

用数字人文技术提高读者服务水平: 远读文本挖掘分析法

Improve Reader Service through Digital Humanities: Distant
Reading via Text Mining Tools

图书馆与社会发展: 新环境 新理念 新服务学术研讨会
中国上海浦东图书馆与
美国华人图书馆员协会21世纪图书馆员学术研讨会系列联合举办
2017年6月12-14日

The seminar on “Library and Social Development, New Setting, New Ideas and New Services”
Jointly Organized by Pudong Library, Shanghai, China, and the CALA 21st Century Librarian Seminar Series
June 12-14, 2017

报告人：包学鸣，教育博士(Xue-Ming Bao, Ed.D.)

电子资源馆员(Electronic Resources Librarian)

西东大学(Seton Hall University Libraries)

<http://blogs.shu.edu/baoxuemi>

提纲 (Outline)

- ▶ 数字人文与图书馆服务
(Digital Humanities and Library Services)
- ▶ “远读”的概念 (Distant Reading)
- ▶ 文本挖掘工具 (Text Mining Tools)
- ▶ 文本分析案例 (Examples of Text Analytics)
- ▶ 结语 (Conclusion)

数字人文与图书馆服务 (1)

Digital Humanities and Library Services

- ▶ 数字时代的人文学科
The humanities in a digital age
- ▶ 数字人文: 创造和保存可扩展的数字文字档案, 创造工具
用于和材料文化的互动
Digital Humanities: the creation and preservation of extensible digital archives to document, and tools to interact with, material culture
- ▶ 数字人文是人文研究和数字工具的重叠和结合
Digital humanities defines the overlap between humanities research and digital tools
- ▶ Source: Credo Reference <http://search.credoreference.com/>

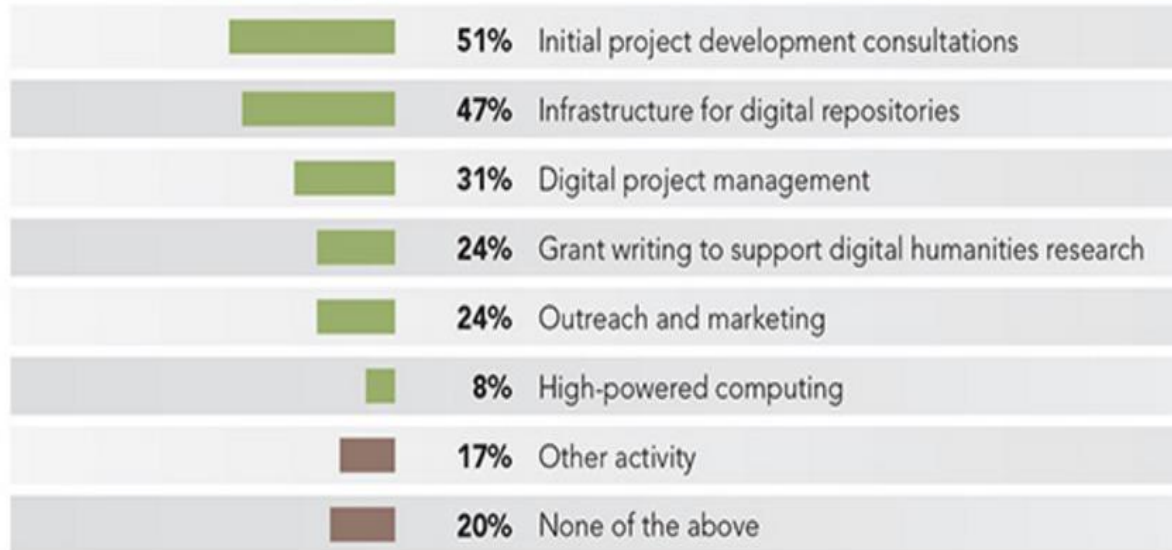
数字人文与图书馆服务 (2)

- ▶ 数字人文是图书馆新兴的资源和服务领域
- ▶ 图书馆创造数字馆藏，订购数库
- ▶ 图书馆如何利用数字人文的新技术提高读者服务水平？
- ▶ 我的演讲探讨如何利用文本挖掘工具，更好更快地了解向读者所推荐的新书

数字人文与图书馆服务 (3)

- ▶ 数字人文在图书馆是一个正在发展中的领域，充满了不定因素；图书馆需要适应。Digital Humanities in Libraries is a growing field with uncertainty and needs adaptation.

Which of the following types of services does your library offer users who are engaged in digital humanities projects?



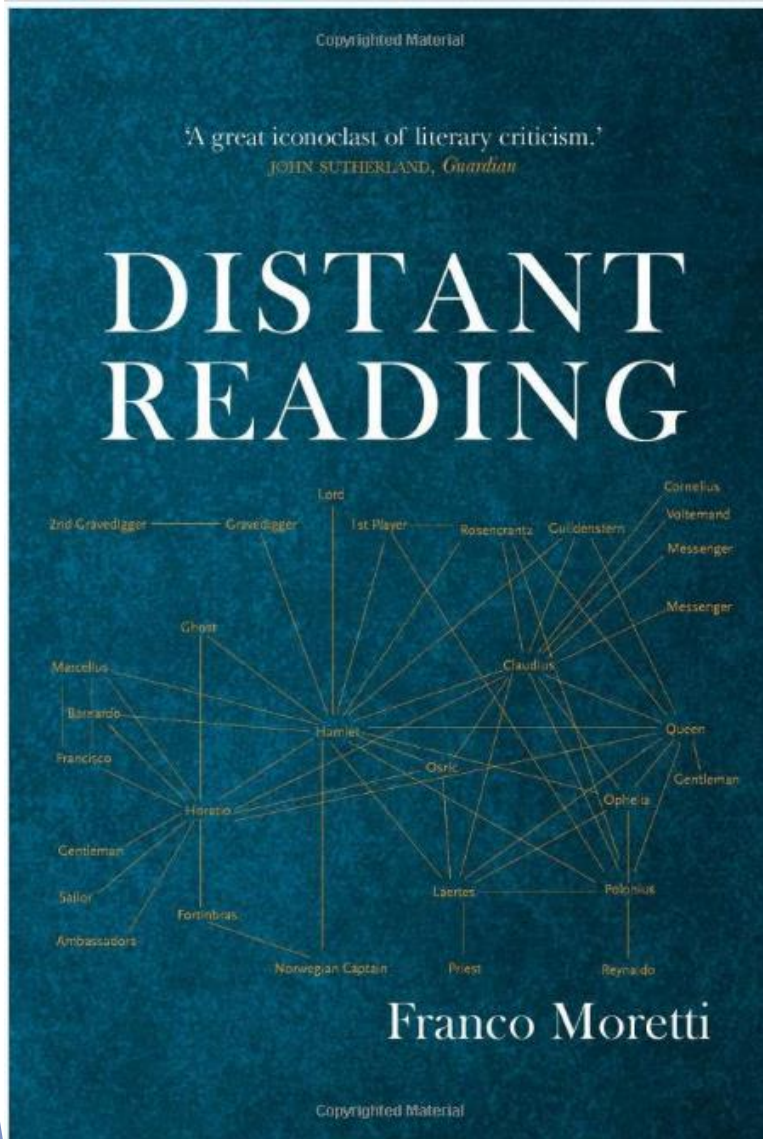
Which of the following types of services does your library offer users who are engaged in digital humanities projects? (click to enlarge)

Image 3 of 4

美国图书馆协会2016问卷调查：
图书馆为用户提供怎样的数字人文服务？
51% 项目初期的咨询
47% 数字知识仓库的基础设施
31% 数字项目的管理
24% 申请项目资金来支持数字人文研究
24% 外联与市场
8% 高性能计算机
17% 其它活动
20% 没有上述的活动

Source: <https://americanlibrariesmagazine.org/2016/01/04/special-report-digital-humanities-libraries/>

“远读”的概念 (Distant Reading)



“远读：距离是获取知识的条件之一：它能使我们集中到比文本更小或更大的单位上：如文学手法、主题、比喻、文类和系统。”

“Distant reading: where distance, let me repeat it, *is a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes - or genres and systems” (Moretti 2013, 48-49).

“远读”与图书馆服务

- ▶ 美国图书馆协会2016问卷调查报告中，专门有一段写弗朗科·莫瑞狄(Franco Moretti) 2001年提出的“远读”概念，即用电脑来研究大量的书籍。
- ▶ 报告指出：远读可以看到大量文本中的共同的特点，独特的模式，和进化的迹象。远读是相对近读而言（close reading）。近读指分析一个段落中，文字的细微差别。
- ▶ 报告还认为，图书馆可以成为远读这项工作的关键合作伙伴。图书馆员可以帮助用户进入数字馆藏促进远读。图书馆可以把训练使用文本分析软件工具作为教学内容之一。

文本挖掘工具 (Text Mining Tools)

- ▶ 文本挖掘工具是数字人文的工具之一，主要是用于非结构数据的分析
Text mining tools are a part of digital humanities tool collections, used for analyzing unstructured data like ebooks
- ▶ 常用的文本挖掘工具 Popular text mining tools
 - ▶ Voyant tools <https://voyant-tools.org/> (别名：看透你的文本)
 - ▶ NLPIR汉语分词系统 <http://ictclas.nlpir.org/>
 - ▶ LIWC <https://liwc.wpengine.com/> (语言查询与计算字数)
 - ▶ 北京版 <http://ccpl.psych.ac.cn/textmind/> (“文心(TextMind)”中文心理分析系统)
 - ▶ 台湾版 <http://cliwc.weebly.com/>
 - ▶ R <https://www.r-project.org/>





see through your text

Add Texts



Type in one or more URLs on separate lines or paste in a full text.

 Open

Upload



Reveal



UNIVERSITY LIBRARIES
SETON HALL UNIVERSITY

文本分析案例 (Examples of Text Analytics)

- ▶ 《三体》是刘慈欣创作的系列长篇科幻小说，由《三体》、《三体Ⅱ·黑暗森林》、《三体Ⅲ·死神永生》组成。
- ▶ 第一部于2006年5月起在《科幻世界》杂志上连载，第二部于2008年5月首次出版，第三部则于2010年11月出版。
- ▶ 作品讲述了地球人类文明和三体文明的信息交流、生死搏杀及两个文明在宇宙中的兴衰历程。其第一部经过刘宇昆翻译后获得了第73届雨果奖最佳长篇小说奖。
- ▶ 美国前总统奥巴马非常喜欢这部小说。



UNIVERSITY LIBRARIES
SETON HALL UNIVERSITY

Voyant 文本词频分析图案



- 这个词云（cirrus 或 word cloud）图案显示《三体》这部小说充满了对话，是一部关于宇宙、太空、地球和人类的小说。

Voyant文本语境分析

Documents

Phrases

Reader

Contexts

Bubblelines

在以后一段时光中读者朋友们将走过我在过去一年中走过精神历程坦率地说我不知道你们将在这条黑暗诡异迷途上看到什么我很不安但科幻写到今天能够与大家同行这么长一段也是缘份

1 疯狂年代

中国1967年

红色联合 对 四 二八兵团 总部大楼攻击已持续了两天他们旗帜在大楼周围躁动地飘扬着仿佛渴望干柴火种

红色联合 指挥中心急加梵他并不相柏士楼守卫者那一百多名 四 一八 战士与诞生于1966

Document	Left	Term	Right
1) voyan...	在过去一年中走过精神历程坦率地	说	我不知道你们将在这条黑暗诡异
1) voyan...	能使其惊醒过来机械地重复那已	说	过无数遍认罪词 然后他们中一部分人
1) voyan...	走上批判台时他那神情分明在	说	让我背负十字架更沉重一些吧 红卫兵们
1) voyan...	今天继续上次大会议程废话就不多	说	了老实回答下面问题 在六二至六
1) voyan...	怎么能不涉及它呢 叶哲泰回答	说	你胡说 旁边一名女红卫兵厉声说 ...
1) voyan...	回答说 你胡说 旁边一名女红卫兵厉声	说	爱因斯坦是反动学术权威他有奶便是
1) voyan...	早躲过了这场灾难 老人家摇摇头	说	我女儿不可能在学术上有什么建树

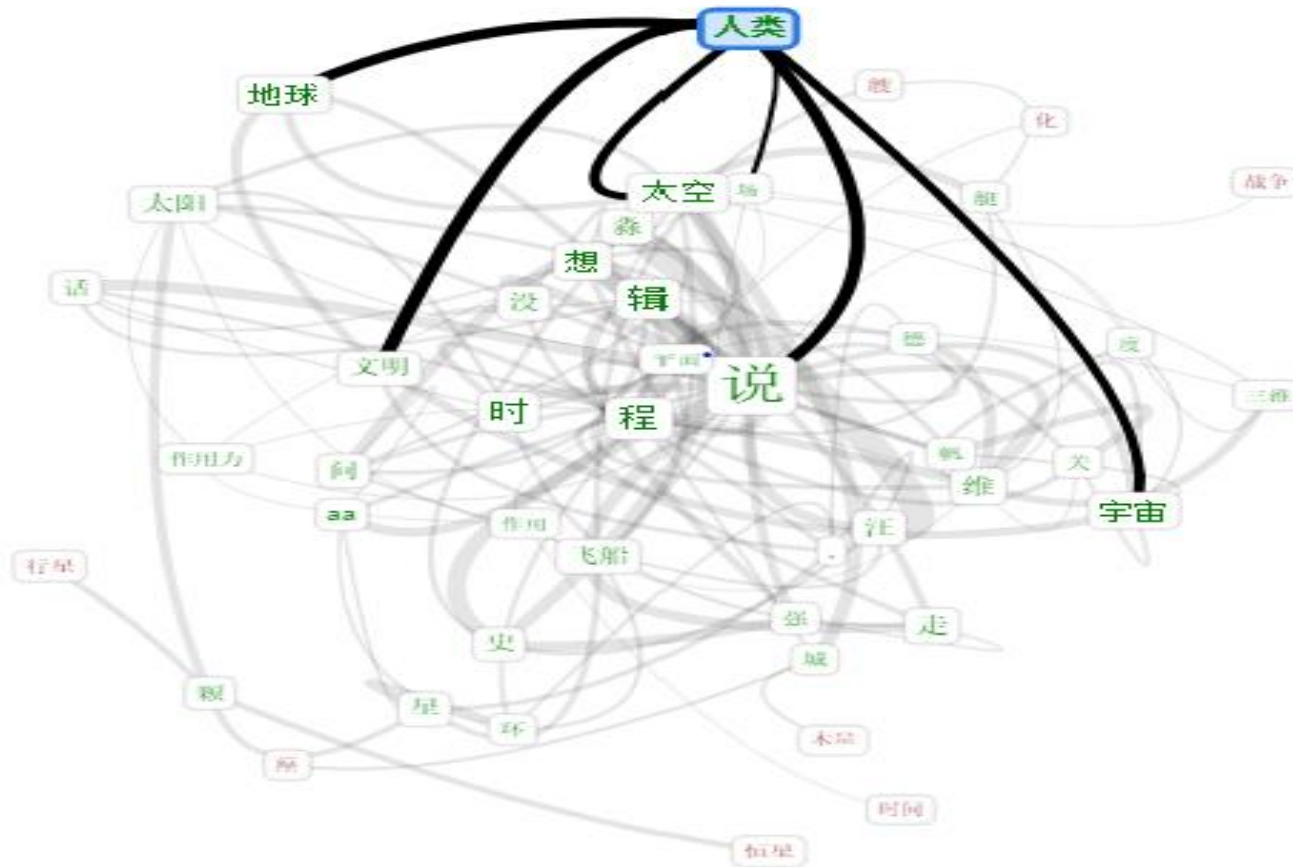
说 X

3,326 context

expand

- ▶ 语境分析可以显示《三体》小说里某个字的用法，如“说”字在小说中怎样出现的。

Voyant 文本关键词连接分析

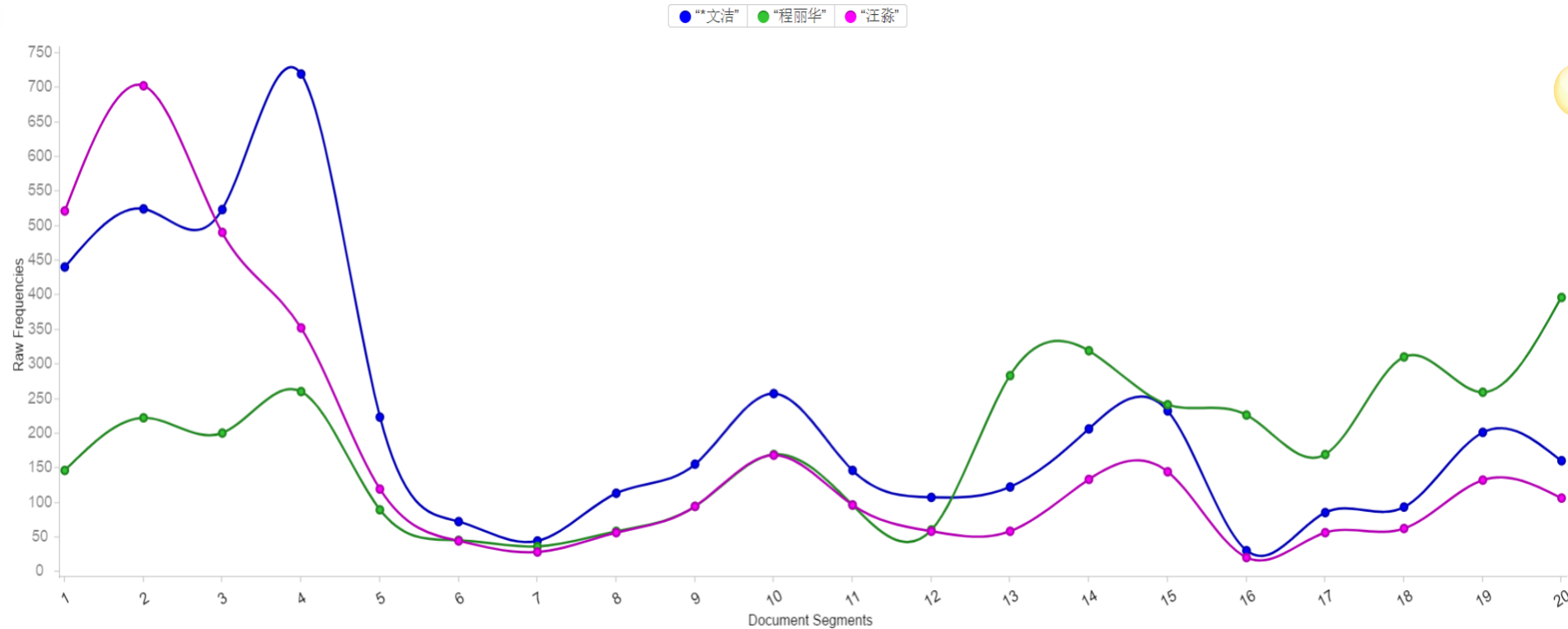


- ▶ 关键词连接分析可以显示在《三体》小说中，人类和宇宙、太空、地球、文明有密切联系。而且人类是这些连接的中心。

Voyant文本人物分析

Trends

?



- 人物趋势图可以显示在《三体》小说中，人物在叙述的顺序中被提到的频率，由此可以看出人物在小说中的份量。汪淼和叶文洁是小说的主人公。程丽华虽然不是主人公，但是小说的后半部分被提到份量很多。

结语 (Conclusion)

- ▶ Voyant 文本挖掘工具使用简单，无需编程知识，可提供20多种分析方法。
- ▶ 图书馆员应利用数字馆藏和数字工具，为用户提供创新服务，提高读者服务水平，更有效地推荐新书。
- ▶ 图书馆员应积极参与，成为数字人文远读工具的使用者和培训者。