Acemap 搜索增强-先导项目报告

1. 数据来源

- **GAKG 子集**:
  - **格式**: Parquet 文件 (`data/gakg_subset.parquet`)
  - **结构**: 包含学术概念的三元组数据 (Subject - Relation - Object)

2. 核心方法

2.1 搜索意图理解 (intent.py)

- **方法**: 基于**规则**的自然语言处理
- **实现**: `IntentParser` 模块使用正则表达式 (Regex) 解析用户查询
  - **关键词提取**: 自动去除停用词 (如 "papers on", "research about")
  - **意图分类**: 识别排序意图 (如 "recent" -> 按时间排序) 和实体类型意图

2.2 知识图谱增强召回 (recall.py)

- **方法**: 1-Hop 邻居扩展
- **流程**:
  i. **实体链接**: 将用户关键词映射到 GAKG 图谱中的节点。
  ii. **邻居查找**: 检索该节点的一跳邻居 (Subject 或 Object)，按共现频率排序，选取 Top-K 相关概念。
  iii. **混合检索**: 并行搜索“原始关键词”与“扩展概念”，将结果融合。

2.3 客户端ui + 排序优化 (search.py & app.py)

- **UI**: 使用**streamlit**构建交互界面
- **问题**: API 不支持直接按引用量或特定日期排序，且单次返回数量受限。
- **解决方案**:
  - **批量获取**: 针对特定排序请求，后台自动通过分页获取更大规模的候选集 (200-500)
  - **内存排序**: 在 Python 客户端对候选集进行基于 `cited_by_count` 或 `publication_date` 的快速排序

3. 运行方式

- 安装依赖: `pip install -r requirements.txt`
- 启动应用: `streamlit run app.py`