为何人文科学学者应该爱上大数据？How？

时间 2016-04-26 17:25:50  [北大社会化媒体研究中心](https://www.tuicool.com/sites/E7f63yF)

原文  [http://www.csmrpku.com/archives/5401](http://www.csmrpku.com/archives/5401?utm_source=tuicool&utm_medium=referral)

主题 [大数据](https://www.tuicool.com/topics/11040019)[数据库](https://www.tuicool.com/topics/11000064)[GIS](https://www.tuicool.com/topics/10200136)

作者 | Song Chen 陈松

翻译 | 张伊妍

编辑 | AKI

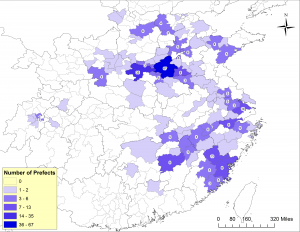


地图总是能展现出历史的延续性。这是一张将2011年中国高铁线路（红线）和明代驿站位置（蓝点）交叠起来形成的地图。显而易见的是，两者之间有着耐人寻味的不同，但这两个相隔几个世纪的交通网络所呈现出的相似性更加令人震惊。

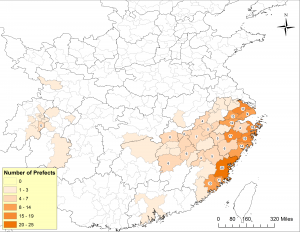


地图 1 红线代表2011年中国高速铁路网络，蓝点代表明代驿站的位置。| 图片来源: Lex Berman, ed., China’s History in Maps（从地图中看中国历史）, 由WorldMap提供。

地图还可以展现出历史的变迁性。看看下面这些地图吧，它们呈现了11世纪40年代（地图2a）和13世纪10年代（地图2b）地方政府的分布。不同的空间特征立刻凸显出来。11世纪40年代的地方长官主要被封在宋都城附近地区（开封和洛阳），而13世纪10年代的地方长官则主要分布在东南沿海地区。对于笔者来说，这样的对比显示了两个时期非常迥异的对政治成功的理解。在11世纪40年代，也就是晚唐时期，都城垄断了所有阶级向上流动所需要的文化和政治资源；与之相反的是，13世纪10年代，经济上的富有才是真正有用的东西。

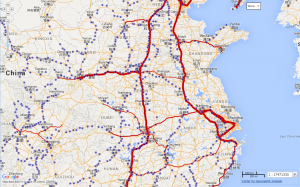


地图2a. 11世纪40年代的地方政府都分布在哪些地方？数据来自中国历代人物传记资料库(CBDB)、中国历史地理信息系统(CHGIS)和其他附加的研究，地图由ArcGIS绘制。CBDB 和CHGIS 都是Peter K. Bol.领导的跨国多领域数字化项目。

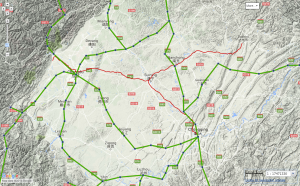


地图2b. 13世纪10年代的地方政府都分布在哪些地方？数据来自中国历代人物传记资料库(CBDB)、中国历史地理信息系统(CHGIS)和其他附加的研究，地图由ArcGIS绘制。（对地图2a和地图2b的详细分析，参见陈松的“Governing a Multicentered Empire: Prefects and Their Networks in the1040s and 1210s,” in State Power in China, 900-1325, edited by Patricia Buckley Ebreyand Paul Jakov Smith. Seattle, WA: University of Washington Press, forthcoming.）

两张地图都引导我们不要“只见树木（单个人或地点），不见森林（宏观模式）”。但“森林”本身就足够有趣。当我们去看宏观模式时，也能够从全新的角度看待每个个体。比如，下面这张地图标记出了郑州、武汉和长沙这几个地点，它们在两个时期都是非常重要的转运枢纽。这张地图也显示出了21世纪四川盆地陆路运输的新模式。这足以让我们好奇许多问题，比如，为什么有些地方在今天的高速铁路线路上，却不在明代驿站的网络中？（比如绥宁/遂宁）这代表了什么？

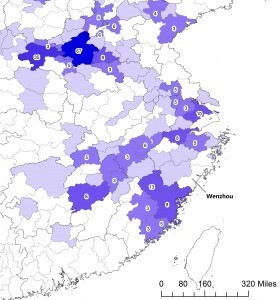


地图3a. 地图1a里中国中部地区的放大图，红线代表2011年中国高速铁路网络，蓝点代表明代驿站的位置。

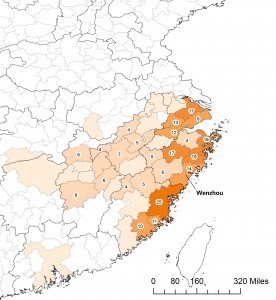


地图3b. 地图1a里四川盆地的放大图，红线代表2011年中国高速铁路网络，蓝点代表明代驿站的位置，绿线代表明代驿道的路线。

我们也可以用同样的方式来看宋代的地方官员的设置。11世纪40年代，没有一个温州人有地方官员的头衔；到了13世纪10年代，有14个温州人获得了官衔。这可是很大的进步，不是吗？但是这是为什么呢？有人可以会将其归因于温州吉利的地形，以及自南宋时期开始就居住在此的书香门第所带来的美德熏陶。解释这些原因，都可以讲出很好的故事来，而这些故事，我们可以轻而易举地在地方志或地方家谱中读到。但是地图展现出来的是，温州的成功并不像看起来那么独特。这其实和地方的优越性没有多大关联，因为有些事情要从更宏观的时间和空间的视角来看待。



地图4a. 地图2a里东南沿海地区的放大图。“11世纪40年代的地方政府都分布在哪些地方？”



地图4b. 地图2b里东南沿海地区的放大图。“13世纪10年代的地方政府都分布在哪些地方？”

宏观视角，模式。这些正是地图所擅长呈现的东西。这也让大数据整体上变得充满魅力。它让我们把注意力放到历史架构的变迁和传承上。我们牺牲了微小的细节，去理解宏观的模式；然后在这种全新的模式下，去重新审视个体的人、地点和事件。

**大数据**

说到世界文学，Franco Moretti(“Conjectures on World Literature”,New Left Review 1(2000): 55)曾极富洞察力地表示它（世界文学）“不是一样东西，而是一个问题，一个亟需新的鉴赏方式的问题”，我们不能“仅仅靠阅读更多文本”。大数据也是这样。近年来，大量文学的、历史的和宗教的文献被电子化，并转化文全文可检索的数据库。比如“电子四库全书项目”有3460个大标题，总计大约8亿个文字。“新汉籍全文资料库”和“国学宝典”也同样如此，均包括几亿文字量，而且都在不断扩充中。出版商爱如生也已经开发了数个关于宗族谱、题字和其他特殊文本的大规模的全文数据库。

但大数据不能仅仅被定义为“大量的数据”，而是应该被看做是大量数据所带来的“问题”。从数据量的角度来看，我们人文科学学者现在所拥有的“大数据”和许多自然科学与社会科学学者所致力于研究的截然不同。举个例子，每年有大约2亿条贴文发表在推特（Twitter）上。2011年，美国国家安全局（NSA）每天收集了数以百万计的文字讯息，并且想办法将千兆位元组级别的数据进行了分析和可视化。（一个千兆位元组=100万个十亿位员组）。

大数据之所以伟大，是因为它提供了新的研究机会，但如何充分利用这些机会带来了新的方法论上的挑战。机会和挑战并存。每个学科领域的成就伟大的门槛都是迥然不同的，但所有的领域中，成功的脚本又非常相似。电脑技术的进步已经极大程度上拓宽了可使用的数据的量级，不管是结构化的还是非结构化的数据，定性的还是定量的数据，这种拓宽让我们可以提出新的问题。但为了回答这些问题，某个学科领域中传统的分析方法已经远远不够。Lev Manovich在Manovich等人撰写的《如何比较100万张图片》(Manovich, et al, “How to Compare One Million Images?” in UnderstandingDigital Humanities. Edited by David M. Berry. New York: Palgrave, 2011)中富于表现力地谈及了这种挑战。当100万张漫画页面中的数字图像只是在几次点击中从我们指尖溜过，我们该如何研究它们？我们的生命如此短暂，我们也没有百眼巨人阿耳戈斯那么多的眼睛，或是记忆女神摩涅莫辛涅那样的好记性。即使我们在翻看这些漫画的时候非常享受其中，我们也很难定格住这些页面之间的细微差别。前所未有的巨大数据量也在呼唤着全所未有的分析处理方案。而同样的，这样的解决方案也将来自于电脑计算方法的世界。

**“词语的口袋”法**

上文中提到的解决方案可以分成两个大的分支。比较一下下面两张图片：



图1. 使用MARKUS制作的对《吴船录》的自动文本标记，仅显示前五篇文章。《吴船录》的文本抽取自Donald Sturgeon的“中国哲学书电子化计划”网站。



图2. 使用 Guan-tin Chien的HTML5词云图生成器（Word Cloud Generator）生成出来的《吴船录》的词云图。为了简明，词频小于4的词语已经被自动剔除。

这两张图是对同一篇文本——范成大的《吴船录》所进行的可视化图像。构造这样的图像，说实话并不难。只要你弄到一份《吴船录》的电子文档（笔者在上文中用到的就来自“中国哲学书电子化计划”网站），你所需要的只是一个现代的网页浏览器、互联网连接和点击几下鼠标而已。哦对了，还有几秒钟的耐心。第一张图片来自MARKUS，第二张来自词云图生成器（word cloud generator）。这两张图片之间的区别显而易见。它们都将其中的一些词语重点标记了出来——要么是用不一样的颜色，要么是用更大的文字尺寸。

但MARKUS和词云用了完全不同的方式来决定哪些词语是需要重点标记的。词云只是进行了简单的计数，计算每个词语在文章中出现的次数。因此，“十里”这个词在词云中显得非常大，因为它是文章中出现次数最多的词语，一共有67次（包括那些两个字拆开出现的情况，比如“二十里”、“三十里”等等）。MARKUS与之相反，对文本使用了实体关系模型，并采用电脑算法去“挖掘”这些“实体”。一个“实体”可以指任何东西，不论是具体的还是抽象的。它可以是一个日期、一个地点、一个官方称号、一个人名、一本书、一个机构或者其他东西的完整名称。MARKUS更关注那些表示一个实体的词语，而忽略非实体词语。也就是说，MARKUS将一个文本看做是标记出了实体及其关系的数据的储藏室——尽管这些数据是以叙述性文字的方式呈现的。

MARKUS和词云对词语在文本中的位置采取了完全不一样的态度。MARKUS将关注点放在词序上，其关注程度至少到获得的词序可以提供线索，让我们了解实体是如何被辨认出来的，它们之间又是如何相互关联的（比如“X推荐了Y”就和“Y推荐了X”表达的意思不同）。而对于词云来说，词序是完全没有关系的。它所采取的方式有一个很恰如其分的形容，就是“装满词语的口袋”。这种方式完全忽略了句法规则，而是将每篇文本看作是词语的简单集合。词云是这种方式的最简化版本的运用。它所做的仅仅是将文本拆解成词语，计算每个词语出现的次数，并按照计数给它们排序。作为一个探索性的工具，它是很有用的（比如，《吴船录》的词云图就告诉我们，这本书提供了大量关于四川的山川与河流，以及地理距离的信息。）然而，词云也建立在一个值得怀疑的假设上。它假设词语和其含义有着直接线性的关联，认为每个词语只有一个含义，这个含义可以脱离文本独立理解，而且出现频率最高的词语能够揭示出文本主题和/或风格上的特点。

不是所有程序都是“词语的口袋”的方式来进行假设。另一个例子是“主题建模”。它采用了一个或然性的模型，从词语的排列搭配来推断文章主题。尽管“主题建模”也忽略了词序，它至少认识到了同一个词语在不同语境下可能有不同的含义，并且尝试去“通过一个词语周围的词语了解它的意思”（也就是说，从上下文中获知这个词语的含义）。由于东亚语言的一个词语常常由两个或两个以上的文字组成，且词与词中间没有空格隔开，一个经常发生的情况就是，使用“词语的口袋”方式的学者要么不得不将一个词语的每个字看做独立的意义单位（比如“陛下”会被看作是两个单独的字，即使它们在文本中紧贴在一起），要么就得找到一个办法把这些字组词更具有实际含义的分析单位（比如N元语法模型或其他斯坦福大学和中央研究院开发的复杂模型）。即使有这样那样的挑战，这种方式还是在识别大部头文本全集散乱的、独特的特点时，生产出了很多令人兴奋的成果。比如在Paul Vierthaler (Digital Approaches to Late Imperial ChineseLiterature research talk delivered at the Fairbank Center for Chinese Studies,Harvard University, 2014 on Youtube:https://www.youtube.com/watch?v=adRlz2HbDD0)的文体学中，就展示出即使是对单个文字的分析也能帮助我们将野史作为一个文学流派来理解。在Donald Sturgeon的成果中，他这种模型与网络分析技术相结合，在大部头文集具有相似性的基础上，去探索中它们的文本之间的复杂关系。(如果你想看看研究中国以外的作品，可以参见 Matthew Jockers 和Matt Erlin的研究)

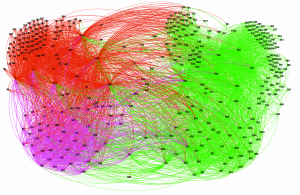


图3.展示了文字在韩国佛教教规中反复使用情况的网络分析图，由Donald Sturgeon制作。佛教教规根据 Lewis R. Lancaster’s 的分类进行标记(http://www.acmuller.net/descriptive\_catalogue/index.html) ，并根据使用频率高低分成了不同的类型（以不同颜色标出）。再现已经过原作者同意。

整体而言，“词语的口袋”法和实体关系法反映出了针对文本的不同假设。实体关系法假设现实的底层（或者至少是文本宣称的“现实”）是通过文本的“调解”被我们所感知的。这种方法希望通过从文本中抽象出不同的实体和它们之间的关系来描述“现实”。文本就意味着一个结束。用庄子的话来说，文本是用来打鱼（鱼指的是实体和它们的关系）的捕鱼栅。所以，一旦我们获得了鱼，捕鱼栅就不再有用了。而截然相反的是“词语的口袋”法，它倾向于将文本本身看成是值得探讨的有意义的主题，希望理解文本的主题和风格特点。当然，两种方法之间的界限不是永远那么明确的。文本主题之间的差别本身就不仅仅反应了作者的个人选择，也反应了社会现实。举例来讲，Ian Miller对《清实录》的研究(“Rebellion, crime and violence in Qing China, 1722-1911: A topicmodeling approach,” Poetics 41 (December 2013): 626-649)就用了主题建模，不仅反映了清朝官僚机构不同的凌乱分类，也认为这些分类反应了现实世界不同类型的暴力行为，每种暴力行为都有自己的时态模式。

**实体关系模型**

让我们回到MARKUS。记住，《吴船录》中的所谓“重点词语”（即那些被认为是代表实体的词语）是被MARKUS标记出来的。所以那又怎么样呢？我们需要按顺序进行一些编辑矫正（我们待会会讨论如何尽可能减少所需的编辑工作）。但一旦完成编辑矫正，我们就可以很容易地将这些标记导入到一个表格里——只需要按页面上的几个按钮而已。然后我们就可以打开Microsoft Excel软件中的一个表格进行筛选，比如只显示文本中提到的人名。这样我们就会得到范成大1177年从成都旅行到杭州途中有过联系的人，以及范成大眼中在12世纪末塑造了四川文化图景的那些人。下图所显示出的是这些人在中国传记数据库(CBDB)中可能对应的ID，你可以在这个数据库中找到对应个人的详细生平信息。

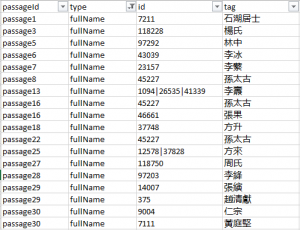


图4. 对《吴船录》中所有人名进行的自动标记（已经过编辑矫正）

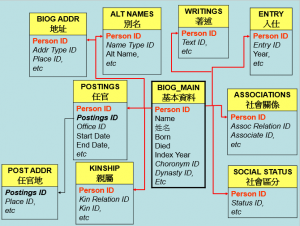
或者我们也可以看看MARKUS标记出的所有地名，并将它们与对应的地理变量匹配起来。对于中国的历史地点，我们可以找到许多数据来源，包括CHGIS 和Ruth Mostern 与Elijah Meeks 开发的“宋代数字地名”（the Digital Gazetteer of the Song Dynasty）。同时，我们也不难利用CartoDB或其他工具找到范成大旅行的路线。这能给我们提供一个对于范成大1177年游历的直观理解。或许也是一个优秀的教学工具。



图5. 自动标记出的《吴船录》中的地名，从MARKUS中导出，并经过编辑勘误。

但实体关系法的真正力量只有在我们将它应用于大部头、高精度（这样可以将编辑工作压缩到最少）的文本的时候才会展现出来。比如对于宏观分析，如果我们将所有12世纪典籍中所有零散的关于游历路线的引文全都标记出来，我们就可以重现当时交通网络的情况。像这样的尝试往往需要更加专门设计的算法（因此需要一些编程知识）、一个相当大的编辑团队（因为在一个大部头的文本中，即使是很小比例的错误也实际上意味着非常繁重的工作量）和/且愿意接受数据中仍存在适当程度上的错误（换言之，必须拥有足够的自信认为少数的错误不会影响我们从上千条数据中得出的结论）。

在这方面最合适的一个案例是中国传记数据库（CBDB）的团队在处理中国历史上的传记数据时所做的一切（也是中国传记数据库提供了上文中地图2a和2b的数据）。通过中国专家和计算机科学家的通力协作，这个团队正在开发适应性的专门设计的算法来系统地标记大量历史文献中所有人物生平相关数据（包括官职、亲属关系和社会关系等等），并提取出这些数据，将它们存储在一个关系数据库中（见表格6）。这个数据库的架构同样建立在实体关系模型上。时间、地点、机构、人物和书籍都是实体。关于一个人生活的每个方面通过实体之间关系的形式被重新呈现出来。因此，一个官职任命就可以被看做一个人、一个机构、一个地点（这个人被派去任职的地点）及（任命）开始和结束的时间之间关系的实例。类似的，婚姻也可以被看做丈夫和妻子在一段特定时间内发生的关系。



表格6. 一个高度简化的CBDB数据结构模型

因此，这个数据库大大便利了针对大量历史人物开展、交叉检验他们生活不同方面，并将他们的生平信息整合起来的研究项目（CBDB将在2016年5月前发布关于大约370,000个历史人物的生平信息），这些研究项目所需的数据因此散乱在不同的历史文献中。比如，我们可以从数据库中查找出16世纪所有的进士以及他们之间的亲属关系。这样的搜索可以把原本需要成百上千次点击的工作压缩到几分钟。而且我们可以将从各个来源收集的数据结合起来，比如科举名录、地方地名索引、断代史和碑文。CBDB的一个尤其重要的强项是可以从现有的记载中推断出没有记载的亲属关系（比如，如果从数据库中查出B是A的儿子，而且是C的父亲，那么CBDB会报告出A是C的祖父）。这样，我们就可以将这些结果导出为合适的格式，用来进行大规模的统计分析、空间分析和网络分析。

对于每个人来说都有价值的东西……

简单来说，实体关系模型基于一个假设，并整合了三个数据方法的力量。它假设文本是实体及它们之间关系数据的储藏室。它利用了数据挖掘技术来识别和提取实体和它们之间的关系（技术上被称作“命名实体识别”或者NER）。它将这些数据整理成一个关系数据库来帮助实现各种复杂的研究需求。数据要么被直接从文本中提取出来，要么通过查找从数据库中找出来，然后被可视化并进行分析，用来理解宏观模型。

今天，这三种方法都提供了很大的选择范围，以满足不同的用户需求并适应不同程度的计算机使用水平。即使是一个技术水平与经验都很有限的用户都可以像我们在前文中对地图1做的那样，点进WorldMap分享的一幅地图，打开或关闭图层，并根据自己选择的话题探索空间模型。或者我们也可以利用MARKUS自动的标记功能，结合CBDB现有的搜索功能和“傻瓜型”可视化工具（比如CartoDB 和Palladio）来回答一些简单的科研问题。只需要在CBDB的桌面软件中点击几下，一个人就能找到1000年到1040年所有已知的科举中举者。如果再经过一点儿调整和CartoDB中的再几次点击，我们就能获得出一副展示了中举者地理空间分布的地图。

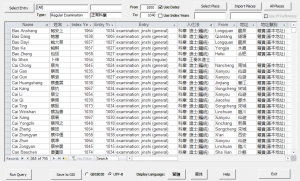


图7. 用CBDB的桌面软件现有的查询功能搜索1000年到1040年所有已知的科举中举者的查询结果

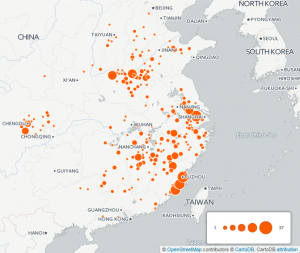


图8. 用CartoDB制作出的地图，展示了上面查找出的中举者都集中来自哪些地方。

用同样的方式，我们也可以利用CBDB自带的查询功能找到在1030年到1070年之间所有发家的人之间的学术关联，然后将这个结果复制粘贴到Palladio里来简单地看看这些联系以及它们的地理分布情况。



图9. 用Palladio制作的网络分布图，展示了用CBDB自带的查询功能找到在1030年到1070年之间所有发家的人之间的学术关联。节点的大小与此节点连接其他节点的数量成正比。

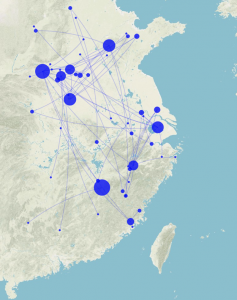


图10. 用Palladio制作的展示了用CBDB自带的查询功能找到在1030年到1070年之间所有发家的人之间的学术关联的地理分布图。节点的大小与此节点连接其他节点的数量成正比。

这些简单的可视化图表只能直观地显示主要的知识分子群体和它们之间的关系。这说明“傻瓜型”工具只有非常有限的分析能力，也只给使用者非常有限的对呈现形式的调整空间。

只要再加上多一点点对数据分析方法的理解，就能够让我们前进很大一步。比如，一些关于正则表达式语法的知识（正则表达式语法，简单来说就是自然语言中计算再现的公式化表达），将在很大程度上加强我们按照特定的研究需求标记文本的能力。举一个例子，在自动标记的基础上，我们可以在MARKUS的“关键词标记”模块提交简单的正则表达式（regexes）来在几秒内快速筛除不需要的地名标签，这样就能大大减少编辑勘误的工作量。



图11. 在MARKUS中提交的正则表达式，能让MARKUS仅标记出1177年范成大游历路线上的地名。



图12. MARKUS中使用了上文中的正则表达式后的《吴船录》关键词标记，仅显示文章的前五页。

同样的，多一点对关系数据库和可视化工具的了解就可以让我们在像CBDB这样的关系数据库中创造出起定制化的搜索请求，根据其他研究调整搜索结果，并用更高级的程序，比如ArcGIS （用来分析地理空间模式））, UCINET和Gephi （用来分析和可视化网络中的结构模式），来对这些结果进行分析。过去的学者已经用这些可视化效果来研究知识分子的活动模式（见Peter Bol对道学运动的研究“GIS，人物传记和历史（GIS, Prosopography, and History）”）和政治成就，但他们只提到了几个案例，这些只关注个体历史人物和个案的研究显得多少有些不够明朗。

由此开始，更多雄心勃勃的使用者或许会开始给自己规划，学习至少一种编程语言并训练自己写出适应研究需求的数据挖掘算法，或开发出面向公众的交互性程序。D3.js, R, 和Python是目前最流行的语言，而且相关的指导对于人文科学学者来说也越来越方便获得。(比如 D3.js in Action,Humanities Data in R, 和theProgramming Historian).

对于希望将数据方法介绍给学生的教师来说，这也意味着我们可以将大量课程加入到课程计划中来满足不同学生的需要。有些学生希望学习数据方法并将它们运用到研究计划中。其他学生可能只是想“了解”这些方法。尽管他们可能没打算成为这个领域的从业者，但是他们同样好奇数据方法对于人文科学领域来说意味着什么，甚至更好奇我们该如何评价它。在过去的几年中，笔者在巴科内尔大学和哈佛大学教授了几门课程。其中一门叫做“人文科学可视化”，就是设计给那些感兴趣的学生的。这门课比起深度更看重广度，它的目标是向本科生介绍人文科学领域大量的可视化项目，并在对数据方法的当代争议的语境下，建立起学生对其的批判性理解。在课堂上，学生需要进行实操，但只是使用简单的、易入门的在线工具。这门课的意义是不是让他们精通这些工具，而是消除数字人文科学的神秘光环。

另一门叫做“数字时代的中国历史”的课程则非常不一样。这门课是开给东亚研究的未来学者的，它更加围绕项目展开，而且有更多技术元素，会带领学生完成整个从数据标记、数据库设计、检索请求设计到发展出一目了然的论点的过程。不难想象，这样类似的课程还可以有很多，比如重点针对不同的成套技术工具的课程（如文本分析）、根据特定研究话题组织和单个项目开展的课程，或是教授某个人文科学研究需求所需要的核心编程技术的课程。

这样的课程正在全国范围内不断增多，无论是在文理学院还是在研究型大学都是如此。我永远不会说“大数据”或“数字人文”是人文学科的未来。它们或许不会成为人文学科的顶梁柱，但毫无疑问，它们会是人文学科的支柱之一。是时候拥抱数字素养，将它纳入我们的人文学科训练当中了。

**作者：**

Song Chen陈松

Assistant Professor of Chinese History中国历史助理教授

Department of East Asian Studies东亚研究系

Bucknell University巴科内尔大学

songchen@post.harvard.edu

**本文所提到的项目和工具的链接：**

* China Biographical Database (CBDB), 由以Peter K. Bol牵头的跨国团队开发:

http://projects.iq.harvard.edu/cbdb/home

* China Historical GIS (CHGIS), 由以Peter K.Bol牵头的跨国团队开发:

http://www.fas.harvard.edu/~chgis/data/chgis/downloads/v4/

* China’s History in Maps (由WorldMap提供), 由哈佛大学的LexBerman 针对地理分析进行修改:

https://worldmap.harvard.edu/maps/china-history

* Chinese Text Project, 由DonaldSturgeon开发:

http://ctext.org/wiki.pl?if=en

* The Digital Gazetter of the Song Dynasty, 由加州大学默塞德分校的RuthMostern 和Elijah Meeks 开发:

http://songgis.ucmerced.edu/

* HTML5 Word Cloud Generator, Guan-tin Chien开发:

http://timdream.org/wordcloud/

* MARKUS, 由莱顿大学的Hilde De Weerdt和Brent Hou Ieong Ho 开发:

http://dh.chinese-empires.eu/beta/index.html

* Stanford Chinese Natural LanguageProcessing (NLP), 由 Dan Jurafsky, Christopher Manning 及其团队开发:

http://nlp.stanford.edu/projects/chinese-nlp.shtml#cws

* Academia Sinica’s ChineseWord Segmentation System, 由台湾「数位典藏国家型科技计画」开发（the National Digital Archives Program）:

http://ckipsvr.iis.sinica.edu.tw/

**本文所提到的概念的介绍（对人文科学学者来说可能已经熟悉）：**

* 对网络分析（network analysis）的介绍，来自：

Scott Weingart:http://www.scottbot.net/HIAL/?tag=networks-demystified&paged=2

* 对正则表达式（regular expressions, regexes）的介绍， 来自：

Shih-peiChen:http://projects.iq.harvard.edu/files/cbdb/files/text\_extraction\_regex\_shihpei.ppt?m=1438182331

* 对话题建模（topic modeling）的介绍，来自：

Ted Underwood:http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

封面图片来源：image.baidu.com