基于HanLP的新闻文本自动化知识图谱构建说明文档

1. 项目背景与意义

背景

- **数据量激增**：随着数据量的指数级增长，特别是在新闻行业，对有效管理和利用这些信息的需求日益增加。知识图谱作为一种组织复杂数据关系的方法，可以帮助更好地理解和利用这些数据。
- **信息提取与组织需求**：新闻文本包含丰富的信息，如事件、人物、地点和时间等。通过自动化技术提取这些信息并有效组织成图谱，可以加速信息检索和分析过程，支持更深入的数据洞察。
- **HanLP的应用**：HanLP是一个高效的中文自然语言处理库，提供了包括词法分析、句法分析、命名实体识别等多种功能。利用HanLP处理新闻文本可以高效准确地抽取关键信息，为知识图谱的构建提供基础。

意义

- **提高信息处理效率**：构建知识图谱，可以快速从大量新闻文本中提取关键信息，提高了数据处理的速度和效率，用户能够更快获得和利用信息
- **增强数据可查询性**：知识图谱通过将信息以结构化的形式呈现，便于机器和人类理解，且易于进行复杂查询和数据分析
- **支持智能化决策应用**：构建的知识图谱可以作为智能搜索引擎、推荐系统等系统的基础。通过理解新闻事件的背景和联系，这些应用可以提供更精准的服务

2. 技术实现

一般来说，知识图谱的构建过程主要包括以下几个步骤：

1. 数据收集：从各种来源获取结构化和非结构化的数据。
2. 数据预处理：包括数据清洗、实体识别等步骤，目的是将原始数据转换为适合构建知识图谱的格式。
3. 知识抽取：使用知识抽取技术，从数据中提取出实体、关系和属性等知识。关键技术包括实体抽取、关系抽取和属性抽取。
4. 知识融合：通过知识抽取和知识融合，已经实现从原始杂乱数据中获得到一系列基本的事实表达。
5. 知识表示：通过知识表示学习技术，将这些知识转化为计算机可以理解和使用的向量表示。

6. 知识存储：针对构建知识图谱设计底层的存储方式，完成各类知识的存储，包括基本属性知识、关联知识、事件知识、时序知识、资源类知识等。
7. 知识计算：对于经过融合的新知识，需要经过质量评估之后（部分需要人工参与甄别），才能将合格的部分加入到知识库中，以确保知识库的质量。
8. 知识应用：知识图谱的应用主要分为用于构建结构化的百科知识的“通用知识图谱”和基于行业数据构建和应用的“领域知识图谱”。

在本项目中，主要涉及数据收集、数据预处理、知识抽取、知识融合、知识表示、知识存储和知识应用步骤。

2.1 中文文本数据处理

在自然语言处理中，中文文本往往更难以进行处理，主要体现在以下两点：

1. 分词问题：中文在每句话中间是不存在分隔符的，而是由一串连续的汉字顺序连接构成了句子。同时，中文的语法结构相对较为灵活，词序和句法结构可以有较大的变化，相比之下，英文的语法结构相对规范和固定。所以，像英文分词那样，依靠一些规则去进行处理在中文语境下有很大的局限性。
2. 语言歧义性：中文中存在大量的歧义性，即一个词或句子可能有多种不同的解释。这增加了中文文本理解的难度。例如，中文中的一词多义现象比较普遍，一个词在不同的上下文中可能有不同的含义。解决这种歧义性需要结合上下文信息和语境进行推理和判断。

团队对比了多种自然语言处理工具（如HanLP、spaCy、Stanford NLP等），考虑到HanLP在中文处理上的优势明显，并且HanLP支持的语义角色标注功能可以准确识别文本中的动作及其执行者和对象，故选择了它进行文本分析。

text = "中共中央组织部日前下发通知，任命辛锋同志为中国核工业集团有限公司党组成员、副总经理。"

HanLP模型

Word: ['中共', '中央', '组织部', '日前', '下发', '通知', '，', '任命', '辛锋', '同志', '为', '中国', '核工业', '集团', '有限', '公司', '党组', '成员', '、', '副总经理', '。'], POS: ['NR', 'NN', 'NN', 'NT', 'VV', 'NN', 'PU', 'VV', 'NR', 'NN', 'VV', 'NR', 'NN', 'NN', 'JJ', 'NN', 'NN', 'NN', 'PU', 'NN', 'PU'], NER: [['中共中央组织部', 'ORGANIZATION', 0, 3], ['辛锋', 'PERSON', 8, 9], ['中国核工业集团有限公司', 'ORGANIZATION', 11, 16], ['党组', 'ORGANIZATION', 16, 17]]

很明显HanLP模型对中文识别效果更好

Stanford NLP模型

Word: 中共, POS: PROPN
Word: 中央, POS: NOUN
Word: 组织, POS: VERB
Word: 部, POS: PART
.....
Word: 为, POS: AUX
Word: 中国, POS: PROPN
Word: 核工, POS: NOUN
Word: 业, POS: PART
Word: 集团, POS: NOUN
Word: 有限, POS: ADJ
Word: 公司, POS: NOUN
Word: 党, POS: NOUN
Word: 组成, POS: VERB
Word: 员, POS: PART
.....

同时引入openai大语言模型，通过自定义提示和对话系统来实现更精确的信息校验和反馈，提高知识图谱的质量和准确性。

```python
client = OpenAI(base_url="http://localhost:1234/v1", api_key="lm-studio")
SYS_PROMPT = '''
你是一名出色的语言学家，你的研究领域是自然语言处理。接下来我将给你一个中文原句与语义角色提取后的句子，你需要仔细理解句子中的各种语法关系，根据原句，判断语义角色提取后的句子是否*完整*、*正确*与*合理*，如果是请回复“T”，否则请回复“F”，请不要回复其他内容。
'''

def filter_triples(sent, triples):
    results = []
    for i in triples:
        srl_text = ""
        if i[0] == "" or i[1] == "" or i[2] == "":
            continue
        srl_text += i[0] + i[1] + i[2]
        USER_INPUT = "原句：" + sent + "\n" + "语义角色提取后的句子：" + srl_text
        completion = client.chat.completions.create(
            model="Qwen/Qwen1.5-7B-Chat-GGUF",
            messages=[
                {"role": "system", "content": SYS_PROMPT},
                {"role": "user", "content": USER_INPUT}
            ],
            temperature=0.3,  # 这里要结合具体模型调整
        )
        if completion.choices[0].message.content == "T":
            results.append(i)
    return results
```

2.2 知识存储

知识图谱中的数据有以下几个特征：

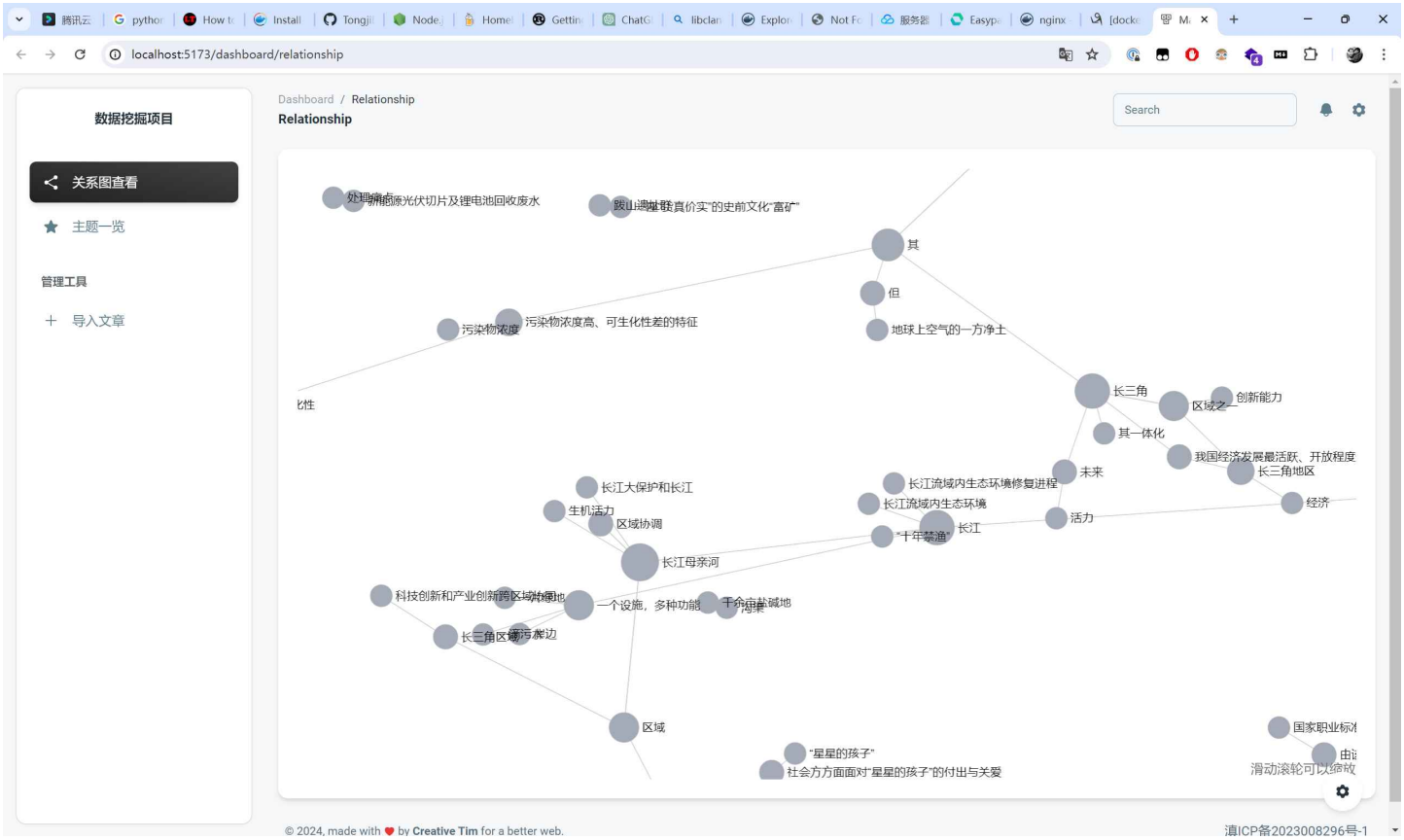
1. 数据规模大：知识图谱按问题领域划分，可分为通用领域和垂直领域，垂直领域知识图谱是基于特定行业数据构建的，规模虽不及通用领域知识图谱，但知识质量高，精度高；而通用领域知识图谱覆盖面更广，规模更大。在新闻文本领域的知识图谱构建中，关系数量和实体数量往往会达到很大的规模。
2. 更新速度快：由于新闻的时效性，知识图谱的数据更新速度要求较高，需要及时更新和维护新闻实体和关系，保证知识图谱的实时性和准确性。
3. 数据关系丰富：知识图谱中的数据通常包含丰富的关系信息，如实体之间的关联关系、实体与属性之间的关系等，这些关系的正确快速地处理、存储和抽取是知识图谱构建过程中的核心问题。

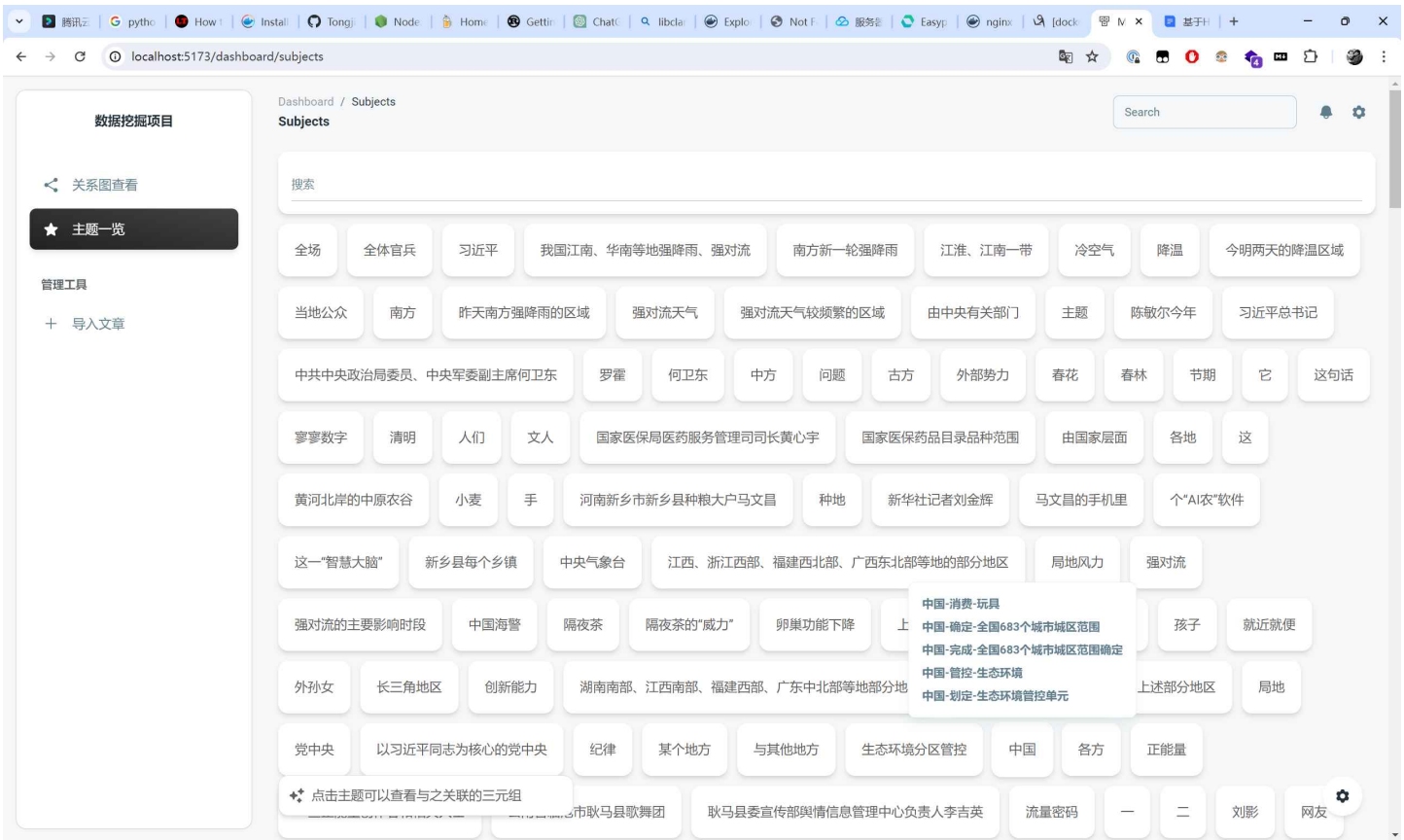
在知识图谱构建的情境下，图数据库是一种非常适合存储知识图谱数据的数据库，它能够高效地存储和查询知识图谱中的实体和关系，提供高效的图查询和图分析功能，满足知识图谱构建的需求。

我们对比多种图数据库工具（如Neo4j、ArangoDB、OrientDB等），最终选用Neo4j图数据库，Neo4j提供的图查询语言Cypher，是为图数据结构定制的，能高效处理复杂的图查询和图算法，同时，Neo4j的图数据库架构支持高并发读写操作，适合处理大规模数据，并能随着数据量增加保持稳定的性能。

2.3 前端应用实现

我们开发了一款用于展示和维护知识图谱、知识节点的前端应用：





技术栈上，我们采用了React作为前端框架。React是一款成熟、稳定、功能强大的前端框架，具有性能优越、功能丰富、易于上手的特点。

3. 图谱评价

一般而言，知识图谱构建的质量通过以下指标进行评判：

1. 准确性：知识图谱中的实体和关系是否准确反映了其数据源中的信息。错误的信息或错误的关系链接会大大降低知识图谱的质量。
2. 完整性：知识图谱是否涵盖了所有相关的实体和关系。一个高质量的知识图谱应该尽可能全面地覆盖特定领域或主题的知识。
3. 一致性：知识图谱中的数据是否一致，没有逻辑矛盾。例如，同一实体的属性在不同部分中应该是相同的。
4. 更新频率：知识图谱的数据应该定期更新，以反映最新的信息和变化。在快速变化的领域，更新频率尤其重要。
5. 互操作性：知识图谱是否能够与其他系统或图谱有效地交互。良好的互操作性意味着知识图谱能够与其他数据源和服务集成，提供更丰富的洞见。
6. 可扩展性：知识图谱应该能够随着数据量的增加而有效地扩展，无论是增加新的实体、关系还是属性。
7. 查询性能：用户或应用程序查询知识图谱的效率和响应时间。一个优质的知识图谱应该能够快速准确地回应查询请求。
8. 用户友好性和可访问性：知识图谱的构建和查询接口是否直观易用，用户是否能够容易地获取和利用图谱中的信息。

本项目的知识图谱构建具有以上标准，但仍具有一些可改进之处：

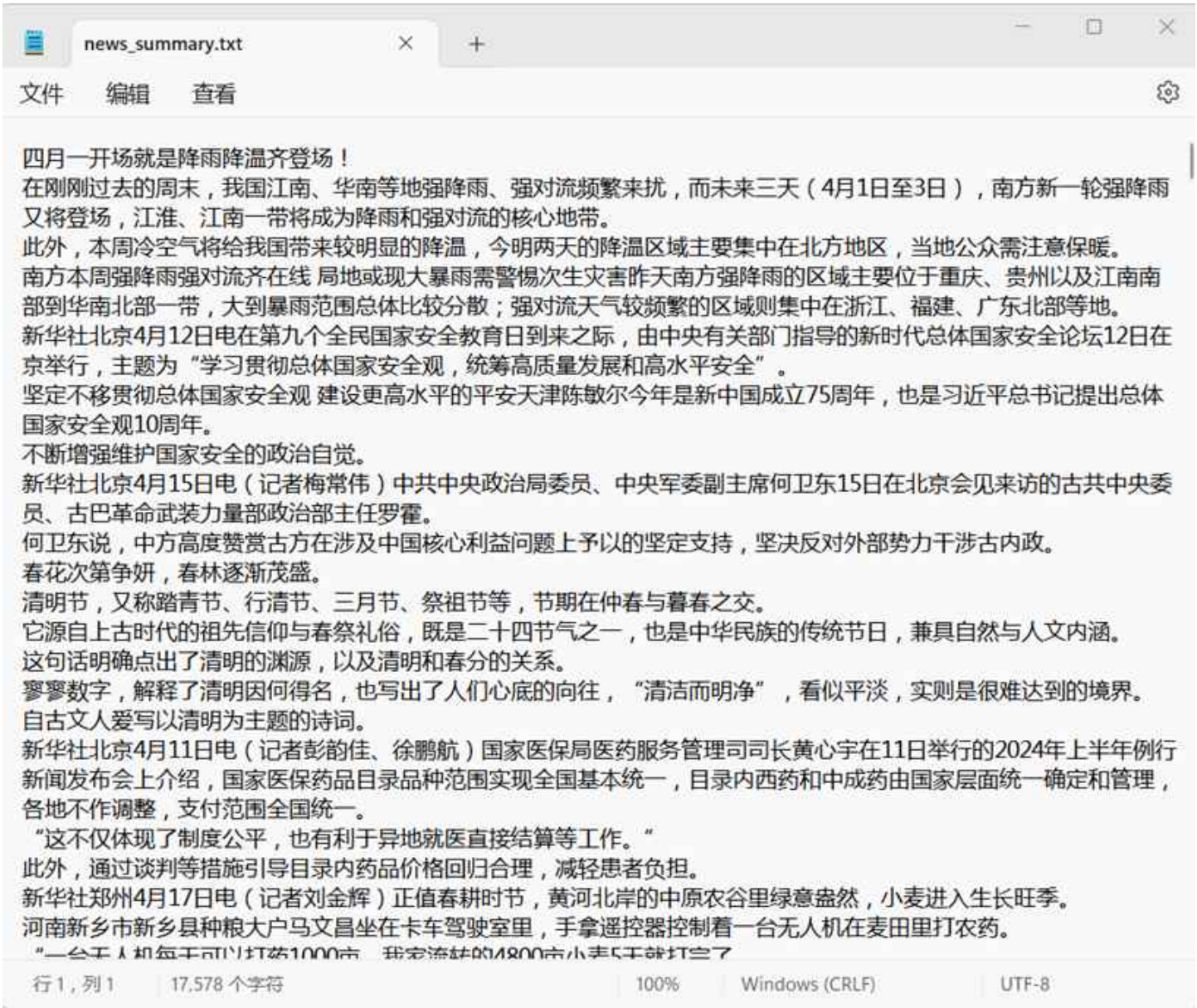
1. 数据清洗和去重：
    - 清理数据中的错误和不一致性，如修正错误的实体属性和关系。
    - 去除重复的实体和关系，确保每个实体和关系的唯一性和准确性。
2. 增强语义连接：
    - 引入更多的本体论和类别系统来增强实体间的语义连接。
    - 通过更复杂的关系类型和属性来增强实体之间的关系表达。
3. 优化查询性能：
    - 改进存储和索引策略，以提高查询效率和响应速度。
    - 使用更高效的图数据库或专用的知识图谱查询语言。
4. 使用机器学习和人工智能：
    - 利用机器学习算法来自动识别和添加新的实体和关系。
5. 增加交互性和可视化：
    - 提供更多的交互式工具和可视化功能，使用户能够更直观地探索和理解知识图谱。
6. 持续维护和更新：

- 定期对知识图谱进行维护和更新，确保其反映最新的知识和信息。
- 建立反馈机制，让用户能够报告错误和建议，不断优化知识图谱的质量。

未来将继续在以上的部分进行改进，以期能够做到一个较为完善的新闻知识图谱系统

4. 结果展示

- 新闻数据提取摘要


news_summary.txt

文件 编辑 查看

四月一开场就是降雨降温齐登场！
在刚刚过去的周末，我国江南、华南等地强降雨、强对流频繁来扰，而未来三天（4月1日至3日），南方新一轮强降雨又将登场，江淮、江南一带将成为降雨和强对流的核心地带。
此外，本周冷空气将给我国带来较明显的降温，今明两天的降温区域主要集中在北方地区，当地公众需注意保暖。
南方本周强降雨强对流齐在线 局地或现大暴雨需警惕次生灾害昨天南方强降雨的区域主要位于重庆、贵州以及江南南部到华南北部一带，大到暴雨范围总体比较分散；强对流天气较频繁的区域则集中在浙江、福建、广东北部等地。
新华社北京4月12日电在第九个全民国家安全教育日到来之际，由中央有关部门指导的新时代总体国家安全论坛12日在京举行，主题为“学习贯彻总体国家安全观，统筹高质量发展和高水平安全”。
坚定不移贯彻总体国家安全观 建设更高水平的平安天津陈敏尔今年是新中国成立75周年，也是习近平总书记提出总体国家安全观10周年。
不断增强维护国家安全的政治自觉。
新华社北京4月15日电（记者梅常伟）中共中央政治局委员、中央军委副主席何卫东15日在北京会见来访的古共中央委员、古巴革命武装力量部政治部主任罗霍。
何卫东说，中方高度赞赏古方在涉及中国核心利益问题上予以的坚定支持，坚决反对外部势力干涉古内政。
春花次第争妍，春林逐渐茂盛。
清明节，又称踏青节、行清节、三月节、祭祖节等，节期在仲春与暮春之交。
它源自上古时代的祖先信仰与春祭礼俗，既是二十四节气之一，也是中华民族的传统节日，兼具自然与人文内涵。
这句话明确点出了清明的渊源，以及清明和春分的关系。
寥寥数字，解释了清明因何得名，也写出了人们心底的向往，“清洁而明净”，看似平淡，实则是很难达到的境界。
自古文人爱写以清明为主题的诗词。
新华社北京4月11日电（记者彭韵佳、徐鹏航）国家医保局医药服务管理司司长黄心宇在11日举行的2024年上半年例行新闻发布会上介绍，国家医保药品目录品种范围实现全国基本统一，目录内西药和中成药由国家层面统一确定和管理，各地不作调整，支付范围全国统一。
“这不仅体现了制度公平，也有利于异地就医直接结算等工作。”
此外，通过谈判等措施引导目录内药品价格回归合理，减轻患者负担。
新华社郑州4月17日电（记者刘金辉）正值春耕时节，黄河北岸的中原农谷里绿意盎然，小麦进入生长旺季。
河南新乡市新乡县种粮大户马文昌坐在卡车驾驶室里，手拿遥控器控制着一台无人机在麦田里打农药。
“一台无人机每天可以打药1000亩，我家流转的4800亩小麦5天就打完了

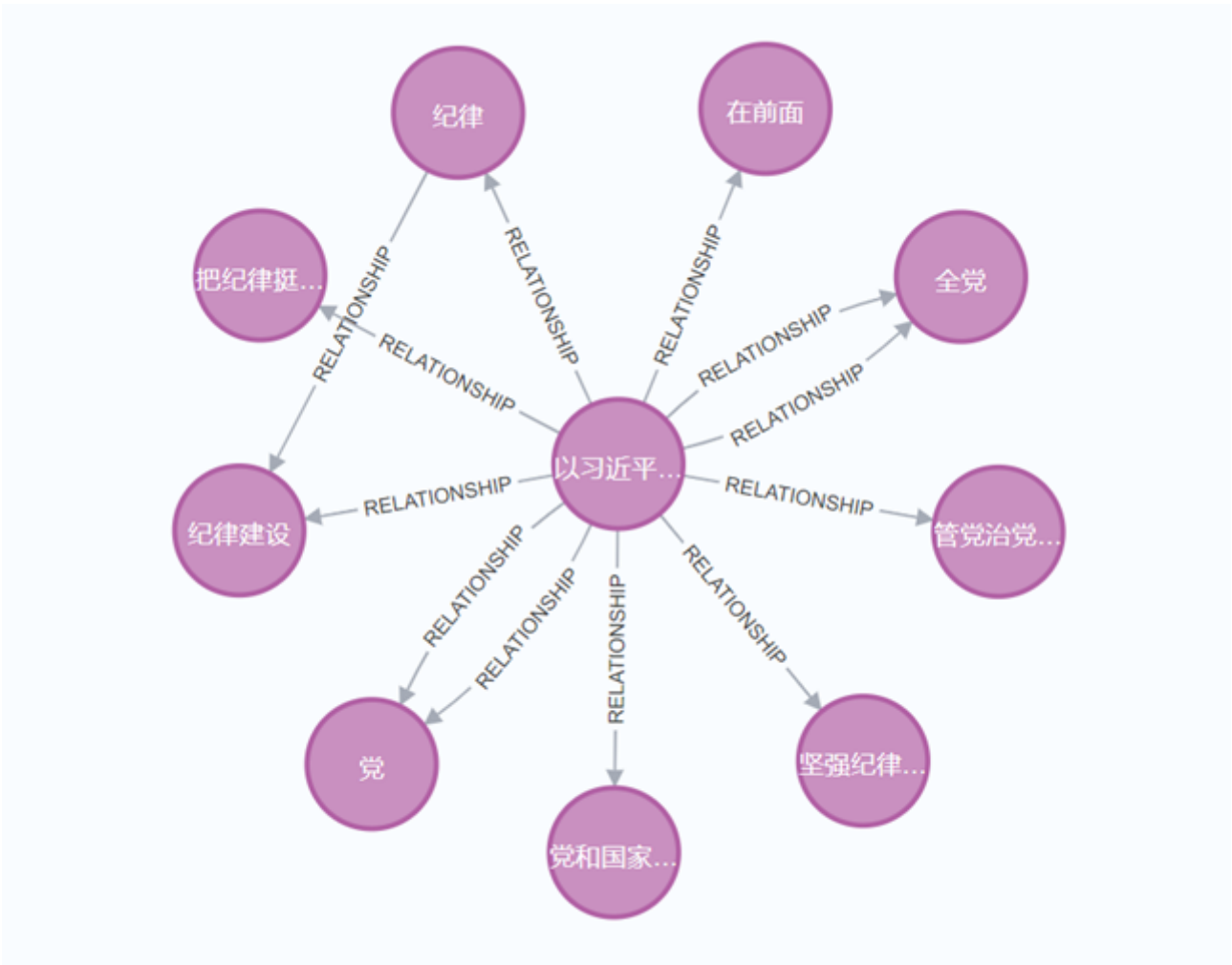
行 1，列 1 17,578 个字符 100% Windows (CRLF) UTF-8

- 关系提取，存入数据库

relation.txt

文件 编辑 查看

[('江淮、江南一带', '成为', '降雨和强对流的核心地带'),
('冷空气', '带来', '较明显的降温'),
('今明两天的降温区域', '集中', '在北方地区'),
('当地公众', '注意', '保暖'),
('南方', '警惕', '次生灾害'),
('昨天南方强降雨的区域', '位于', '重庆、贵州以及江南南部到华南北部一带'),
('强对流天气较频繁的区域', '集中', '在浙江、福建、广东北部等地'),
('由中央有关部门', '举行', '由中央有关部门指导的新时代总体国家安全论坛'),
('主题', '为', '“学习贯彻总体国家安全观，统筹高质量发展和高水平安全”'),
('习近平总书记', '提出', '总体国家安全观'),
('中共中央政治局委员、中央军委副主席何卫东', '会见', '来访的古共中央委员、古巴革命武装力量部政治部主任罗霍'),
('何卫东', '说', '中方高度赞赏古方在涉及中国核心利益问题上予以的坚定支持，坚决反对外部势力干涉古内政'),
('中方', '赞赏', '古方在涉及中国核心利益问题上予以的坚定支持'),
('中方', '反对', '外部势力干涉古内政'),
('春花', '争', '妍'),
('节期', '在', '仲春与暮春之交'),
('它', '源自', '上古时代的祖先信仰与春祭礼俗'),
('这句话', '点出', '清明的渊源，以及清明和春分的关系'),
('国家医保局医药服务管理司司长黄心宇', '介绍', '国家医保药品目录品种范围实现全国基本统一，目录内西药和中成药由国家层面统一确定和管理，各地不作调整，支付范围全国统一'),
('这', '有利于', '异地就医直接结算等工作'),
('这一“智慧大脑”', '整合', '卫星遥感和地面监测站的数据，墒情、虫情、苗情、灾害预警等')
('中央气象台', '发布', '强对流天气橙色预警'),
('中央气象台', '预计', '4月2日20时至3日20时，安徽南部、江苏南部、上海、湖北东南部、湖南东部和中南部、江西、浙江西部、福建西北部、广西东北部等地的部分地区将有10级以上雷暴大风或冰雹天气，局地风力可达12级以上，最大冰雹直径20毫米以上；安徽南部、江苏南部'),
('江西、浙江西部、福建西北部、广西东北部等地的部分地区', '有', '10级以上雷暴大风或冰雹天气'),
('局地风力', '达', '12级以上'), ('强对流', '影响', '强对流的主要影响时段为今天夜间'),
('强对流的主要影响时段', '为', '今天夜间'),
('中国海警', '开展', '维权巡航活动'),
('卵巢功能下降', '导致', '“断崖式衰老”'),
('上海市民吴女士', '来到', '离家步行只需10分钟的上海市普陀区真如镇街道社区“宝宝屋”'),



- 网站展示：dam.accr.cc

5. 新闻知识图谱应用前景展望

新闻知识图谱作为连接数据点、提供深度分析和增强用户体验的技术，未来的应用前景广泛且充满潜力，作为一种新兴的技术，其在新闻领域的应用前景广阔，以下是一些可能的应用方向：

1. 个性化推荐：
   知识图谱通过分析新闻内容中的实体（如人物、地点、组织等）和它们之间的关系，结合用户过去的阅读历史和点击行为，构建起用户的兴趣画像。这使得新闻推荐系统能够根据用户的兴趣偏好提供更加定制化的内容。例如，如果用户经常阅读有关某个特定政治人物的新闻，系统可以优先推荐涉及此人物的最新报道。
2. 事件追踪和串联：
   在发生连续或相关的新闻事件时，知识图谱能够识别并链接各个报道中的相同实体和事件，形成一个连续的事件线索。这对于理解复杂事件的发展过程非常有用，比如在政治选举或自然灾害中，用户可以通过事件图谱追踪整个事件的发展历程。
3. 自动摘要：
   知识图谱提取新闻报道中的核心实体和关键动作，使用这些信息来生成精炼的新闻摘要。这种方法不仅提高了摘要的相关性和信息丰富性，还能确保摘要中包含报道的主要点，帮助用户在有限的时间内获取最重要的信息。
4. 舆情分析：
   通过分析知识图谱中实体的关系和属性，结合自然语言处理技术分析文本情感，可以定量地监测和分析公众对于特定事件的情绪反应。这种分析帮助政府、企业和媒体机构在必要时调整其策略或响应。
5. 错误检测和事实核查：
   知识图谱中包含了大量经过验证的事实数据，新闻报道在发布前可以通过与知识图谱中的数据进行比对，来检查报道中的数据和事实是否准确。这种做法可以有效减少误报和假新闻，提升新闻的公信力。
6. 交互式探索：
   知识图谱不仅是静态的信息存储库，还可以作为一个交互式的探索工具，使用户能够通过点击不同的实体了解更多相关信息。例如，用户可以点击一个企业的名称，探索该企业的历史、关联人物、相关事件等，从而获得更全面的背景信息。