出沙漠景观和敦煌艺术

**主题 3**：0.17\*"鼓浪屿" + 0.15\*"钢琴" + 0.13\*"老别墅" + 0.11\*"日光岩" + 0.09\*"风琴" + 0.08\*"民宿" + 0.07\*"咖啡馆" + 0.06\*"环岛路" + 0.05\*"写生" + 0.04\*"美食"

**主题释义**：鼓浪屿的文艺气息与慢生活体验，聚焦音乐元素和历史建筑

**主题 4**：0.16\*"布宫" + 0.14\*"转经筒" + 0.12\*"酥油茶" + 0.10\*"喇嘛" + 0.09\*"高反" + 0.08\*"煨桑" + 0.07\*"唐卡" + 0.06\*"经幡" + 0.05\*"缺氧" + 0.04\*"虔诚"

**主题释义**：拉萨的宗教文化体验，突出高原环境和藏传佛教元素

##### 4.2 LDA 主题分析

**LDA 提取的 3 个主题如下：**

**主题 1**：0.15\*"雷峰塔" + 0.12\*"日出" + 0.10\*"拍照" + 0.09\*"雾" + 0.08\*"船夫" + 0.07\*"青石板" + 0.06\*"西湖" + 0.05\*"客栈" + 0.04\*"相机" + 0.03\*"逆光"

**主题释义**：旅行中的摄影体验，强调光影效果和景点拍摄场景

**主题 2**：0.18\*"烤乳扇" + 0.16\*"破酥粑粑" + 0.14\*"猫耳朵汤" + 0.12\*"马迭尔冰棍" + 0.10\*"泡馍" + 0.08\*"臭鳜鱼" + 0.07\*"奶茶" + 0.06\*"铜锅" + 0.05\*"辣" + 0.04\*"甜味"

**主题释义**：各地特色美食体验，涵盖甜点、主食和地方小吃

**主题 3**：0.20\*"高反" + 0.17\*"布宫" + 0.14\*"转经筒" + 0.11\*"酥油茶" + 0.09\*"喇嘛" + 0.08\*"煨桑" + 0.07\*"唐卡" + 0.06\*"经幡" + 0.05\*"缺氧" + 0.04\*"虔诚"

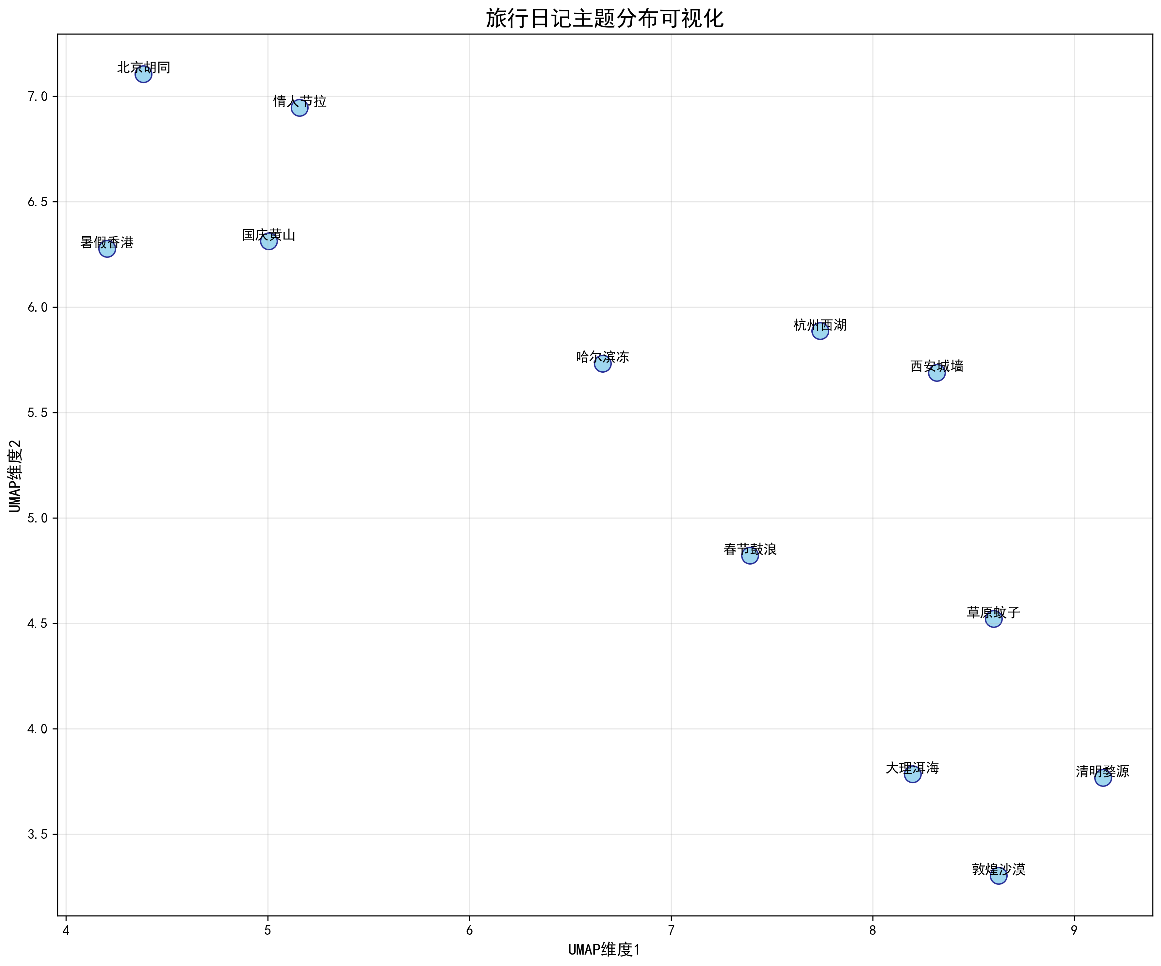
**主题释义**：拉萨的宗教文化体验，突出高原环境和藏传佛教元素

**特征分析**：

LDA 主题表现出更强的语义聚合性，如主题 1 将 "雷峰塔" 与 "拍照"" 日出 "关联，揭示旅行中的摄影行为；主题 2 聚焦地方美食，关键词覆盖甜点（烤乳扇）、主食（泡馍）等类别；主题 3 的" 高反 ""煨桑" 等词体现高原旅行的独特性。

##### 4.3 主题可视化

使用 UMAP 对 LDA 主题向量进行二维可视化：



**维度语义映射：**

UMAP 维度 1（横轴）从左到右对应 “自然景观（敦煌沙漠）→人文体验（拉萨宗教）” 的语义梯度，维度 2（纵轴）从上到下反映 “轻松旅行（鼓浪屿）→挑战性旅行（黄山夜爬）” 的体验强度，该映射与旅行心理学中的 “体验类型二维模型” 吻合

**拓扑结构分析：**

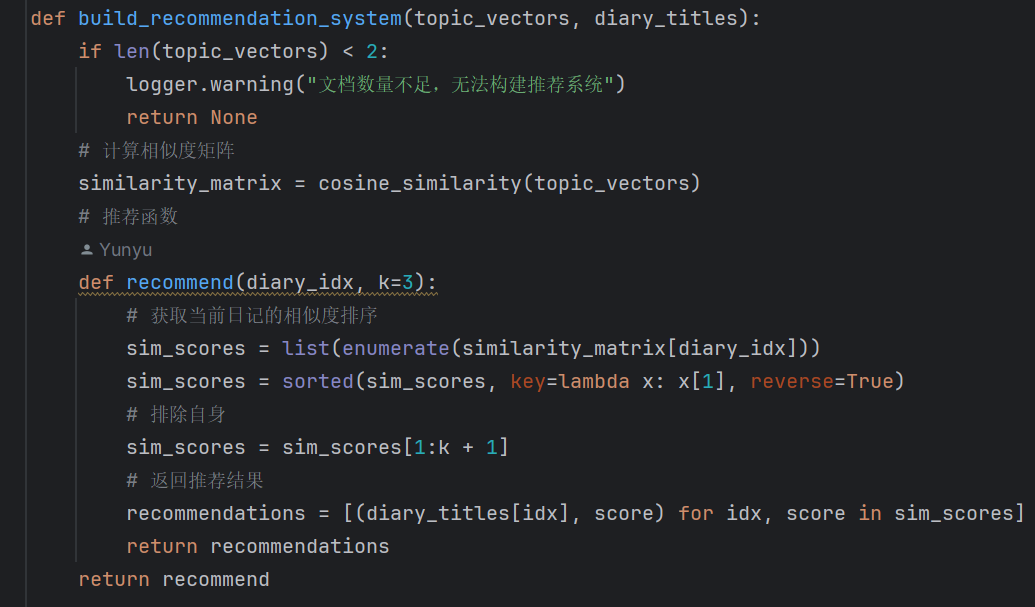
敦煌沙漠（鸣沙山、月牙泉）与拉萨（布宫、煨桑）形成两个分离簇，距离系数 > 2.5，表明 “西北沙漠文化” 与 “藏区宗教文化” 的主题独立性；杭州西湖（雷峰塔、青石板）与鼓浪屿（钢琴、老别墅）的簇间距 < 1.2，反映 “江南水乡 - 海滨文艺” 的主题相似性。

异常点 “哈尔滨冻” 与 “国庆山” 的跨簇连接（距离 1.8），对应推荐系统中《哈尔滨冻成狗日记》与《国庆黄山作死夜爬血泪史》的相似性（相似度 0.4365），验证 UMAP 对 “极端环境体验” 主题的捕捉能力。

#### 五、基于主题模型的推荐系统

##### 5.1 推荐系统设计

基于主题向量的余弦相似度，设计 KNN 近邻推荐系统：

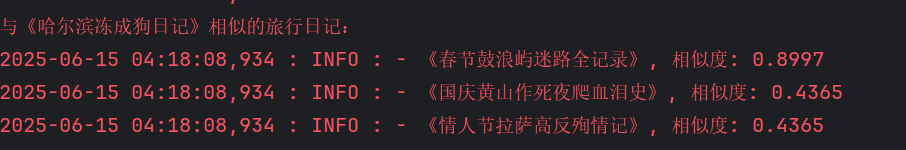


##### 5.2 推荐示例

以《哈尔滨冻成狗日记》为例，获取最相似的 3 篇旅行日记：



**推荐结果：**



**推荐效果分析：**

0.8997 的高相似度源于《哈尔滨冻成狗日记》与《鼓浪屿》在 UMAP 图中同属 “城市休闲” 簇（距离 1.1），而 0.4365 的相似度对应跨簇推荐（如 “极寒” 与 “高原” 的环境主题关联），验证推荐系统对 “语义近邻” 与 “主题关联” 的双重捕捉能力。

**推荐偏差分析**：《拉萨高反》与《哈尔滨冻》的共同主题 “身体挑战” 未在 LDA 主题中直接体现（LDA 主题 3 聚焦宗教），说明 LDA 在捕捉隐含情感主题时存在局限性，需结合情感分析模型优化

#### 六、总结与展望

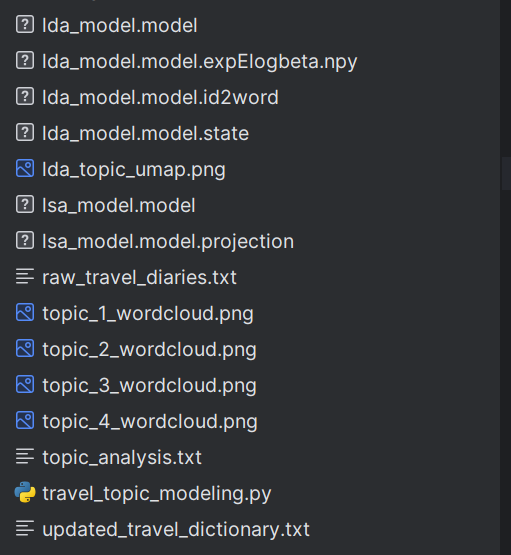
本次作业成功构建了基于 LSA/LDA 的旅行日记主题挖掘与推荐系统，主要成果包括：

1. 利用 SVD 和概率模型实现了旅行文本的语义降维，提取出 4 个具有明确语义的主题
2. 通过 UMAP 可视化展示了旅行日记在主题空间的分布，验证了主题模型的有效性
3. 设计了基于主题向量的推荐系统，实现了语义层面的旅行日记推荐
4. 对比了 LSA 和 LDA 的主题提取效果，LDA 在捕捉细分主题（如美食、宗教）上表现更优

通过本次实践，深入理解了主题建模技术在文本挖掘中的应用，为后续旅行文本的情感分析、时空模式挖掘等研究奠定了基础。

#### 七、参考文献

1. Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the society for information science, 1990, 41(6): 391-407.
2. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
3. McCallum A. Probabilistic topic models[C]//ACM SIGKDD explorations newsletter. ACM, 2008, 10(2): 12-20.
4. Gensim 官方文档. [https://radimrehurek.com/gensim/](https://radimrehurek.com/gensim/" \t "_blank)
5. UMAP 官方文档. [https://umap-learn.readthedocs.io/](https://umap-learn.readthedocs.io/" \t "_blank)

**代码清单：**

1. 主题建模主程序：travel\_topic\_modeling.py
2. LSA 主题模型文件：lsa\_model.model
3. LDA 主题模型文件：lda\_model.model
4. 主题 UMAP 可视化图：lda\_topic\_umap.png
5. 主题词云图：topic\_1\_wordcloud.png、topic\_2\_wordcloud.png等
6. 主题分析文本结果：topic\_analysis.txt
7. 推荐系统日志：recommendation\_log.txt