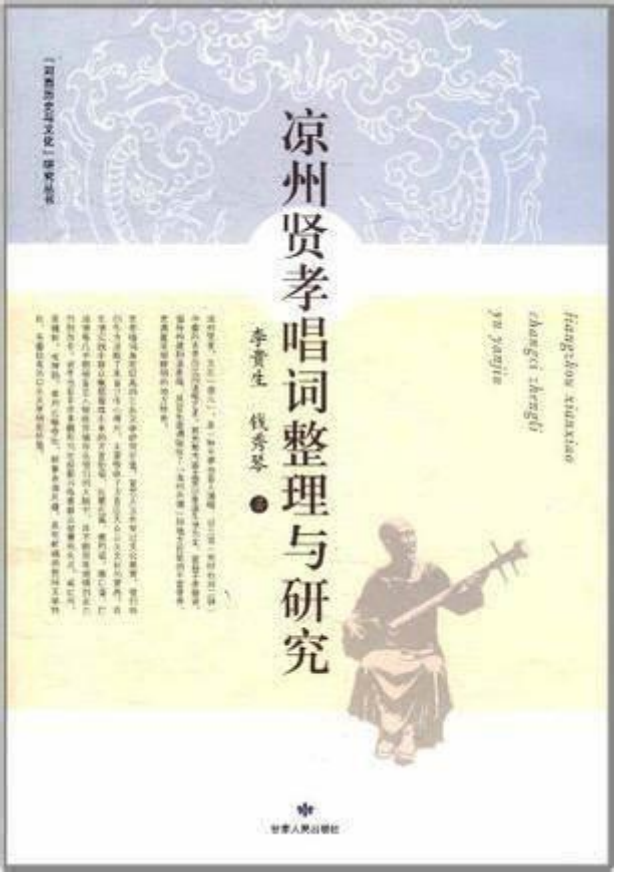
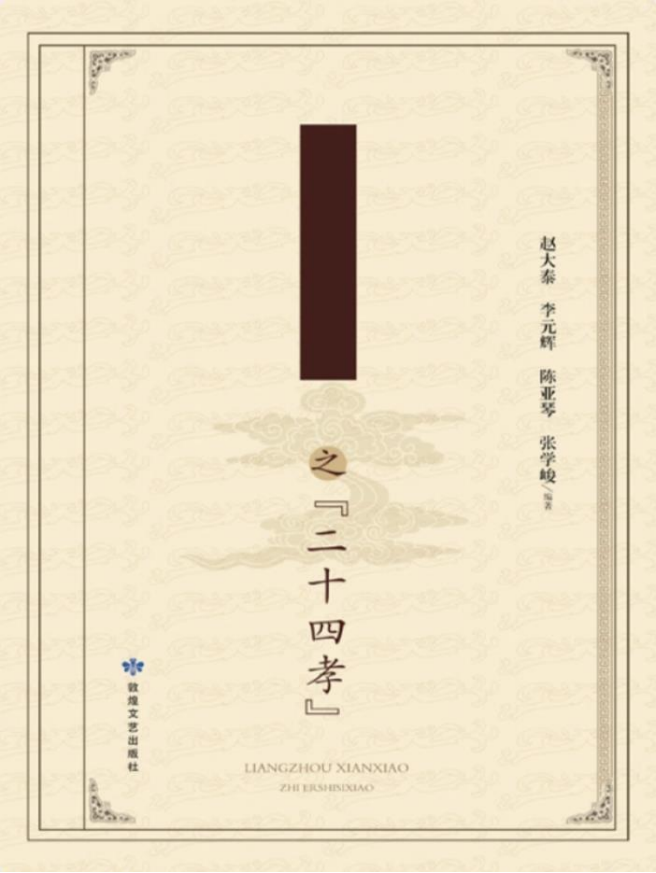
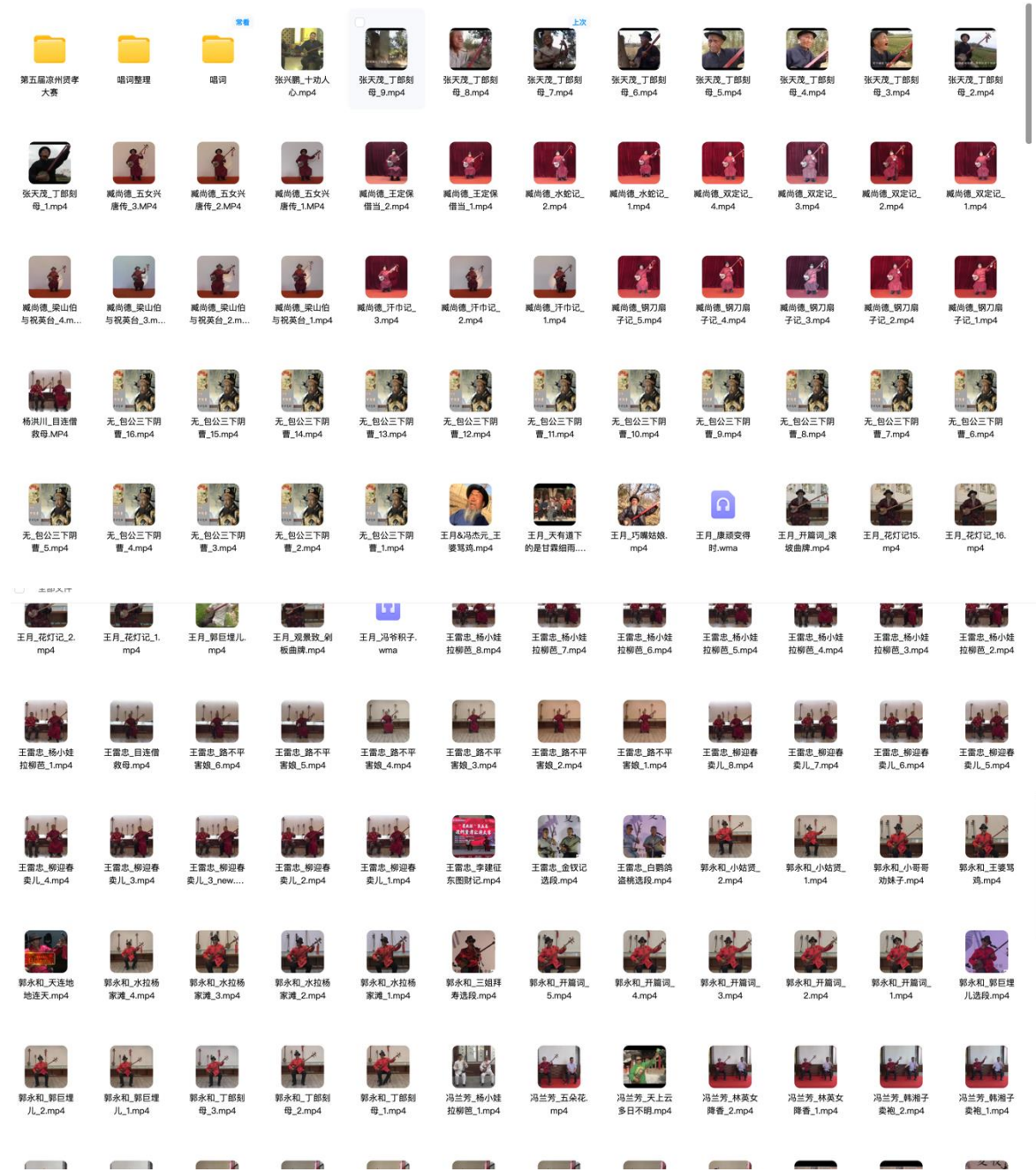
6.8组会





第一章　凉州贤孝曲目考证与赏析

文件名	总字符数	中文字数
二十四孝_1_丁郎刻母_明朝_山东济南.txt	15604	13228
二十四孝_2_杨小娃拉柳笆_王雷忠_明朝_山西林平.txt	8137	6813
二十四孝_3_郭巨埋儿_东汉_河南林州.txt	14884	12456
二十四孝_4_扒肝孝母_明朝_晋中地区.txt	9206	7825
二十四孝_5_三子分财_明朝_甘肃武威.txt	7392	6161
二十四孝_6_路不平害娘_明朝_甘肃武威.txt	21383	17979
二十四孝_7_任仓埋母_明代_四川成都.txt	10302	8691
二十四孝_8_小姑贤_元朝_苏州虎丘南.txt	8300	6962
二十四孝_9_白鹦鸽盗桃_北宋_河南开封.txt	14549	12224
二十四孝_10_三姐拜寿_明朝_山西省黄沙县.txt	11570	9826
二十四孝_11_二郎神劈山救母_明朝_甘肃武威.txt	3743	3146
二十四孝_12_柳迎春卖儿_王雷忠_宋朝_山西壶关县.txt	29427	24477
二十四孝新编_1_虞舜孝感动天_上古五帝时期_山西永济.txt	2171	1833
二十四孝新编_2_汉文帝亲尝汤药_西汉_陕西西安.txt	791	663
二十四孝新编_3_曾参啮指痛心_春秋时期_山东嘉祥.txt	625	493
二十四孝新编_4_仲由百里负米_春秋时期_山东泗水.txt	1471	1210
二十四孝新编_5_郯子鹿乳奉亲_春秋时期_山东郯城.txt	1109	918
二十四孝新编_6_闵子骞单衣顺亲_春秋时期_山东济宁.txt	1309	1086
二十四孝新编_7_老莱子戏彩娱亲_春秋时期_湖北荆门.txt	1075	895
二十四孝新编_8_陆绩怀橘遗亲_三国时期_江西九江.txt	937	777
二十四孝新编_9_王祥卧冰求鲤_晋朝_山东临沂.txt	1459	1215
二十四孝新编_10_黄香扇枕温衾_东汉_湖北云梦.txt	1252	1059
二十四孝新编_11_孟宗哭竹生笋_三国时期_湖北孝感.txt	1324	1091
二十四孝新编_12_董永卖身葬父_西汉_山东博兴.txt	2014	1682
三十六记_9_水蛇记_唐朝_江南水乡.txt	12889	10539
三十六记_3_灯盏记_贾福德_唐朝_山东五常.txt	10089	8026
三十六记_7_梁山伯与祝英台_臧尚德_东晋_浙江.txt	15722	12587
三十六记_6_吕祖买药_李林洪_唐朝_浙江杭州.txt	9599	7766
三十六记_5_韩湘子探家_唐朝_陕西西安.txt	2868	2322
三十六记_10_关云长单刀赴会_三国时期_湖北荆州.txt	2205	1671
三十六记_8_康王变得时_王月_唐朝_中原.txt	9563	7665
三十六记_1_图财记_贾福德_宋代_中原.txt	21280	16694
三十六记_2_水拉杨家滩_郭永和_清朝_河南洛阳.txt	17642	14664
三十六记_4_韩湘子卖袍_唐朝_陕西西安.txt	3814	3185
无_0_小哥哥劝妹子_郭永和_清朝_甘肃武威.txt	870	733
无_0_目连僧救母_王雷忠_东汉_印度比哈尔邦.txt	1277	1077



研究贤孝保留了多少文化内核

1.寻找不同贤孝作品叙事单元承转起合的共性，归纳叙事逻辑下吸引人的原因（和童话故事、歌剧比较）

2.叙事单元中曲牌匹配（悲音/苦音/甜音/哭音/平述音/紧述音/长述音/起述音/贤孝调/古儿词调）根据曲牌区分不同情绪在故事（叙事单元占比，引入曲调元素）

3.思想源流分析，从（儒家/佛教/道教/三教融合）三个方面分析句子出处（三教关键词出发），可以适当结合故事发生时间和地点分析宗教影响，或者分析思想源流对在故事情节发展中的推动作用

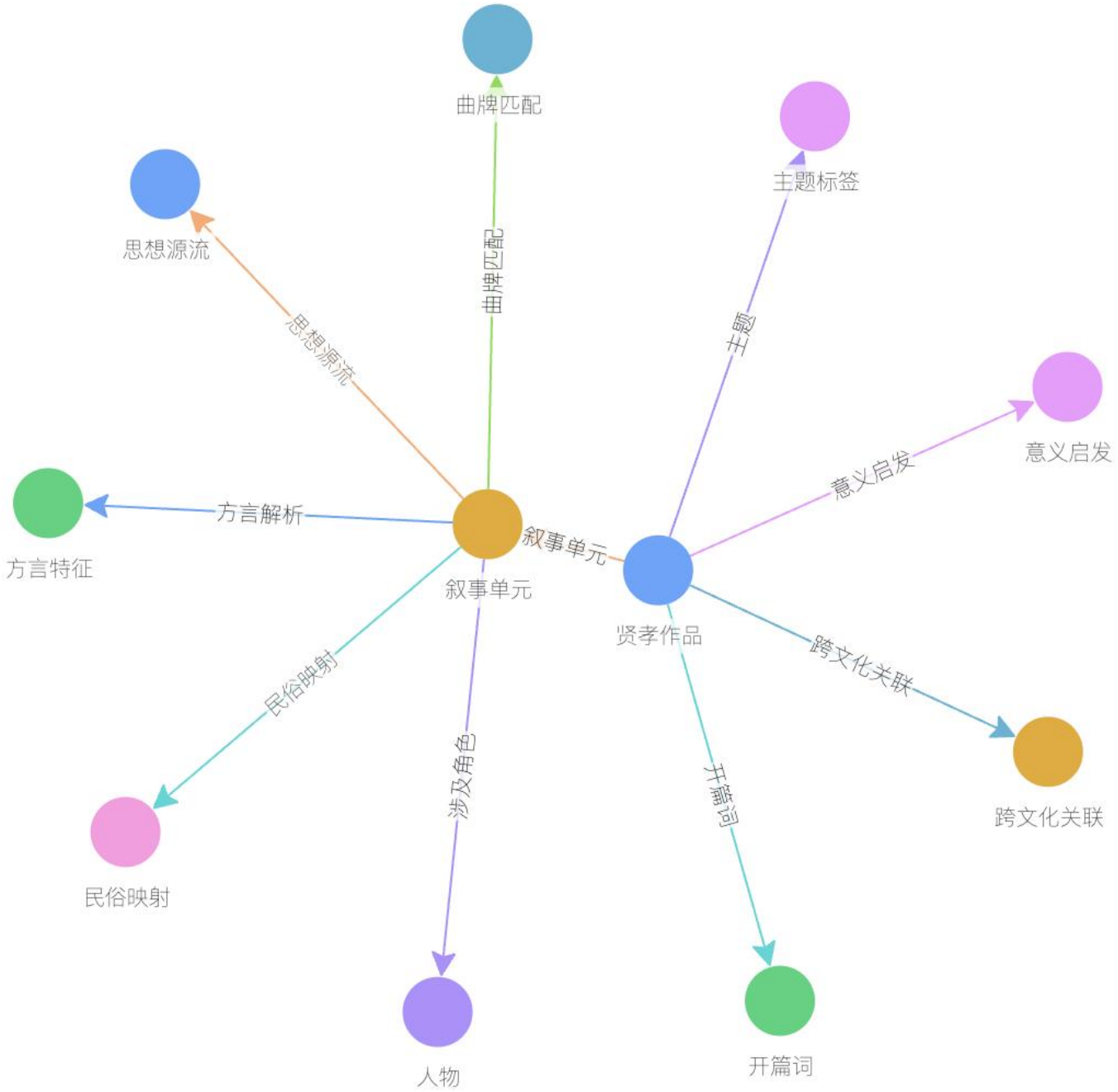
4.方言特征，研究方言俚语的艺术特征（同样意思使用俚语的地方和不使用俚语地方对比）、

5.民俗映射，从（婚俗/丧仪/宗法/节庆）四个方面寻找民俗特征，（构建民俗实体词典）

6.人物形象塑造，从人物形象塑造上分析（孝子/贤媳/恶婆/等）民间感兴趣话题的一致性。从角色功能模型（普洛普理论）出发，归纳贤孝

7.从主题/开篇词中分析伦理主题（孝道/贤德/因果报应）和社会映射（阶级矛盾/女性地位/礼法制度）

8.跨文化互联分析：从跨文化关联和同源故事中分析文化同源性



角色功能模型（普洛普理论）

2. 角色（Role）的抽象化

- 普洛普将纷繁的角色简化为**7种固定类型**，每种角色对应特定功能范围 6 7 11：

角色类型	功能范围	典型人物
反派	制造冲突、伤害主角	恶龙、巫婆
施助者	提供魔法工具或信息	智者、精灵
帮手	协助主角完成任务	战友、动物伙伴
公主（被寻求者）	被拯救/被追求的对象	公主、落难者
派遣者	派遣主角踏上旅程	国王、父亲
主角（英雄）	承担核心任务并战胜反派	骑士、平民英雄
假主角	冒充英雄最终被揭露	骗子、篡位者

- **关键**：同一角色可承担多个功能（如主角同时完成“战斗”和“胜利”），同一功能也可由不同角色执行（如“施助者”和“帮手”均可提供援助）。

从跨文化关联

郭巨埋儿

日本文化
文献移植：郭巨故事最早于日本南北朝时期（14世纪）传入，收录于《今昔物语》等典籍中，成为日本孝道教育的素材。
本土化改造：部分版本淡化"埋儿"情节，突出"天赐黄金"的因果报应，以适应佛教"孝为功德"的理念
价值观冲突：日本传统更强调"父母在世时尽孝"，而非牺牲后代。

高丽文化
•**文献移植**：元代高丽儒官权准编纂《孝行录》（1346年）
•直接复制宋金元"画像二十四孝"内容，郭巨故事位列其中**6**
•**本土化叙事**：新罗时期《三国遗事》记载"孙顺埋儿"故事，情节与郭巨高度相似（埋儿得金），但主角改为韩国人，体现文化挪用

朝鲜文化
•朝鲜王朝将郭巨故事纳入官方伦理教材《三纲行实图》（1431年），但删除血腥细节，强化"孝感天应"主题**6**
•对比日本，韩国更少批判性重构，倾向保留原故事内核，作为儒家伦理的正面典范。

Total Samples: 1488
Original Classes : 248
Clustered Classes : 271
Excluded Clusters (>15kg): 97
Valid Clusters : 174
Correct Clusters : 162
Incorrect Clusters : 12
Correct Cluster Ratio : 0.9310
Correct Cluster Avg MAE : 20.3992
Global MAE : 21.0421

全局MAE（原始） : 21.0232
取id质心最优MAE平均值 : 20.0947
类内平均预测值MAE : 17.4456
各类MAE最小值平均值 : 7.0153
总样本数 : 1487
有效类别数 : 248

Total Samples: 1542
Original Classes : 257
Clustered Classes : 263
Excluded Clusters (>15kg): 59
Valid Clusters : 204
Correct Clusters : 185
Incorrect Clusters : 19
Correct Cluster Ratio : 0.9069
Correct Cluster Avg MAE : 22.6353
Global MAE : 23.1175

全局MAE（原始） : 23.1240
取id质心最优MAE平均值 : 21.1392
类内平均预测值MAE : 19.8117
各类MAE最小值平均值 : 8.7123
总样本数 : 1541
有效类别数 : 257

Total Samples: 1500
Original Classes : 250
Clustered Classes : 279
Excluded Clusters (>15kg): 108
Valid Clusters : 171
Correct Clusters : 145
Incorrect Clusters : 26
Correct Cluster Ratio : 0.8480
Correct Cluster Avg MAE : 21.2770
Global MAE : 22.1013

全局MAE（原始） : 22.0999
取id质心最优MAE平均值 : 19.8192
类内平均预测值MAE : 18.6185
各类MAE最小值平均值 : 7.6378
总样本数 : 1499
有效类别数 : 250

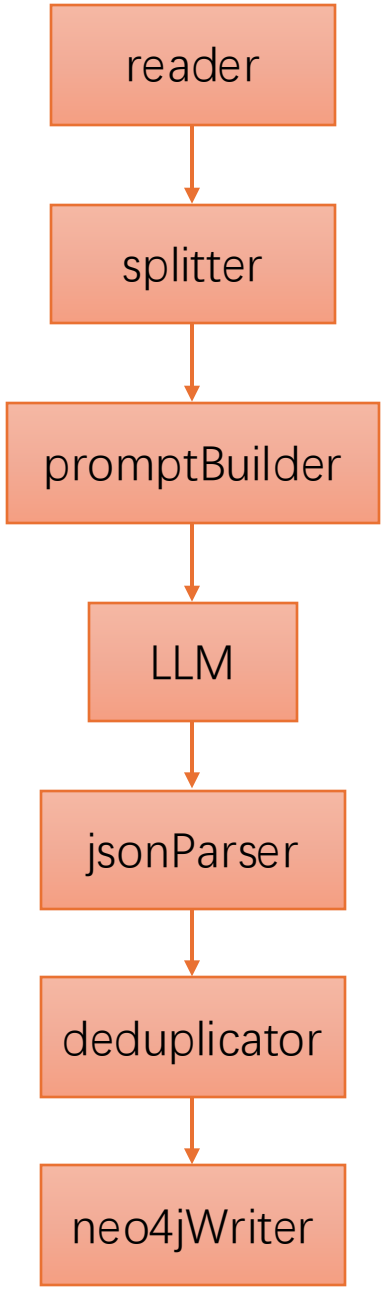
Total Samples: 1578
Original Classes : 263
Clustered Classes : 289
Excluded Clusters (>15kg): 124
Valid Clusters : 165
Correct Clusters : 141
Incorrect Clusters : 24
Correct Cluster Ratio : 0.8545
Correct Cluster Avg MAE : 24.2366
Global MAE : 23.9897

全局MAE（原始） : 23.9740
取id质心最优MAE平均值 : 21.5665
类内平均预测值MAE : 19.9769
各类MAE最小值平均值 : 7.7504
总样本数 : 1577
有效类别数 : 263

Total Samples: 1524
Original Classes : 254
Clustered Classes : 265
Excluded Clusters (>15kg): 65
Valid Clusters : 200
Correct Clusters : 188
Incorrect Clusters : 12
Correct Cluster Ratio : 0.9400
Correct Cluster Avg MAE : 22.3468
Global MAE : 22.3516

全局MAE（原始） : 22.3646
取id质心最优MAE平均值 : 20.3278
类内平均预测值MAE : 18.6195
各类MAE最小值平均值 : 8.2071
总样本数 : 1523
有效类别数 : 254

6.16组会



1.数据阅读器：遍历文件夹，读入txt。返回“文件名:文件内容”

2.文件划分器：将文件进行清洗（去换行、空格），以句号为分界线，20个句子为一组，划分段落。

文学类文档之间没有跳转关系，将所有文档拼接然后按照句子划分，存在拼接处意思误解的可能性。

3.提示词构建器：包含被解析内容、schema定义、例子(json格式)。

Schema格式需要设计成json的风格会跟容易被理解。

4.模型调用器：使用deepseek接口调用deepseek-v3。

基于schema进行信息提取是一个信息密集型的任务，即需要模型针对有限的段落尽可能从多个维度分析，事实证明，r1-14b提取内容少于v3满血版。

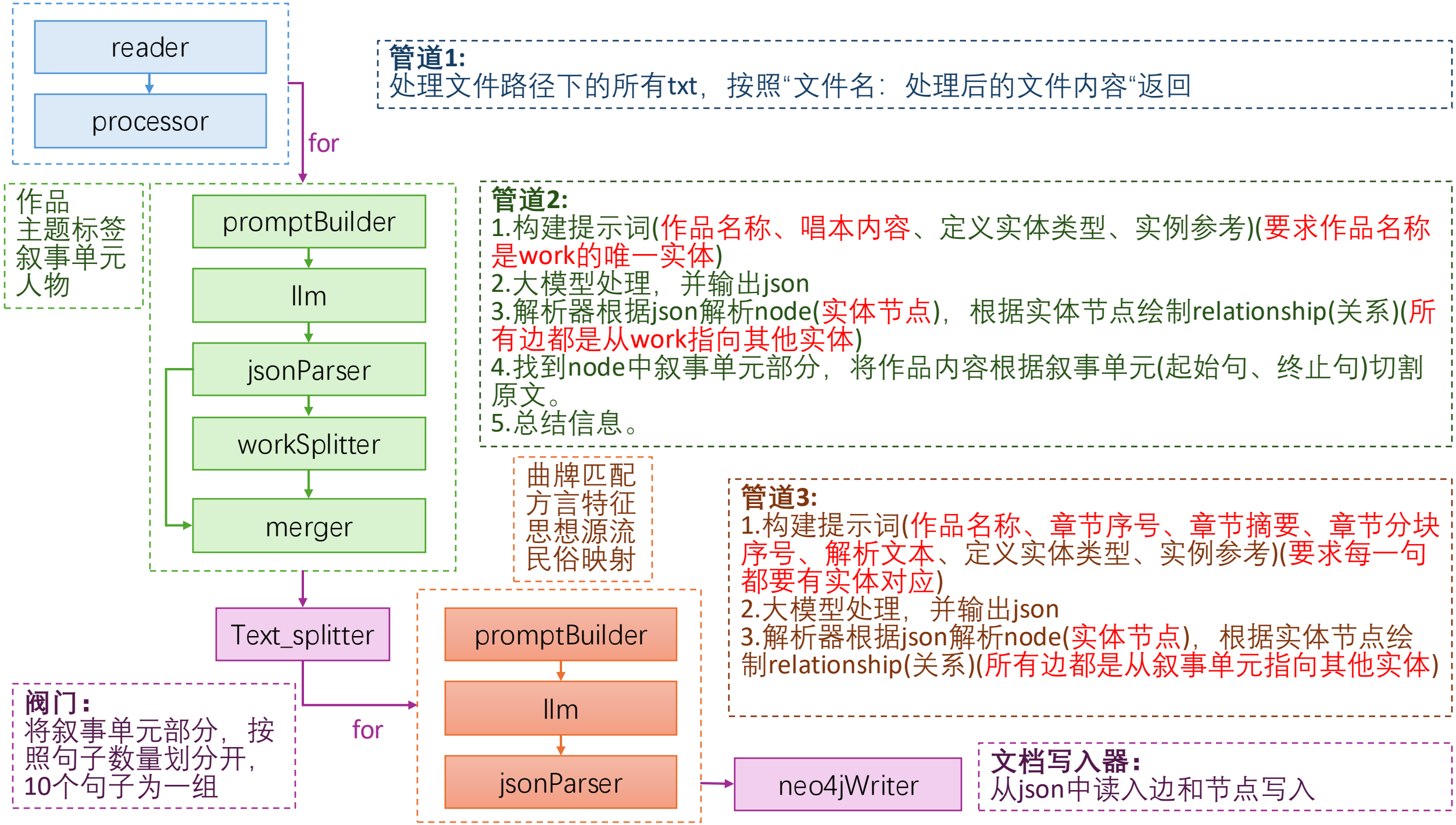
5.Json解析器：从返回内容中解析json格式的内容（包括节点和关系）

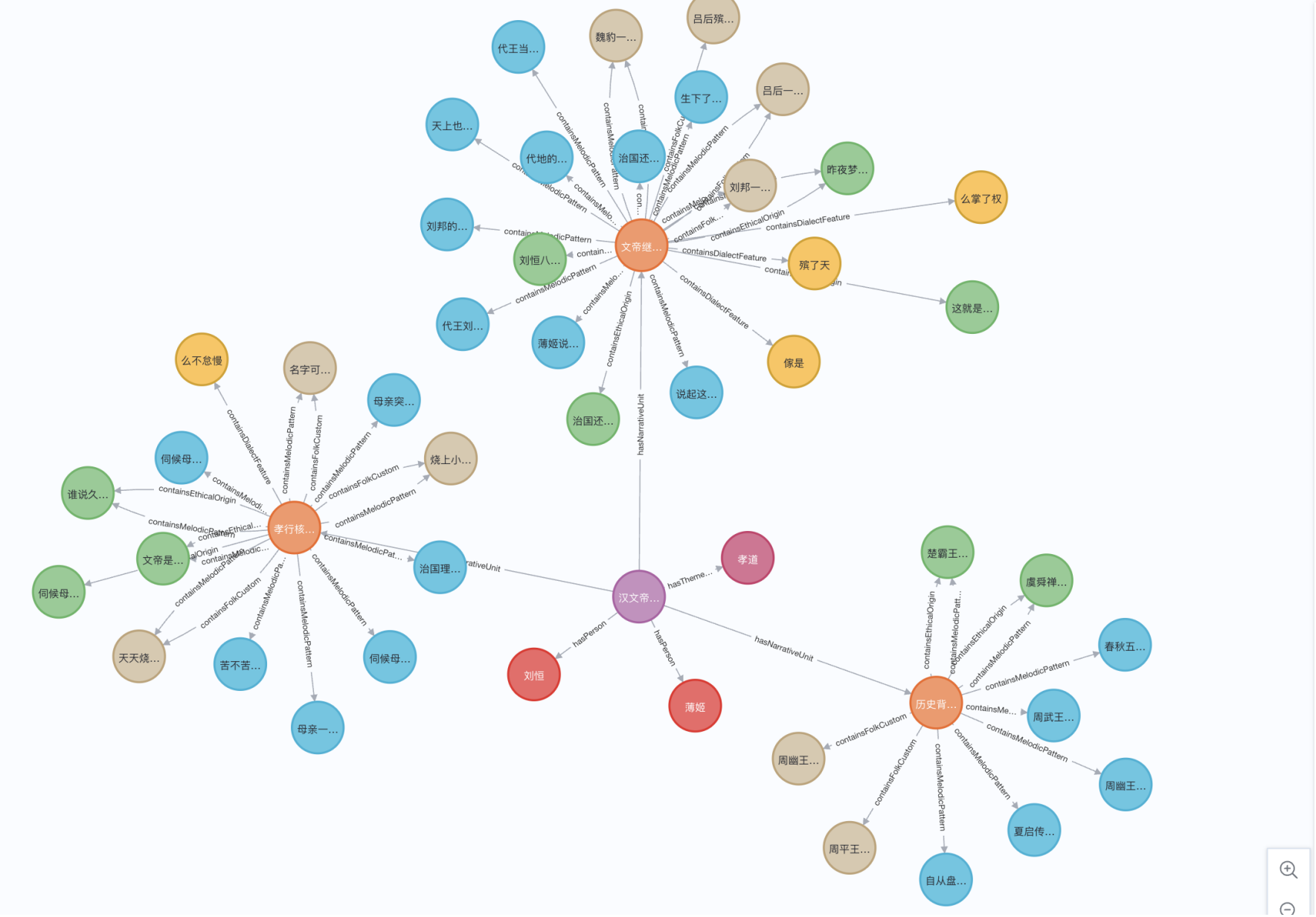
基于schema进行信息提取是一个信息密集型的任务，即需要模型针对有限的段落尽可能从多个维度分析，事实证明，r1-14b提取内容少于v3满血版。

6.重复化简器：将名字相同的实体进行消除

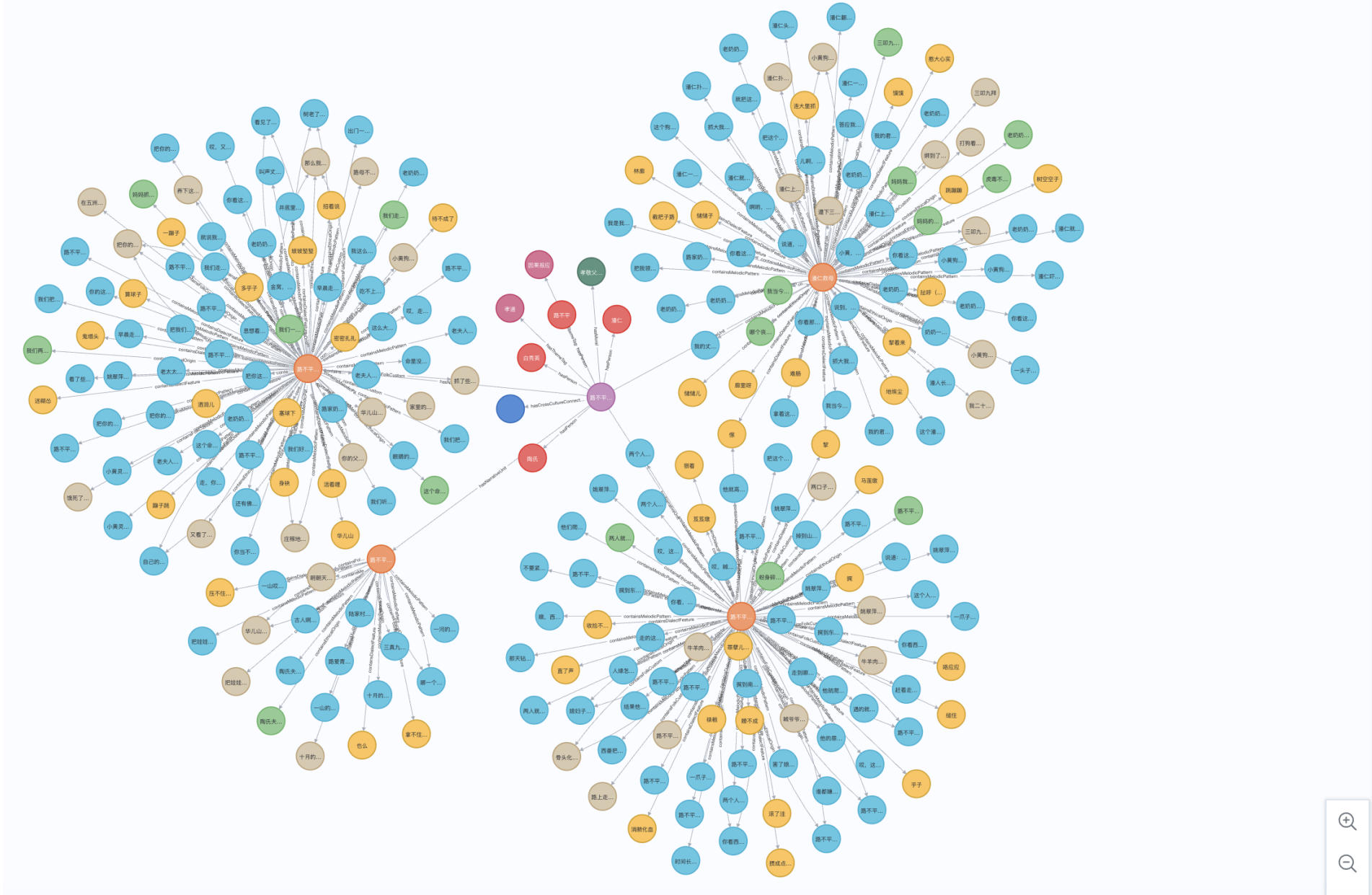
基于schema进行信息提取是一个信息密集型的任务，即需要模型针对有限的段落尽可能从多个维度分析，事实证明，r1-14b提取内容少于v3满血版。

7.图谱知识库写入器：将节点和边写入知识库


reader
processor
for
管道1:
处理文件路径下的所有txt，按照“文件名：处理后的文件内容“返回
作品
主题标签
叙事单元
人物
promptBuilder
llm
jsonParser
workSplitter
merger
管道2:
1.构建提示词(作品名称、唱本内容、定义实体类型、实例参考)(要求作品名称是work的唯一实体)
2.大模型处理，并输出json
3.解析器根据json解析node(实体节点)，根据实体节点绘制relationship(关系)(所有边都是从work指向其他实体)
4.找到node中叙事单元部分，将作品内容根据叙事单元(起始句、终止句)切割原文。
5.总结信息。
曲牌匹配
方言特征
思想源流
民俗映射
Text_splitter
for
promptBuilder
llm
jsonParser
管道3:
1.构建提示词(作品名称、章节序号、章节摘要、章节分块序号、解析文本、定义实体类型、实例参考)(要求每一句都要有实体对应)
2.大模型处理，并输出json
3.解析器根据json解析node(实体节点)，根据实体节点绘制relationship(关系)(所有边都是从叙事单元指向其他实体)
阀门：
将叙事单元部分，按照句子数量划分开，10个句子为一组
neo4jWriter
文档写入器：
从json中读入边和节点写入



汉文帝亲尝汤药 791字



Overview

Node labels

* (299)
Work (24)
NarrativeUnit (7)
EthicalOrigin (39)
ThemeTag (2)
Person (4)
CrossCultureConnection (1)
Moral (1)
MelodicPattern (173)
DialectFeature (49)
FolkCustom (35)

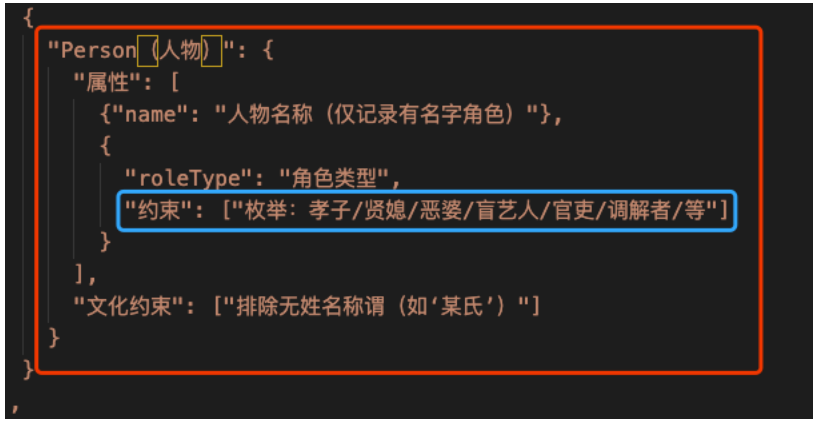
Relationship types

* (312)
hasNarrativeUnit (7)
hasThemeTag (2)
hasPerson (4)
hasCrossCultureConnection (1)
hasMoral (1)
containsEthicalOrigin (38)
containsMelodicPattern (173)
containsDialectFeature (51)
containsFolkCustom (35)

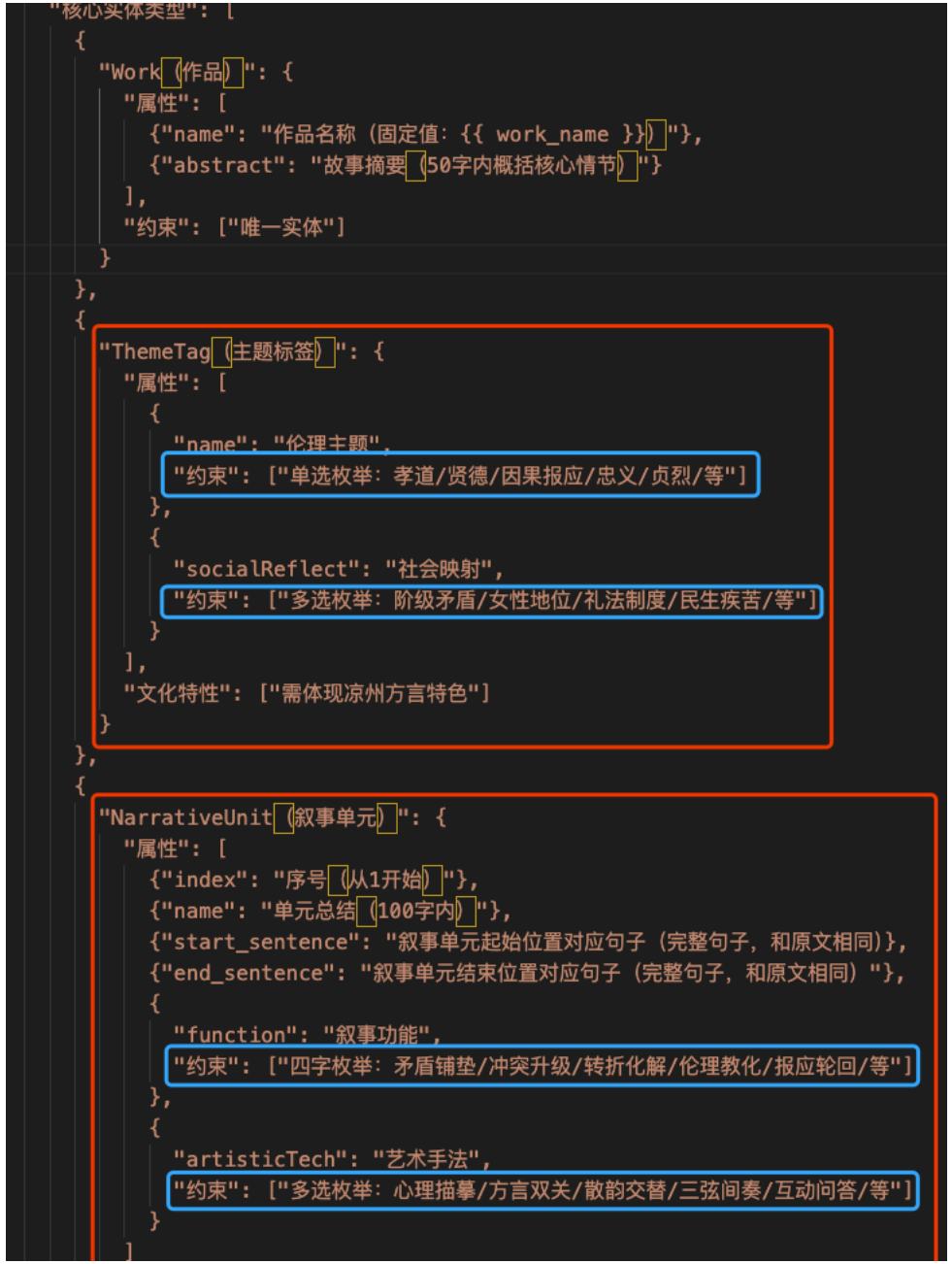
Displaying 299 nodes, 312 relationships.

路不平害娘 21383字

提示词缺乏学术依据：



```
{
  "Person（人物）": {
    "属性": [
      {"name": "人物名称（仅记录有名字角色）"},
      {
        "roleType": "角色类型",
        "约束": ["枚举：孝子/贤媳/恶婆/盲艺人/官吏/调解者/等"]
      }
    ],
    "文化约束": ["排除无姓名称谓（如'某氏'）"]
  }
},
```



```
"核心实体类型": [
  {
    "Work（作品）": {
      "属性": [
        {"name": "作品名称（固定值：{{ work_name }}）"},
        {"abstract": "故事摘要（50字内概括核心情节）"}
      ],
      "约束": ["唯一实体"]
    }
  },
  {
    "ThemeTag（主题标签）": {
      "属性": [
        {
          "name": "伦理主题",
          "约束": ["单选枚举：孝道/贤德/因果报应/忠义/贞烈/等"]
        },
        {
          "socialReflect": "社会映射",
          "约束": ["多选枚举：阶级矛盾/女性地位/礼法制度/民生疾苦/等"]
        }
      ],
      "文化特性": ["需体现凉州方言特色"]
    }
  },
  {
    "NarrativeUnit（叙事单元）": {
      "属性": [
        {"index": "序号（从1开始）"},
        {"name": "单元总结（100字内）"},
        {"start_sentence": "叙事单元起始位置对应句子（完整句子，和原文相同）},
        {"end_sentence": "叙事单元结束位置对应句子（完整句子，和原文相同）"},
        {
          "function": "叙事功能",
          "约束": ["四字枚举：矛盾铺垫/冲突升级/转折化解/伦理教化/报应轮回/等"]
        },
        {
          "artisticTech": "艺术手法",
          "约束": ["多选枚举：心理描摹/方言双关/散韵交替/三弦间奏/互动问答/等"]
        }
      ]
```

研究内容：
1.共性研究：叙事单元的过程（统计方法）
2.共性研究：情绪与曲牌和情节关系（统计手法）
3.共性/差异研究：思想源流的作用（统计手法）
4.共性/差异研究：民俗和地区的映射（统计手法）
5.共性/差异研究：人物形象作用研究（统计手法）
6.差异研究：形成时间，故事演进方面研究：贤孝传播途径以人传为主，在古代没有文本形式留存，相关内容也证明盲人艺人大胆创造故事，吸引人。
凉州宝卷大多都有底本记载，念卷人是“照本宣科”，不需要记忆。宝卷底本既有古老的木刻本，手抄本，也有现代的油印本，打印本。由于有底本存在，宝卷在演化的过程中，与变文之间的承继关系表现更为明显。宝卷的宗教性、仪式感更为强烈。念卷人念卷要进行神圣的仪式：净手——漱口——焚香——念卷。念卷的人通常为当地文化水平高、德高望重的人，念卷时坐于上首，正襟危坐，表情严肃。念卷人唱韵词的时候，旁边的人要“接佛声”，也叫“和佛声”，还叫“接下音子”，和佛声的词一般为“阿弥陀佛弥陀佛”或“南无阿弥陀佛”。
凉州贤孝的表演则更为随意，娱乐性更强。凉州贤孝本来是盲艺人讨饭时依赖的一门技艺，其传承主要是靠盲艺人的口传心授，一般没有文字性的底本。盲艺人在演唱过程中，为了吸引更多的观众，就要增强故事的趣味性、生动性和戏剧性，他们往往会在原有基本情节的基础上进行大胆的个性化创造，可能会采用增加故事情节、利用宗教因果报应的思想、注入地方民俗风情、联系现实、增强细节描摹等手段。

1.围绕贤孝（二十四孝、三十六记寻找各种艺术风格的同源故事）

2.贤孝相关研究划分依据：

伦理主题：孝道/贤德/因果报应/忠义/贞烈/

社会映射：阶级矛盾/女性地位/礼法制度/民生疾苦

叙事单元划分：是否有相关的故事线框架（起承转合）

叙事单元的叙事功能：矛盾铺垫/冲突升级/转折化解/伦理教化/报应轮回

叙事单元包含的艺术手法：心理描摹/方言双关/散韵交替/三弦间奏/互动问答

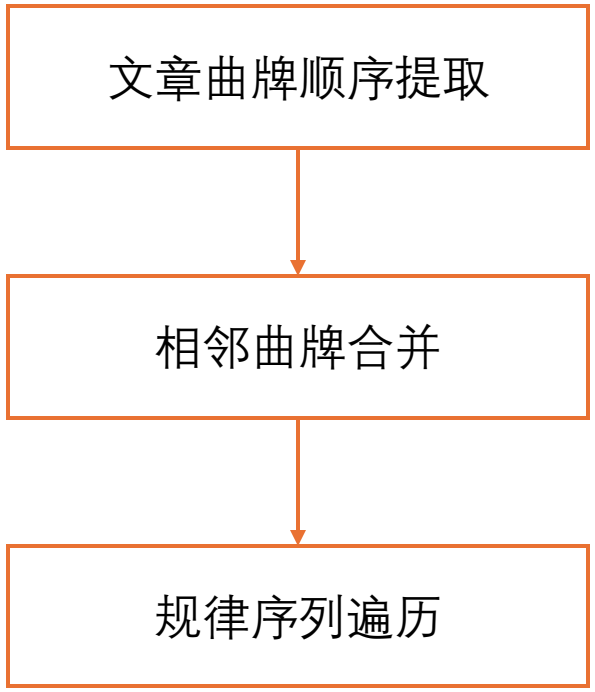
人物角色类型：孝子/贤媳/恶婆/盲艺人/官吏/调解者

句子曲牌类型：根据句子情感能否找到对应曲牌

曲牌划分：悲音/苦音/甜音/哭音/平述音/紧述音/长述音/起述音/贤孝调/古儿词调/河西老调/凉州杂调

民俗映射：婚俗/丧仪/宗法/节庆/农耕礼仪

思想流派：儒家/佛教/道教



1=悲音 2=苦音 3=甜音 4=哭音 5=平述音
6=紧述音 7=长述音 8=起述音 9=贤孝调
10=古儿词调 11=河西老调 12=凉州杂调

序号组合频率排行榜

TOP 1: (3 → 5 → 3) = 45次
TOP 2: (5 → 3 → 5) = 44次
TOP 3: (5 → 6 → 5) = 37次
TOP 4: (2 → 4 → 1) = 37次
TOP 5: (6 → 5 → 6) = 28次
TOP 6: (4 → 2 → 4) = 28次
TOP 7: (5 → 6 → 2) = 28次
TOP 8: (6 → 2 → 6) = 26次
TOP 9: (5 → 2 → 5) = 25次
TOP 10: (3 → 5 → 3 → 5) = 24次
TOP 11: (2 → 5 → 3) = 24次
TOP 12: (5 → 3 → 2) = 24次
TOP 13: (1 → 2 → 4) = 24次
TOP 14: (4 → 1 → 2) = 23次
TOP 15: (9 → 5 → 9) = 23次
TOP 16: (2 → 4 → 2) = 21次
TOP 17: (3 → 5 → 6) = 21次
TOP 18: (2 → 4 → 5) = 21次
TOP 19: (6 → 5 → 2) = 19次
TOP 20: (5 → 2 → 6) = 19次

序号组合频率排行榜

TOP 1: (3 → 5 → 3 → 5) = 24次
TOP 2: (5 → 3 → 5 → 3) = 16次
TOP 3: (2 → 4 → 2 → 4) = 14次
TOP 4: (5 → 3 → 5 → 3 → 5) = 13次
TOP 5: (6 → 5 → 6 → 5) = 10次
TOP 6: (5 → 3 → 5 → 6) = 10次
TOP 7: (1 → 2 → 4 → 1) = 10次
TOP 8: (5 → 6 → 5 → 6) = 9次
TOP 9: (4 → 2 → 4 → 2) = 9次
TOP 10: (4 → 1 → 2 → 4) = 8次
TOP 11: (5 → 6 → 2 → 6) = 8次
TOP 12: (2 → 4 → 1 → 2) = 8次
TOP 13: (2 → 4 → 5 → 6) = 8次
TOP 14: (4 → 2 → 4 → 1) = 7次
TOP 15: (4 → 1 → 4 → 1) = 7次
TOP 16: (3 → 5 → 3 → 5 → 3) = 6次
TOP 17: (2 → 5 → 3 → 2) = 6次
TOP 18: (6 → 5 → 2 → 6) = 6次
TOP 19: (9 → 3 → 5 → 3) = 6次
TOP 20: (2 → 4 → 2 → 4 → 2) = 6次

A Sentence-Based Stylistic History of the Hungarian Novel

基于句子的匈牙利小说风格史

研究核心

论文提出一种**自动识别匈牙利语复合句/复杂句类型**的方法，通过分析连词位置和功能，结合统计分析与语义解读，揭示小说文本的**修辞逻辑结构、认识论态度及文学史风格演变。**

关键创新点

1.**双重分析视角**

1. **统计价值**：连词作为功能词提供大量数据
2. **语义价值**：连词隐含从句间逻辑关系（如转折、条件）

2.**句法-语义结合**

突破传统语法分类（如主从/并列二分法），关注连词的**语法化意义**（如因果、对立）对文本深层结构的塑造。

方法论

1.**语料库**

1. **150部匈牙利经典小说**（1832-2005年），覆盖文学史关键时期
2. 选材标准：再版次数+文学史提及率，确保经典性

2.**从句关系分类**

1. 基于连词位置和语义，划分 **12类从句关系**（如并列、转折、条件、比喻、关系从句等）
2. 正则表达式匹配标点+连词（匈牙利语标点严格标记从句边界）

3.**量化分析**

1. 相对频率（从句类型占比）
2. 统计检验（ANOVA）与可视化（LOESS趋势线、PCA主成分分析）
3. 异常值处理与聚类分析（k-means）

核心发现

1. 文学史风格演变趋势

•**19世纪早期**：

- **关系从句主导**（图4高频区）
- **修辞性"圆周句"**：多层嵌套从句（例：Eötvös小说），增强描述细节与讽刺效果
- 成因：古典修辞教育传统

•**20世纪中期**：

- **比喻从句显著增加**（图6）
- 服务于**内心描写与自由间接引语**（例："他感觉如困玻璃罩中"）

•**20世纪后期**（1970s后）：

- **推理/解释性连词激增**（图5）
- **反讽逻辑**：用严谨推理揭示现实荒诞性（例：Kertész集中营描写中的"自然推论"）

2. 句长演变（图1-3）

•**19世纪**：句长显著下降（线性回归显著）

•**20世纪后期**：两极分化（长句派：Krasznahorkai；短句派：Mándy）

•成因：教育普及、新闻业兴起、书写工具革新

3. 风格传统三维模型（PCA分析）

传统类型	句法特征	代表时期	认识论态度
描述固定传统	高频关系从句	19世纪	语义锚定现实
逻辑开放传统	推理/解释连词主导	20世纪后期	逻辑解构现实
简洁意象传统	短句+比喻从句+非标记关系	20世纪中期	主观感知现实

结论与意义

1.**句法风格反映认识论**

1. 从句类型偏好揭示作者组织现实的方式（如关系从句的"确定性" vs 推理连词的"可重释性"）**量化验证文学史分期**
2. 匈牙利风格演变滞后西欧（如圆周句延续至1868年教育改革）。

2.**作者风格识别局限**

1. 连词分布受创作意识影响（半可控），不适合独立用于作者归属（图8-9聚类效果有限）。

Extracting Geographical References from Finnish Literature

从芬兰文学中提取地理参考：纯文本语料库的全自动处理

研究目标

开发自动化流程，从**1870-1940年芬兰语文学作品**中提取地理信息（地名），通过命名实体识别（NER）、地理编码（Geocoding）和关联开放数据（Linked Open Data）技术，构建交互式文学地图，探索芬兰文学的空间叙事特征。

创新点

1.**小语种适配**：首套针对**芬兰语历史文学**的NER-NEL全流程
2.**文学特异性处理**：
   1. 虚构地名与真实地名的模糊性保留（不强制二值分类）
   2. 复合地名格变解析算法（属格+主格组合）
3.**开源与可扩展**：
   1. 代码公开（GitHub/Zenodo）
   2. 模块化设计支持扩展至瑞典语等北欧语言

关键技术流程

1. 语料库构建
•**数据源**：
   • **Projekti Lonnrot**（志愿者校正的芬兰公共领域文本，噪声少）
   • **Project Gutenberg**（补充文本）
•**筛选原则**：
   • 仅芬兰语原创作品（排除翻译与非虚构）
   • 覆盖小说、戏剧、诗歌等体裁
•**规模**：848部作品，2,035万词汇（表1按体裁与年代分布）
•**挑战**：
   • 编码格式混乱（UTF-8/Win-1252/ISO-8859-1）→ 统一转Unicode
   • 多作品合集的分割（依赖6空行分隔符，需启发式处理）

2. 文本结构化（TEI/XML转换）
•**步骤**：
   • **解析**：用EBNF语法+Instaparse库分割文本
   • **转换**：XSLT将解析结果转为TEI/XML
   • **标注**：为每个词汇分配唯一ID（如lonnrot-0585-1-token3139）
•**成果**：每部作品独立TEI文件，含元数据（作者、标题、年代）、章节结构、词汇级标注。

3. 命名实体识别（NER）
•**工具**：TurkuNLP的芬兰BERT模型（基于OntoNotes分类）
•**实体类型**：

OntoNotes类型	对应TEI标签	示例
GPE（政体地名）	<name type="GPE">	赫尔辛基
LOC（自然地标）	<name type="LOC">	波罗的海
FAC（建筑设施）	<name type="FAC">	教堂

•**词形还原（Lemmatization）**：
   • 解决芬兰语14种格变问题（如"Berliinissa"→"Berliini"）
   • 工具：Turku Neural Parser Pipeline（TNPP）

4. 地名链接（NEL）与地理编码
•**数据库**：Wikidata（替代Getty/GeoNames/DBpedia，因覆盖更佳）
•**流程**：
   • 复合地名处理（如"Suomen Suuriruhtinaskunta"→属格+主格组合）
   • SPARQL查询匹配Wikidata ID（如赫尔辛基→Q984931）
   • 存储坐标（WGS84）与语义关联（国家、地理类型等）
•**成果**：TEI中地名添加ref属性链接Wikidata（例：<name key="/GPE/GPE_Siuntio?wikidata_id=Q984931"...>）

5. 工作流管理
•**挑战**：大规模数据处理与错误恢复
•**方案**：RabbitMQ消息队列系统
   • 分布式任务分配（本地服务器+国家超算中心）
   • 错误隔离：故障文件记录，重试机制

Connecting the Dots Variables of Literary History and Emotions in German-language Poetry

连接点：德语诗歌中文学史和情感的变量

研究目标

通过**量化方法**探究德国诗歌（1850-1920年，现实主义至早期现代主义过渡期）中**情感表达的影响因素**，构建文学史变量（如时期、作者属性、诗歌形式）与情感之间的统计模型，挑战传统文学史研究的定性假设。

方法论创新

1. 数据基础

•**语料库**：6,249首德语诗歌，选自20部当代权威诗集（现实主义8部，现代主义12部）。

•**情感标注**：

- 人工标注1,352首诗，识别6类情感（爱、喜悦、悲伤、愤怒、恐惧、躁动），基于心理学情感模型（Ekman, Plutchik）。
- 训练BERT模型自动标注全语料，F1值0.62-0.79（表1）。

•**特征变量**：

变量类型	具体特征	数据来源
时期	现实主义（1859-1882） vs. 现代主义（1885-1911）	诗集出版时间
作者属性	性别（男/女）、职业（8类，如诗人、哲学家、自然科学家）	德国国家图书馆权威数据
文本特征	主题类型（爱情诗、自然诗等）、押韵比例、诗行平均长度	人工标注 + 工具 Metricalizer（准确率>0.91）

2. 统计模型：贝叶斯分层广义线性模型

•**模型结构**：

- 因变量：6类情感是否存在（二元变量）。
- 自变量：5组特征（时期、性别、职业、主题类型、诗歌形式）。
- **分层设计**：允许变量效应随文学时期变化（如“现代主义爱情诗”的情感模式可能不同于现实主义）。

$$\text{logit}(p) = x_f(s_{f,e} + h_f) + c_f$$

其中 $s_{f,e}$ 表示时期 e 对特征 f 的斜率修正。

•**优势**：处理变量交互效应，避免虚假关联（如“押韵减少恐惧”需控制时期影响）。

关键发现

1. 情感分布特征

•**高频情感**：爱（25%）、喜悦（20%）、悲伤（18%）主导诗歌情感表达（图1）。

•**低频情感**：愤怒（8%）、恐惧（7%）、躁动（6%）较少出现。

2. 变量影响力排序（表5）

通过贡献度公式计算各变量对情感预测的重要性：

$$\text{Score} = \sum_i \left(|p(f)h_{if} + c_{if}|\right) p(em^i)$$

$p(f)$：特征在语料中的频率。

$p(em^i)$：情感在语料中的频率。

1.**主题类型**（Score=0.18）

1. 最强预测因子（如历史题材诗表达愤怒的概率是其他诗的2.59倍，图7）。

2.**作者职业**（0.16）

1. 哲学家写作的诗歌愤怒值显著更高（现代主义时期更明显，图4）。

3.**押韵比例**（0.13）

1. 押韵比例与**恐惧**呈负相关（押韵越少，恐惧表达越多，图6）。

4.**作者性别**（0.12）

1. 影响微弱，反驳“女性诗人更常表达爱”的传统假设（图5）。

5.**诗行长度**（0.10）

1. 效应最弱，与情感关联不显著。

3. 与传统文学研究的对比

•**一致点**：主题类型是核心影响因素（文学史著作常按主题分类）。

•**矛盾点**：

- 职业影响被低估（传统研究忽视职业，但数据显示其重要性堪比性别）。
- 押韵的效应被忽略（数据显示其影响力与性别相当）。

突破性贡献

1.**首建多变量文学史模型**：量化揭示职业、形式特征等“非传统因素”对情感的影响。

2.**方法论创新**：贝叶斯分层模型处理文学变量的交互效应（如时期×主题）。

3.**数据开源**：提供标注数据集与代码

What Do Characters Do? The Embodied Agency of Fictional Characters

角色会做什么？小说人物的具身代理

研究目标

通过**大规模文本分析**探究虚构人物的行为模式，挑战传统文学理论中“小说核心是心理描写”的假设（如Zunshine的“心理理论”），揭示**身体动作**（而非心理活动）才是虚构人物区别于非虚构叙事的核心特征。

方法论创新

1. 数据与工具

•语料库：

- **CONLIT**：2,754部当代英语散文（2001年后出版，含12种文体）
- **Hathi1M**：167万页英语散文随机样本（1800-2000年）

•分析工具：**BookNLP流水线**（专为文学文本优化）

- 步骤：人物识别 → 指代消解 → 主语定位 → 动作动词提取 → **超义原标注**（29类动词标签，如"motion"、"cognition"）
- 准确率：超义原标注整体76%，关键类别（身体/认知）F1值达0.86（表3）

2. 行为分类框架

•具身动作（Embodied Actions）：身体运动（motion）、接触（contact）、身体状态（body）

例：微笑、行走、触摸

•认知动作（Cognitive Actions）：思考（cognition）、情感（emotion）

例：思考、希望、爱

•验证：人工标注500个动词样本，组间一致性Fleiss' κ=0.813，证明分类可靠

3. 统计方法

•差异度量：邓宁对数似然比（Dunning's log-likelihood）量化虚构/非虚构行为差异

•效应量：Cohen's d评估文体影响力（如d=1.82表示具身动作在小说中极显著）

核心发现

1. 虚构人物的核心特征：身体主导

•高频具身动作：虚构叙事中占比47.3%（非虚构仅28.1%），效应量**d=1.82**（图2）

•最显著行为（表2）：

- 接触（如站立、坐下）：G2=+116,743（虚构最显著）
- 身体动作（如微笑）：G2=+47,340
- 感知（如看见）：G2=+39,404

•认知无差异：虚构与非虚构的思考/情感动作频率接近（d=-0.17）

2. 跨文体与历史趋势

•文体差异（图3）：

言情小说具身动作最多（比科幻高21%），第一人称叙事比第三人称多5%

•历史演变（图4-5）：

- 1800-2000年：小说中具身动作增长50%（运动类+46%，身体类+58%）
- 认知动作保持稳定（仅+3%）

3. 对“心理理论”的挑战

•传统观点：小说通过复杂心理描写培养读者“心理理解能力”（Theory of Mind）

•本文反证：虚构人物通过**身体与环境互动**传递心理状态（如“微笑”隐含喜悦，“颤抖”暗示恐惧），而非直接描写内心（图6）

理论贡献

1.**提出“具身代理”模型**：

虚构人物本质是**身体化的行动者**（embodied agents），其行为通过身体与空间交互传递心理状态。

2.**修正文学价值论**：

小说价值在于模拟**具身认知**（embodied cognition）——思维产生于身体与环境的互动，而非抽象推理。

3.**方法论启示**：

需开发新标注框架（如结合动作、环境、对话）量化“心理深度”（图6）。

BookNLP 动词超义原标签（29类）

类别	英文标签	典型动词示例	论文中是否出现
身体动作	body	smile（微笑）、laugh（笑）、wear（穿戴）、sleep（睡觉）	✓
情感	emotion	want（想要）、like（喜欢）、love（爱）、hope（希望）	✓
状态变化	change	start（开始）、die（死亡）、become（成为）、grow（生长）	✓
移动	motion	go（去）、walk（走）、turn（转身）、leave（离开）	✓
认知	cognition	know（知道）、think（思考）、remember（记得）、believe（相信）	✓
感知	perception	see（看见）、look（看）、hear（听见）、feel（感觉）	✓
交流	communication	say（说）、ask（问）、tell（告诉）、call（呼叫）	✓
持有	possession	have（有）、get（得到）、give（给予）、lose（失去）	✓
竞争	competition	fight（战斗）、play（玩）、win（赢）、shoot（射击）	✓
社会活动	social	do（做）、try（尝试）、work（工作）、help（帮助）	✓
消费	consumption	eat（吃）、drink（喝）、use（使用）、need（需要）	✓
静态状态	stative	be（是）、keep（保持）、wait（等待）、live（生活）	✓
接触	contact	stand（站立）、sit（坐）、put（放置）、touch（触摸）	✓
天气现象	weather	rain（下雨）、snow（下雪）、blow（吹）、burn（燃烧）	✓
创造	creation	make（制造）、write（写）、build（建造）、create（创造）	✓
存在	existence	exist（存在）、appear（出现）、disappear（消失）	✗
位置	location	contain（包含）、cover（覆盖）、fill（填充）	✗
量度	quantity	weigh（称重）、measure（测量）、cost（花费）	✗
时间	time	begin（开始）、end（结束）、last（持续）	✗
因果关系	causation	cause（引起）、allow（允许）、prevent（阻止）	✗
感知描述	perception_body	taste（品尝）、smell（闻）	✗
身体内部	body_internal	breathe（呼吸）、bleed（流血）	✗
移动方式	motion_directional	rise（上升）、fall（落下）	✗
社会关系	social_relation	marry（结婚）、divorce（离婚）	✗
认知过程	cognition_epistemic	doubt（怀疑）、prove（证明）	✗
情感表达	emotion_expression	cry（哭）、sigh（叹气）	✗
言语行为	communication_speech	whisper（低语）、shout（喊叫）	✗
持有转移	possession_transfer	buy（买）、sell（卖）	✗
天气影响	weather_atmospheri	freeze（冻结）、melt（融化）	✗

Repetition and Innovation in Dramatic Texts: An Attempt to Measure the Degree of Novelty in Character's Spee

戏剧文本中的重复与创新：衡量人物言语新意程度的尝试

研究目标

通过**句嵌入技术**量化莎士比亚戏剧人物台词的创新性（Innovation）与重复性（Repetition），揭示人物在戏剧信息流中的功能差异，并构建新型人物关系网络。

方法论创新

1. 核心指标：最大余弦相似度（MCS）

•**技术基础**：使用 **SBERT 模型**（all-MiniLM-L6-v2）生成句子嵌入向量

•**计算逻辑**：

对目标角色的每句台词，计算其与**源角色所有先前台词**的余弦相似度，取**最大值**作为 MCS 值

→ *低 MCS = 高创新性*（台词语义新颖）

→ *高 MCS = 高重复性*（台词语义相似）

•**数据处理**：

- 过滤短句（<4词，如"Yes, sir"）
- 按幕（act）加权校正时序偏差（图2c）

2. 人物网络构建

•**节点关系**：箭头从 **创新者（Source）** 指向 **重复者（Target）**

- 边权重 = 平均 MCS（值越大表示重复性越强）
- 节点大小 = 角色作为 Source 的频率（创新性得分）

•**网络特性**：

- 使用 **ForceAtlas2 算法**布局
- 揭示"谁重复谁"的层级关系（图3）

3. 语料与验证

•**数据来源**：莎士比亚戏剧 TEI-XML 语料库（DraCor）

•**人工验证**：

- 标注 500 组句子对，Fleiss' κ = 0.813（高一致性）
- 定性分析极端 MCS 值句子（表1-5）

核心发现

1. 创新性角色的两种类型

类型	代表角色	低 MCS 台词特征	功能
事件报告型	霍拉旭（《哈姆雷特》） 卡西乌斯（《凯撒大帝》）	描述具体事件 *例："钟敲了三下"*	推动情节发展 连接不同场景
表达方式型	哈姆雷特（《哈姆雷特》） 赫米娅（《仲夏夜之梦》）	表达怀疑/情感/抽象隐喻 *例："时间脱节了"*	挑战既定认知 深化主题内涵

2. 体裁差异与性别模式

•**悲剧/非喜剧**：

- 创新性呈**层级化分布**（男性主导，如哈姆雷特、麦克白）
- 网络脆弱：核心角色失败导致叙事崩溃

•**喜剧**：

- 创新性呈**环形分布**（女性角色更创新，如《皆大欢喜》罗莎琳德）
- 网络抗毁性强：多重路径维持信息流
- **数据支撑**：14部喜剧中，6部女性创新性最高（图3）

3. 创新性动态规律

•**抽象 vs 具体**：

抽象/诗性台词创新性更高（如奥赛罗结局台词："贞洁的星辰啊，莫让我向你们说出它的名字"）

•**权力操控**：

伊阿古（《奥赛罗》）通过**语言控制**使他人重复自己台词（例："人应当表里如一" → 奥赛罗重复）

理论贡献

1.**提出"语义差异"驱动创新**：

台词创新性取决于其与先前话语的**语义距离**（非词汇重复），通过句嵌入量化实现。

2.**构建新型人物网络**：

突破传统"共现网络"，用 **MCS 定向关系**揭示信息流动方向。

3.**修正莎士比亚性别表征认知**：

证明喜剧中女性角色是核心创新者，挑战悲剧男性中心的叙事模式。

Small Worlds: Measuring the Mobility of Characters in English-Language Fiction

小世界：衡量英语小说中人物的流动性

方法论创新

数据与工具

•语料：13,383本英语书籍（1789-2021），含小说/非虚构（表1）

•技术栈：

- **BookNLP**：识别角色、地点共现（10词窗口）
- **BERT关系分类**：过滤“位于”(IN)关系（如“她在厨房”）

•空间分类：

- **地理实体（GPE）**：可测绘地点（纽约、加州）→ 计算球面距离
- **通用空间（Generic）**：功能地点（房间、街道）→ 用GloVe词向量测语义相似度

关键指标

指标	定义	揭示问题
移动距离	主角跨GPE的球面距离总和	虚构角色移动范围更小
跳跃次数	地点变更次数（如伦敦→巴黎=1跳）	虚构路线固化（表3高频路线）
语义距离	通用空间词向量余弦相似度均值	虚构空间语义单调（厨房→卧室相似度高）
指代词比例	“这里/那里”在通用空间的占比	虚构文本更依赖模糊空间指代

颠覆性发现

1. 性别与移动性的反常识结论

•移动距离无性别差异（表4）：

- 男性角色距离均值 31,134英里 vs 女性 29,943英里（p=0.199）
- **推翻“女性角色被空间限制”假设**

•但空间类型分化显著（图3c-d）：

- 女性关联**私密空间**（厨房、卧室 z-score↑）
- 男性关联**权力场所**（白宫、战场 z-score↑）

2. 虚构性的空间表征差异

•虚构文本：

- 高频词：厨房、车道、外星（Valhalla, Mars）→ **私密与幻想空间**
- 依赖指代词（“这里”频率↑）

•非虚构文本：

- 高频词：白宫、参议院、巴格达→ **权力与冲突空间**（图3a-b）

核心结论：文学中的“小世界效应”

1.**地理局限**

1. 虚构角色移动距离**不足非虚构角色的一半**（均值：38,024英里 vs 131,263英里）
2. 路线高度固化：75%移动集中于欧美核心城市（纽约-伦敦-巴黎三角，图2）

2.**语义局限**

1. 虚构文本**依赖通用空间**（厨房/卧室/街道），地理实体（GPE）占比低（Generic/GPE ratio↑）
2. 空间语义相似度高（室内场景重复出现）

3.**历时性扩张有限**

1. 1789-2021年间角色移动距离翻倍（图1），但主要因**文本变长**（控制文本长度后增幅微弱，图4）

Towards an Event Based Plot Model: A Computational Narratology Approach

基于事件的情节建模：一种计算叙事学方法

核心目标

提出一种**基于事件（Event-based）**的计算叙事学方法，用于建模情节（Plot），核心是将**叙事性（Narrativity）**和**可述性（Tellability）**操作化，通过事件标注与量化分析实现情节可视化。

关键概念

1.**事件分类**

1. **事件 I**：基础事件，指文本中明确或隐含的**任何状态变化**（如“格里高尔变成甲虫”）。
2. **事件 II**：需满足额外条件（如意外性、重要性），依赖主观解读（如“父亲用苹果砸伤格里高尔”代表父子冲突升级）。
3. **操作化方案**：
    1. **状态事件**（静态描述，叙事性=2）
    2. **过程事件**（动态但无状态变化，叙事性=5）
    3. **状态变化事件**（关键转折，叙事性=7）
    4. **非事件**（无虚构世界指涉，叙事性=0）

2.**叙事性（Narrativity）**

1. 定义为事件的**可叙述程度**，具有**标量特性**（如状态变化事件 > 过程事件 > 状态事件）。
2. 通过事件标注生成**叙事性时间线**（Narrativity Graph），刻画情节起伏。

3.**可述性（Tellability）**

1. 事件是否“值得讲述”的程度，通过**读者摘要**量化：
    1. 收集多份读者摘要（每文本9-11份），标注摘要中提及的文本段落。
    2. 计算每个事件在摘要中被提及的频率，作为可述性值。

方法论

1.**事件标注流程**

1. **单元**：以限定动词为核心的短语（如“他醒来发现自己变成了甲虫”）。
2. **分类规则**：基于动词类型判断事件类别（如“醒来”属状态变化，“思考”属过程事件）。
3. **自动化**：模型已实现事件自动标注（F1值达0.71）。

2.**叙事性时间线生成**

1. **步骤**：
    1. 为每个事件分配叙事性值（如状态变化=7）。
    2. **加权平滑处理**：使用余弦加权窗口（默认100个事件宽度）生成连续曲线，弱化局部波动。
2. **输出**：可视化曲线，峰值对应关键情节（如《变形记》中格里高尔死亡事件）。

3.**模型优化**

1. **目标**：使叙事性时间线与读者摘要的可述性分布一致。
2. **方法**：
    1. 调整事件类型的叙事性权重（如测试状态变化事件值=3/7/50）。
    2. 优化平滑窗口大小（测试10-190事件宽度）。
3. **关键发现**：
    1. 最优窗口：**50-100个事件**（过小则噪声多，过大则细节丢失）。
    2. 事件权重需显著区分：**状态变化事件 > 过程事件 > 状态事件**（如7-5-2优于等权重）。

理论创新

1.**话语层（Discourse）导向**

1. 聚焦事件**如何被表述**（而非虚构世界“发生了什么”），避免过度依赖主观解读（如《变形记》开篇的“变形”仅标注为感知过程事件）。

2.**情节的双重建模**

1. **微观**：事件级别的叙事性标量分析。
2. **宏观**：通过摘要可述性识别核心情节（如高频提及的事件=关键转折点）。

Operationalization and Interpretation Dependence in Computational Literary Studies

计算文学研究中的操作化与解释依赖性

在计算文学研究（CLS）中，**操作化**指将抽象的文学概念转化为可计算或可测量的具体步骤的过程。其目的是让文学理论中的概念（如“不可靠叙事”）能被计算机或研究者系统化地识别和分析。

解释依赖性是文学概念或文本陈述的核心属性，指其**依赖主观解释的程度**

核心论点

1.**解释依赖性阻碍操作化**:
高解释依赖性概念（如不可靠叙事）难以被完全还原为可计算规则，因其依赖主观推理与语境。
2.**操作化需兼顾理论与实践**:
1. 定义阶段需平衡精确性与文学研究意图（避免过度简化）。
2. 决策步骤需结合实践反馈（如标注争议揭示隐藏假设）。

3.**当前局限性**:
完全的计算操作化对高解释依赖性概念尚不可行，需接受**近似操作化**并明确其与原概念的关联。

方法论建议

1.**迭代协作**：通过标注实践与论证分析，逐步优化操作化流程。
2.**分轨策略**:
1. 人类操作化（核心/深度轨）注重理论严谨性。
2. 计算操作化（近似轨）作为补充，需明确其近似性。

3.**评估标准改革**：用“论证合理性”替代传统指标（如标注者一致性），以兼容文本模糊性。