数字人文

——数字环境下的人文学科

王军

北京大学数字人文实验室

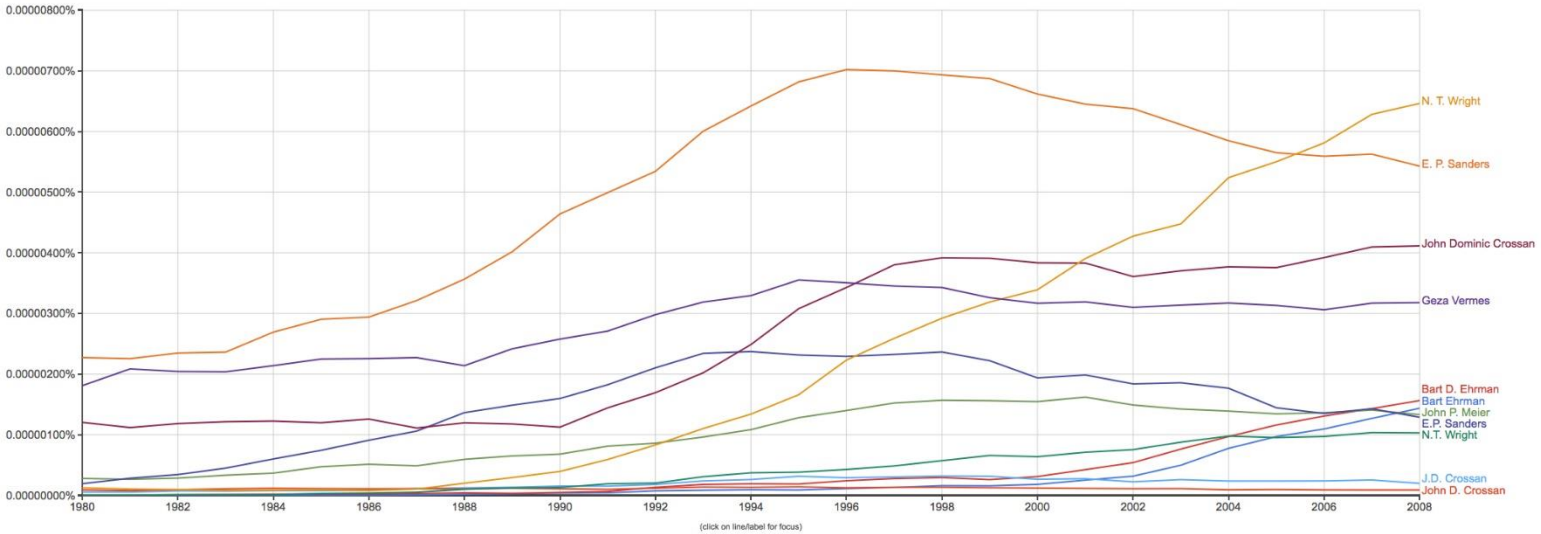
北京大学信息管理系

1. 何谓“数字人文”

- 信息环境的演变
  - 纸质文献环境 → 数字化网络化信息环境
  - 数字化网络化的学术生态环境

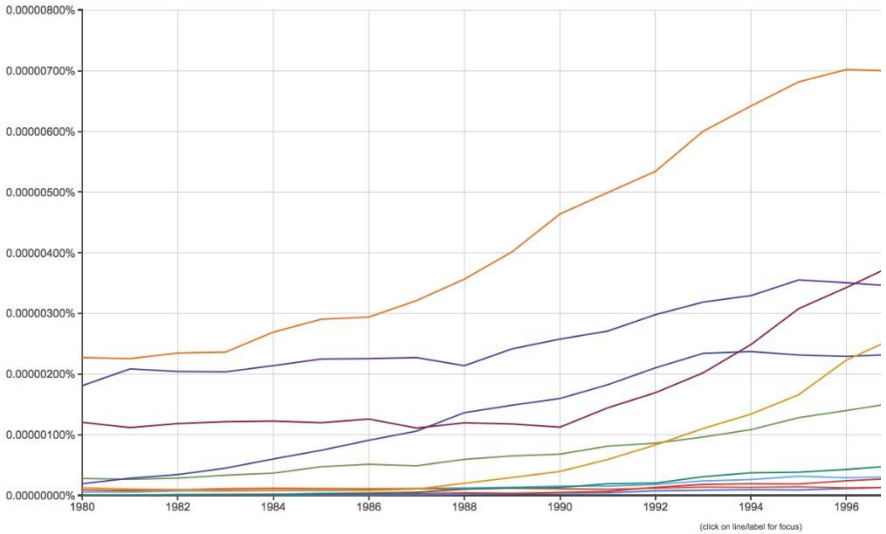
- 数字化环境对人文学科带来的影响才刚刚开始
  - 数字化 + 搜索
  - 人文计算：在人文领域应用数字工具和计算方法

1. 何谓“数字人文”——人文計算

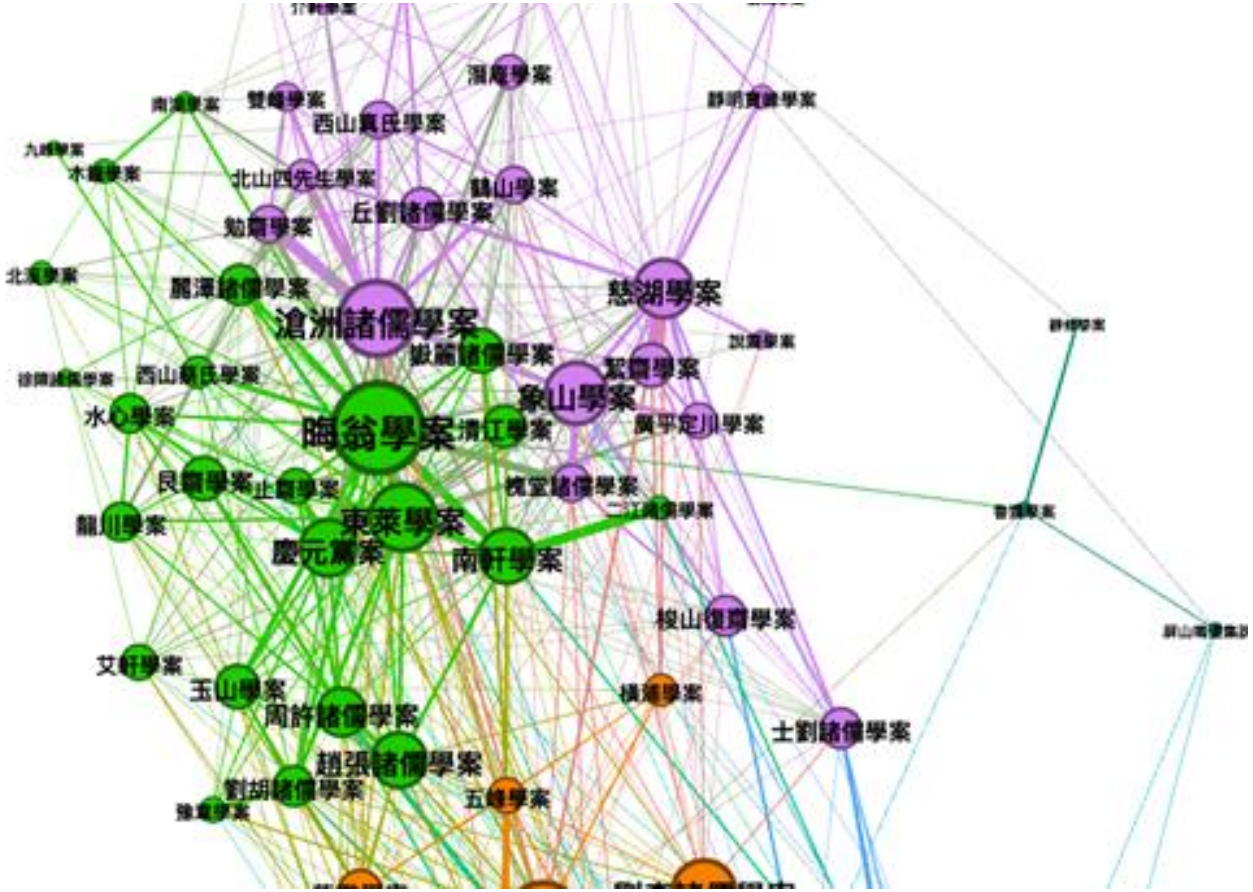


Big Names in Biblical Studies,
https://www.jasonstaples.com/bible/big-names-in-biblical-studies-on-precise-ngrams-and-search-engine-syntax/

1. 何谓“数字人文”——人文計算



Big Names in Biblical Studies,
https://www.jasonstaples.com/bible/big-names-on-precise-ngrams-and-search-engine-syntax/



杨海慈,王军. 宋元学案网络分析，2019，北京大学信息管理系

1. 何谓“数字人文”

- 信息环境的演变
  - 纸质文献环境 → 数字化网络化信息环境
  - 数字化网络化的学术生态环境

- 数字化对人文学科带来的影响才刚刚开始
  - 数字化+搜索
  - 人文计算：在人文领域应用数字工具和计算方法

  - **人文智能：深度学习与知识图谱在人文领域的应用**

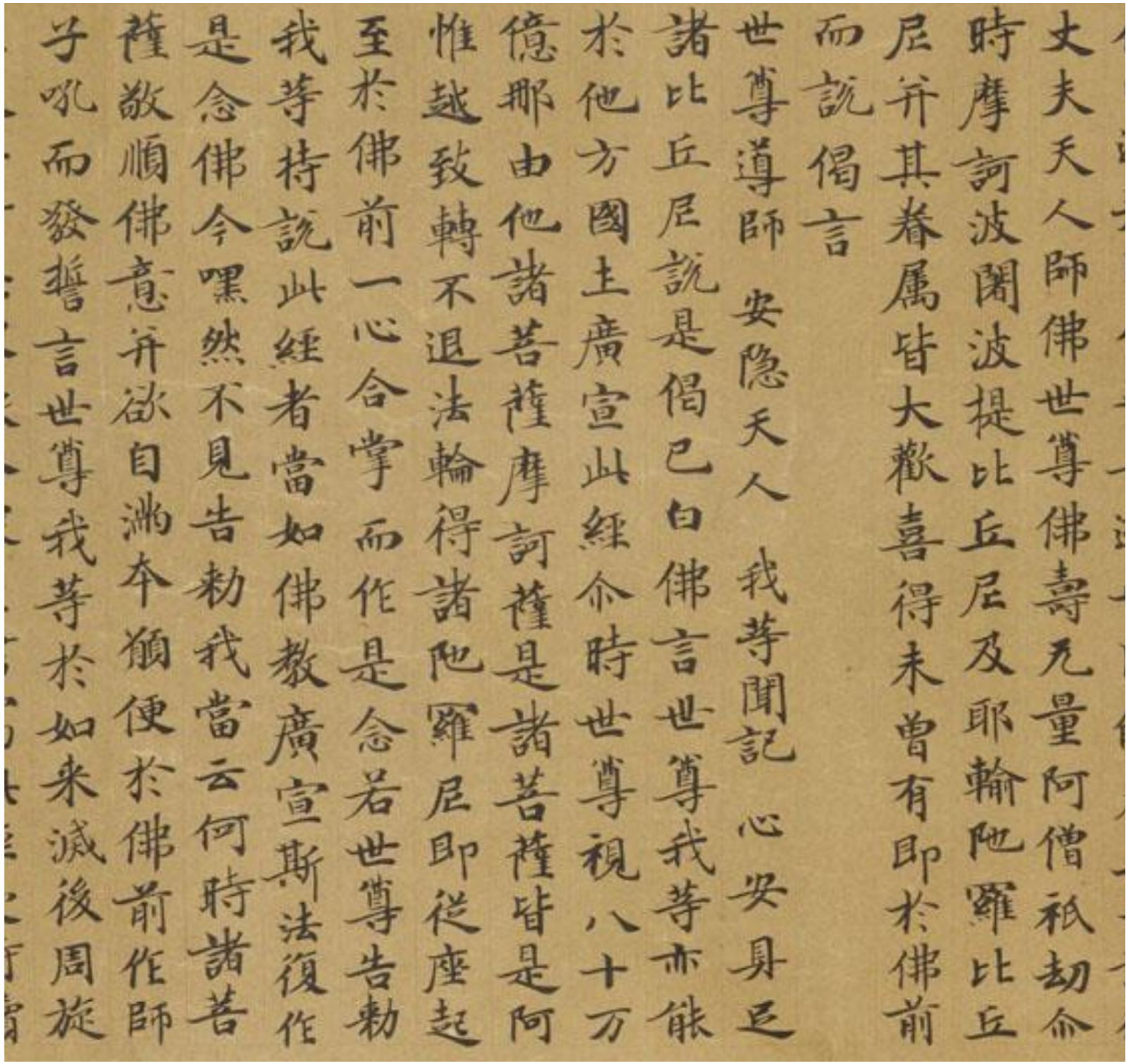
AI在古籍整理领域

- OCR 识别
- 自动句读
- 命名实体识别

OCR识别

以“书同文”为例

https://dzcj.unihan.com.cn/Ocr


丈夫天人師佛世尊佛壽无量阿僧祇劫今
時摩訶波闍波提比丘尼及耶輸陀羅比丘
尼并其眷屬皆大歡喜得未曾有即於佛前
而說偈言
世尊導師安隱天人我等聞記心安身定
諸比丘尼說是偈已白佛言世尊我等亦能
於他方國主廣宣此經今時世尊視八十万
億那由他諸菩薩摩訶薩是諸菩薩皆是阿
惟越致轉不退法輪得諸陀羅尼即從座起
至於佛前一心合掌而作是念若世尊告勅
我等持說此經者當如佛教廣宣斯法復作
是念佛今默然不見告勅我當云何時諸菩
薩敬順佛意并欲自滿本願便於佛前作師
子吼而發誓言世尊我等於如来滅後周旋

丈夫天人師佛世尊佛壽无量阿僧祇劫亦
丈夫天人師佛世尊佛壽无量阿僧祇劫尒
時摩訶波闍波提比丘尼及耶輸陀鄔比丘
時摩訶波闍波提比丘尼及耶輸陀鄔比丘
尼并其春屬皆大歡喜得未曾有即於佛前
尼并其春屬皆大歡喜得未曾有即於佛前
而說偈言
而說偈言
世尊導師安隱天人我等聞記心安具足
世尊導師安隱天人我等聞記心安具足
諸比丘尼說是偈已自佛言世尊柰等亦能
諸比丘尼說是偈已自佛言世尊柰等亦能
於他方國土廣宣山鈺介時世尊視八十万
於他方國土廣宣山鈺介時世尊視八十万
億鄀由他諸菩薩摩訶薩是諸菩薩皆是阿
億鄀由他諸菩薩夾訶荷是諸菩桂皆是阿
惟越致轉不退法輪得諸弛羅尼即從座起
惟越致轉不退去輪得諸弛羅尼即從座起
至於佛前一心合掌而作是念若世尊告勑
至於佛前一心合掌而仵是念若世尊告勑
我等持說此經者當如佛教廣宣斯法復作
我等持說此挺者當如佛教廣宣斯法復作
是念佛今哩然不見告勑我當云何時諸菩
是念佛今哩然不見告勑我當云何時諸菩
薩敬順佛意并欲自洮夲願便於佛前作師
薩敬順佛意并欲自洮夲願便於佛前作師
子吼而發誓言世尊我等於如来滅後周旋
子吼而發誓言世尊我等於如来滅後周旋

古汉语文本的自动句读

- 北京师范大学中心信息处理研究所 https://seg.shenshen.wiki/

- 梅花发寒梢挂着瑶台月瑶台月和羹心事履霜时节野桥流水声呜咽行人立马空愁绝空愁绝为谁凝伫为谁攀折 （朱熹《忆秦娥》）

- 李十一郎行修初娶江西廉史王仲舒女贞懿贤淑行修敬之如宾王女有幼妹尝挈以自随行修亦深所鞠爱 （冯梦龙《情史类略》）

- 此即昔人所谓东坡诗如大家妇女大踏步走出山谷便不免花面丫头屏角窥人扭捏作态之意 （柳亚子《磨剑室杂拉话》）

古詩文斷句 v2.1

點擊瞭解更多

有任何問題或需大規模使用請聯繫：shen@mail.bnu.edu.cn

古文

皇祖虽尝抉精防征辞防著论百余首亦惟析疑正陋垂教后世耳于其书则一仍厥旧无所笔削也故全书篇幅虽多而议论乃什倍于事实即如前编之中总论史论音释辨疑考证纷不一家正编之中凡例发明书法考异集览考证正误质实滥觞益甚至于续编之作成于有明诸臣其时周礼沿尹起莘例作发明而广义则出于张时泰效刘友益书法而为之者夫发明书法其于歷朝兴革正统偏安之际已不能得执中之论而况效而为之者哉且以本朝之臣而纪其开国之事自不能不右本朝而左胜国此亦理之常也况三编中嬗代崛起之际称太祖而系以我者不一而足亦非体例也故命儒臣纂歷代通鉴辑览一书尽去歷朝臣各私其君之习而归之正自隆古以至本朝四千五百五十九年事实编为一部全书于凡正统偏安天命人心系属存亡必公必

斷句結果

皇祖虽尝抉精防 O 征辞防 O 著论百余首 O 亦惟析疑正陋 O 垂教后世耳 O 于其书则一仍厥旧 O 无所笔削也 O 故全书篇幅虽多 O 而议论乃什倍于事实 O 即如前编之中 O 总论 O 史论 O 音释 O 辨疑 O 考证 O 纷不一家 O 正编之中 O 凡例 O 发明 O 书法 O 考异 O 集览 O 考证 O 正误 O 质实 O 滥觞益甚 O 至于续编之作 O 成于有明诸臣 O 其时周礼沿尹起莘例作发明 O 而广义则出于张时泰效刘友益书法而为之者 O 夫发明书法 O 其于歷朝兴革 O 正统偏安之际 O 已不能得执中之论 O 而况效而为之者哉 O 且以本朝之臣而纪其开国之事 O 自不能不右本朝而左胜国 O 此亦理之常也 O 况三编中嬗代崛起之际 O 称太祖而系以我者 O 不一而足 O 亦非体例也 O 故命儒臣纂歷代通鉴辑览一书 O 尽去歷朝臣各私其君之习 O 而归之正 O 自隆古以至本朝 O 四千五百五十九年事实 O 编为一部全书 O 于凡正统偏安 O 天命人心系属存亡 O 必公必

命名实体识别 （北大数字人文实验室）

例句：唐故朝議郎內供奉守慶州司馬上柱國賜紫金魚袋賈公故夫人穎川縣太君陳氏墓誌銘並序承務郎前太常寺協律郎李參撰公諱光,其先武威人也。

例句：考志悌皇長安縣尉、贈吏部郎中;(下渺)中府君之少子,蔡州吳房縣人,博陵崔公孟陽之外孫也。

命名实体识别 （北大数字人文实验室）

RF算法 | 深度学习 ▾ 人名 地名 职官名

句子标注的结果：

唐故朝議郎内供奉守慶州司馬上柱國賜紫金魚袋賈公故夫人潁川縣太君陳氏墓誌銘並序承務郎前太常寺協律郎李參撰公諱光,其先武威人也。考志悌皇長安縣尉、贈吏部郎中,下洎中府君之少子,蔡州吴房縣人,博陵崔公孟陽之外孫也。

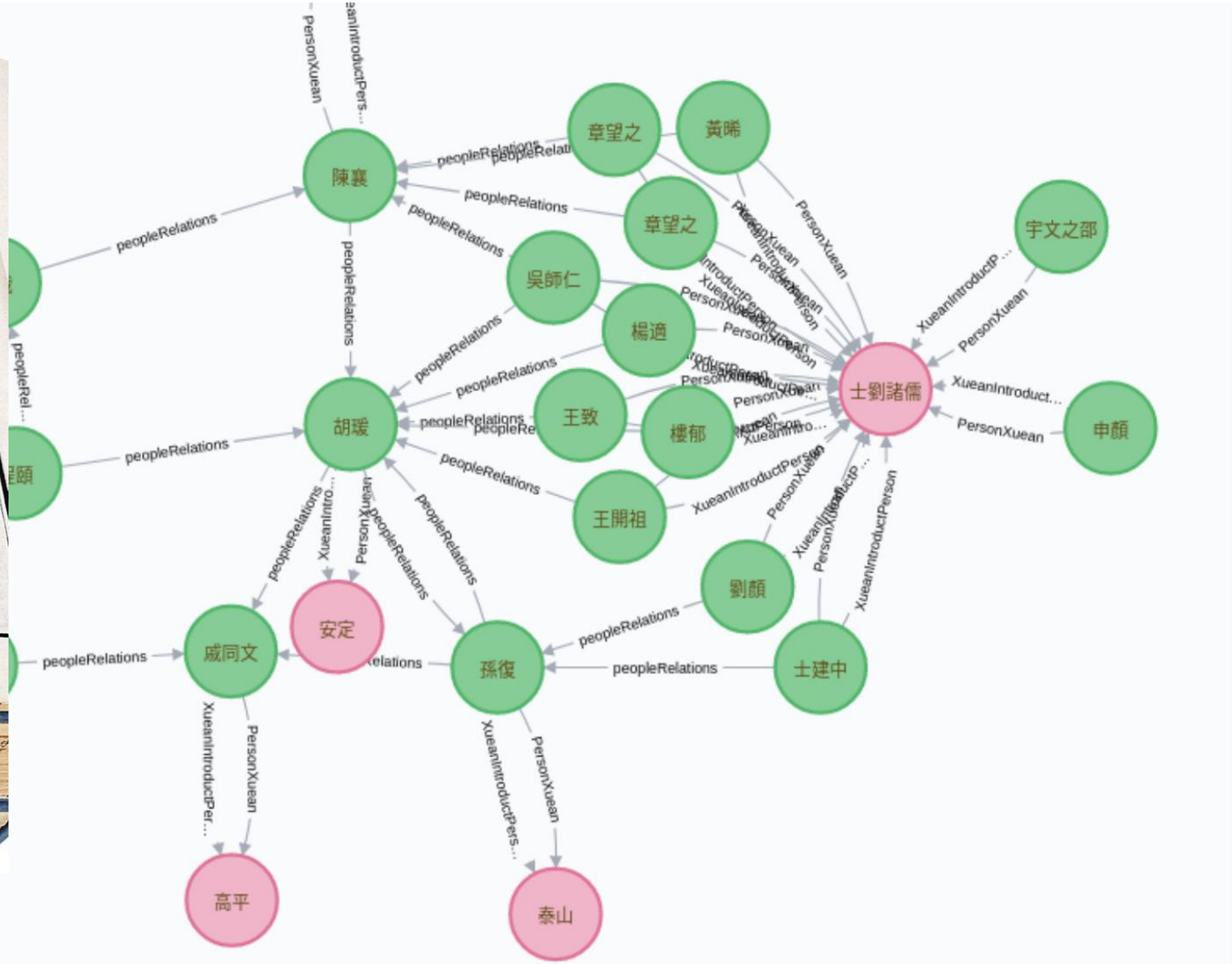
training_num	method
15078	CRF
15078	CRF
15078	CRF

ning_num	method
78	BiLSTM-CRF
78	BiLSTM-CRF
78	BiLSTM-CRF

	precision	recall	f1-score	support	training_num	method
micro avg	0.904045	0.889998	0.896966	21045	15078	CNN-BiLSTM-CRF
macro avg	0.898830	0.837511	0.864874	21045	15078	CNN-BiLSTM-CRF
weighted avg	0.903584	0.889998	0.896197	21045	15078	CNN-BiLSTM-CRF

知识图谱


宋元學案
第四帙
宋元學案卷二
慈谿馮雲濠
鄞縣王梓材校刊
餘姚黃宗羲原本
黃百家纂輯
道州何紹基
鄞縣全祖望修定
高平講友
泰山學案
祖望謹案泰山之與安定同學十年而所造各有不同安
定冬日之日也故如徐仲車宛有安定
風格而泰山高弟爲后守道以振頑懦則嚴嚴氣象倍有
力焉抑又可以見二家淵源之不紊也述泰山學案
卷與安定學案離山所修梨洲本原底並載盧氏又案
泰由著述莫重于春秋胥王發微故從黃氏補本錄之
橫渠
濂溪


陳襄
章望之
黃晞
章望之
吳師仁
楊適
胡瑗
王致
樓郁
王開祖
士劉諸儒
宇文之邵
申顏
劉顏
士建中
孫復
安定
威同文
高平
泰山
peopleRelations
PersonXuean
XueanIntroduct...

真正的“图书集成” ➔ 知识集成

- 工具书
  - 职官表、历史年代表、舆图
  - 字典、辞书
- 各类典籍
  - 史籍、地方志、年谱、政书、学案、etc.

对人文学科的影响

- 信息环境向知识环境的演进
  - 知识表示 & 知识集成
  - 知识生产的目标：不仅仅是为了让人读，而且要让机器懂

- 人文学科的人机结合模式：
  - “探赜索隐，钩深致远，以定天下之吉凶”

- 更多的可能性
  - 如何帮助人文学者解决研究问题?
  - 重大问题的突破性进展?

数字人文——

智能信息环境下人文学科的研究范式

为二十年之后的北大培养人文学者

数字人文研究中心(筹)

- 突破原有的学术藩篱与组织边界，融合文科、理科和技术学科，为多学科的交流与协作提供平台。

- **数字人文研究中心：**
  - 以人文学科为导向的跨学科平台，esp. 信息技术&管理学科+人文学科
  - 组织跨学科的团队，从事数字人文的研究
  - 培养数字人文人才，推动数字人文发展

- **数字人文开放实验室**
  - 人文需要实验室?

目标： 提供数字人文研究资源的导航（inc. 方法、工具、资源、案例、资讯）
成员： 刘姝然
王林旭，李晓煜，桑宇辰，陈雨航，王睿，季佳雯，唐震怡
法： 问卷调查、访谈、调研、网站开发

北京大学数字人文开放实验室

- **打造人文实验室**
  - 开放接收各院系师生加入
  - 提供平台，组织跨学科的活动
  - 建设数字人文虚拟社区

- **人文计算：** 推广数字人文的研究方法
- **人文设计：** 打造新媒体环境下的人文产品
- **人文创新：** 鼓励同学自主创新活动

工作坊计划

- **授课团队：** 台大 + 北大
- **授课内容：** Docusky, Markus, GIS, Text Ming
- **时间安排：** 四月中旬~五月中旬
- **后续通知：** 请关注 pkudh.org 网站通知，及实验室公众号

数字人文作品评选

- 5月：启动
  - 各院系同学启动自己的兴趣研究，DH-Lab协助组织跨学科团队
- 9月：提交作品
- 10月：作品评审、评奖、展示、出版
- 11月：“北京大学数字人文论坛”

北京大学数字人文开放实验室

公众号：pkudhlab
网　站：pkudh.org
公　邮：gdhc@pku.edu.cn