**数据可视化**

**基于知识图谱的网页数据可视化**



姓 名：

学 号：

专 业： 人工智能

指 导 教 师：

2023年 6 月

目录

[一、 项目背景及意义 3](#_Toc1711830614)

[1.1 需求背景 3](#_Toc898710639)

[1.2 文本可视化 3](#_Toc1377220322)

1.3基于图的文本可视化 [4](#_Toc1363204488)

[1.4 知识图谱简介 4](#_Toc2022283620)

[1.5 知识图谱与数据可视化 5](#_Toc297120271)

[二、 核心技术简介 6](#_Toc2061208615)

[2.1 后端技术 6](#_Toc1674482548)

[2.2 前端技术 6](#_Toc254990301)

[2.3 ReAct Prompting 7](#_Toc1392113142)

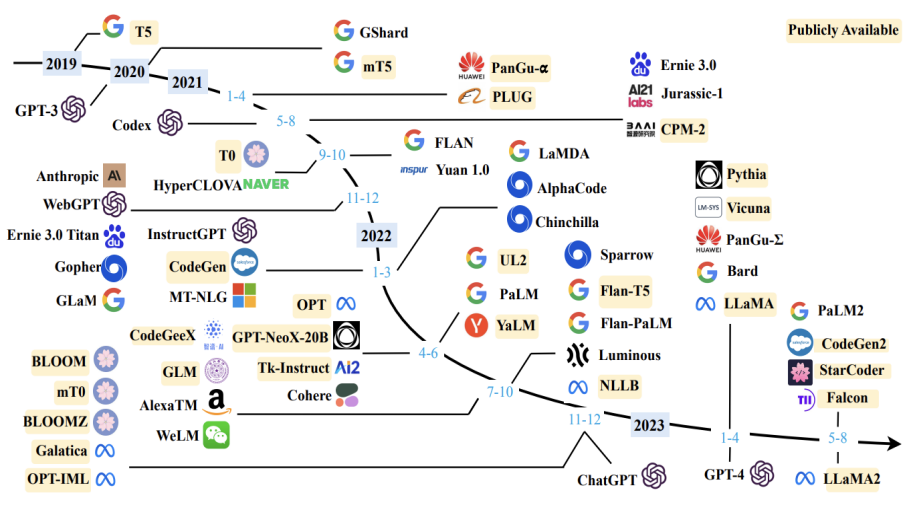
[2.4 Agent Workflow 7](#_Toc411243529)

[三、实现流程及交互设计展示 8](#_Toc1823130504)

[四、分析结论 9](#_Toc884667981)

1. 项目背景及意义

**1.1需求背景**

在信息量爆炸的时代，各类新闻和文章像潮水一样涌入我们的视线。使用GPT、Kimi、Qwen等大语言模型能够快速总结文章，让我们在短时间内了解文章的表层内容，但若想深入理解文章的细节和内部关系，则会需要其它信息数据表达方式，而知识图谱正是这种需求的最佳解决方案。

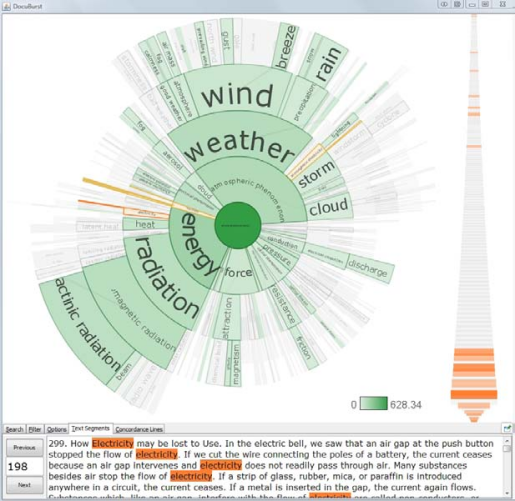
**1.2文本可视化**

文本可视化有多个角度，依据可视化表现的文本信息特征可分为文本内容可视化，文本关系可视化，文本多层面信息的可视化。这三方面并非相互独立，而是相辅相成、互相依赖。从文本内容可视化的角度进行划分，有

基于关键词的文本内容可视化方法，例如文档散、文档卡片法等

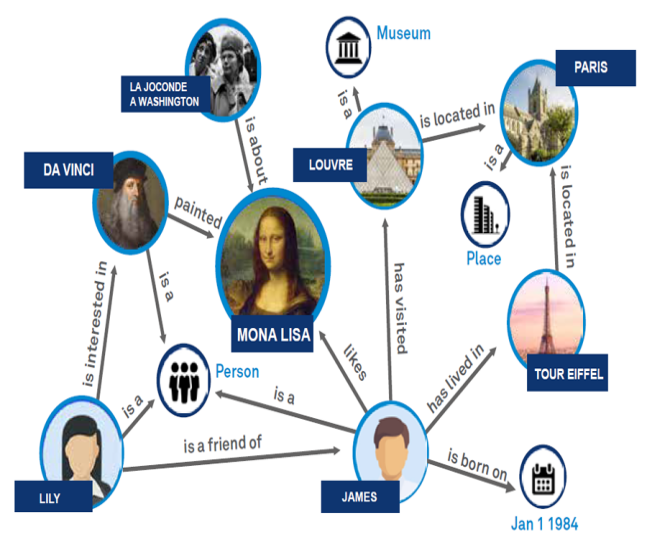
对于时序文本的内容可视化，如主题河流、TIARA、历史流等。

**1.3基于图的文本可视化**

基于文本关系的可视化旨在可视表达文章、网页内蕴含的关系信息，包括文本之间的引用、元素之间的链接关系、文本间内容的相似性和文档集合内容的层次性等。各种图的布局和投影就是常用的表达文本关系的可视化方法。

例如单词树，其利用树型结构来可视化文本中的句子。树的根节点是用户感兴趣的一个词，子节点是原文中搭配在父节点后面的词或短语。字体大小反映了词或短语在文中出现的频率。

**1.4 知识图谱简介**

知识图谱是一种表示实体（如人、地点、事件等）及其关系的图形结构。可以采用节点-链接图展示无结构文本中语义单元彼此间的关系，如“X is Y”。节点代表语义单元，如词或短语，边代表用户指定的关系，箭头代表关系的有向性，边的宽度指示这对短语关系在文中出现的频率。通过这种方法可以直观的展示文本中各个实体的关联关系。

知识图谱构建的一般流程为：

首先确定知识表示模型（类似于确定房子的图纸）；

然后根据数据来源选择不同的知识获取手段把知识导入模型中（类似于建房子），这一步就是构建知识图谱；

接着综合利用知识推理 、知识融合 、知识挖掘等技术对构建的知识图谱进行质量提升 （类似于装修）

最后根据场景需求设计不同的知识访问方法 ，如语义搜索、问答交互等（类似于入住）。

使用知识图谱完成对数据的可视化有以下优势：

结构化数据：通过将数据中的关键信息节点化，并展示它们之间的关系，帮助用户更直观地理解数据内容。

揭示隐藏关系：不仅仅是表现数据内容，还能展示不同概念和实体之间的关系，揭示出隐藏的关联和逻辑。

提高信息获取效率：用户可以通过知识图谱快速抓取重点信息，而无需逐字阅读整篇文章，极大提高了信息获取的效率。

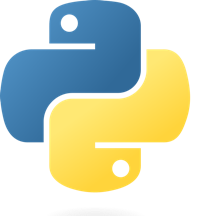
**1.5 知识图谱与数据可视化**

知识图谱与数据可视化的结合可以带来更深刻的洞察和更高效的分析。通过将知识图谱中的复杂关系和结构化数据可视化，用户可以更直观地理解数据之间的关联和模式。例如：

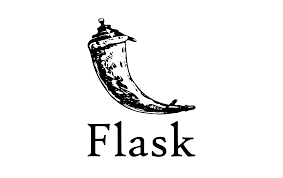
网络图：用于展示知识图谱中的实体和关系。

时间轴可视化：展示事件之间的时间关系。

层次结构图：展示知识图谱中的层次结构和分类。

1. 核心技术简介

**2.1 后端技术**

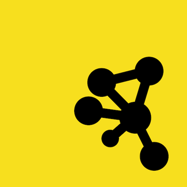
· 开发语言与服务框架：使用Python作为开发语言，便于快速实现数据处理功能。Flask用于构建轻量级Web服务器，处理前后端通信和数据请求，支持数据可视化和网络爬虫的后端服务。

· 数据分析、处理：世界最先进的多模态大模型GPT-4o，用于对数据的分析和信息抽取，最后进行格式化输出JSON数据，用于可视化部分。

**2.2 前端技术**

· 开发语言：使用TypeScript相对于JavaScript更好维护，提供类型检查和更好的代码质量，适用于构建稳定的前端应用。

· 前端框架：基于React的Next.js框架，支持服务器端渲染和静态站点生成，提升前端性能和SEO效果。

·数据可视化库：Cytoscape.js是一个用于图论和网络可视化的JavaScript库，帮助展示复杂的数据关系和结构，增强数据可视化效果。

构建与打包工具： Bun是用Zig语言开发的高性能JavaScript运行时，提升开发和运行速度，适用于现代Web开发中的构建和打包工具。

**2.3 ReAct prompting**

在生成知识图谱的过程中，ReAct Prompting 可以提供实时的指导和提示，实时检测并纠正错误，帮助用户更高效地完成任务。

数据输入：当用户输入实体或关系数据时，系统可以实时验证输入的正确性，并提供自动补全或错误提示。例如，提示用户输入的实体名称是否已存在，或关系类型是否有效。

实体识别：在用户标注文本中的实体时，系统可以实时提供推荐的实体类别和相关信息，帮助用户快速完成标注。

关系验证：当用户建立实体之间的关系时，系统可以根据已有的知识图谱数据验证关系的合理性，并提示可能的错误或冲突。

智能补全：系统可以根据已有的知识图谱数据，智能补全用户输入的实体和关系，提升生成效率。

**2.4 agent workflow**

定义核心函数：

· `deepseek\_react(messages)`：根据输入消息选择合适的工具（如`WebCrawler`或`Condenser`），并产生思考&行动。

· `generate\_text\_completion(model, messages)`：调用指定模型生成文本补全。

· `webScrawler(doc\_url)`：使用`WebBaseLoader`工具从指定URL加载网页内容。

· `generate\_graph(text)`：根据输入文本生成知识图谱。

处理POST请求的逻辑：

· 从请求中获取`text`。

· 调用`deepseek\_react`确定动作。

· 如果动作是`WebCrawler`，则调用`webScrawler`获取网页内容。

· 根据`text`的长度决定是否进行内容压缩（condense）。

· 调用`generate\_graph`生成知识图谱。

· 返回生成的图谱数据或错误信息。

三、实现流程展示

1. 用户在前端页面输入文本数据；

2. 前端通过fetch API将用户输入发送到后端的`/update\_graph`端点；

3. 后端使用GPT-4o处理输入数据，生成知识图谱格式的JSON数据；

4. 后端将生成的知识图谱数据返回给前端；

5. 使用Cytoscape.js将知识图谱数据可视化，前端网页上展示给用户。



四、分模块介绍

4.1 视口设计

视口一：文本输入&高亮 (TextInputHighlightScreen)

负责任务：允许用户输入文本，并在输入过程中高亮显示特定的关键词或结构，以帮助用户更好地理解和组织输入的内容。

展现数据类型：用户输入的文本，关键词或结构的实时高亮显示。

可视化方法：

* 文本输入框：提供一个清晰可见的区域供用户输入文本。
* 关键词高亮：使用不同颜色突出显示输入文本中的关键词或特定结构，增加文本的可读性和互动性。

视口二：知识图谱渲染&交互

(KnowledgeGraphRenderInteractionScreen)

负责任务：展示与输入文本相关的知识图谱，并允许用户通过图谱交互探索相关概念和连接。

展现数据类型：知识点之间的关系和属性，如概念、类别和联系。

可视化方法：

· 图谱展示：以节点和边的形式展现不同的知识点和它们之间的联系。

* 交互功能：用户可以点击或悬停在任何节点或连接上，查看详细信息或进行进一步的操作。

视口三：主题色切换 (ThemeColorSwitchScreen)

负责任务：允许用户根据个人喜好或特定需求更改应用的主题色彩明暗，改善视觉体验和应用的可用性。

展现数据类型：不同的颜色选项和预览。

可视化方法：

· 颜色选择器：提供一个颜色选择器，用户可以从中选择喜欢的颜色。

* 即时预览：更改颜色后，应用的界面能够即时显示新的主题色彩，让用户直观地看到更改效果。

这三个视口的设计注重了用户交互和数据可视化的结合，旨在提升用户体验并提供直观、有效的信息展示。



五、可视化展示

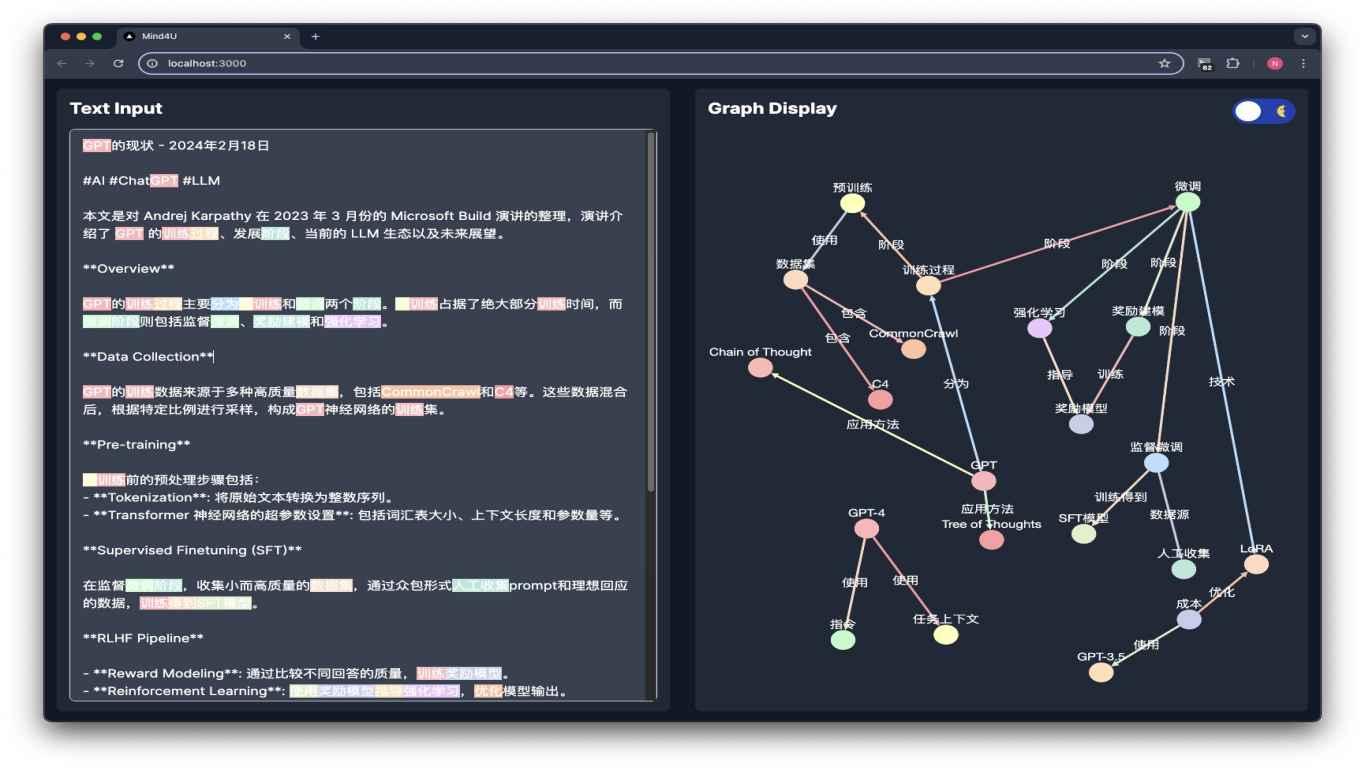
·封装成APP供所有人使用（全平台）

·项目部署Demo链接 <https://kg.nagi.fun/>

·用户前端交互展示：

在原始文段中高亮显示各节点所对应的文本与分词

以知识图谱的形式展示各分词节点之间的关联关系



六、分析结论

该项目通过结合最先进的多模态大模型GPT-4o与知识图谱技术，将复杂的信息结构化并进行可视化，显著提升了信息处理的效率和质量。

知识图谱不仅帮助用户快速理解大量数据的表层内容，还能揭示深层次的关联和逻辑，使得信息的获取和理解更加深入和全面。通过结构化的数据展示和直观的图形化界面，为用户提供了一个全新的视角去分析和解读数据，用户可以轻松抓住关键信息，发现隐藏的关系，从而做出更加明智的决策。

七、未来计划

实现对用户输入的网站内容进行缓存，允许内容的共享，并支持其他用户对已缓存内容的查看和修改，从而促进信息的互动和协作。 展现数据类型：用户输入的网站数据，包括文本、图像、链接等，以及由这些数据生成的知识图谱。

互动式知识图谱：生成的知识图谱不仅呈现信息的结构和关系，还可以由用户进行实时编辑和更新，如添加新的节点或修改现有的连接。

历史版本跟踪：为每个用户操作提供历史记录，允许用户查看和回滚到先前的状态，增强协作的灵活性和数据的完整性。

通过这样的扩展，项目不仅能够提供单向的信息展示，而且通过缓存机制和用户互动，使得每个人的贡献和修改都能被保存和共享。这不仅加深了用户对数据的理解，也促进了集体智慧的形成和信息的全面发展，为用户提供了一个共享和协作的平台，使他们能够在全新的维度上探索和利用数据，进一步发掘信息的价值。