由中国人民大学信息资源管理学院冯惠玲教授、刘越男教授、严承希博士和哲学院王小伟副教授主讲的《数字人文导论》是中国人民大学本科生通识核心课程之一，面向全校本科生开设。课程内容包括数字人文的基本概念、历史源流、发展脉络、基本方法和技术、前沿议题等。本系列推文是该课程学生的学习成果展示。

本期作者

辛净梵 高瓴人工智能学院2020级

马乾朔 外国语学院2020级

黎穎儀 信息资源管理学院2019级

一 案例引入

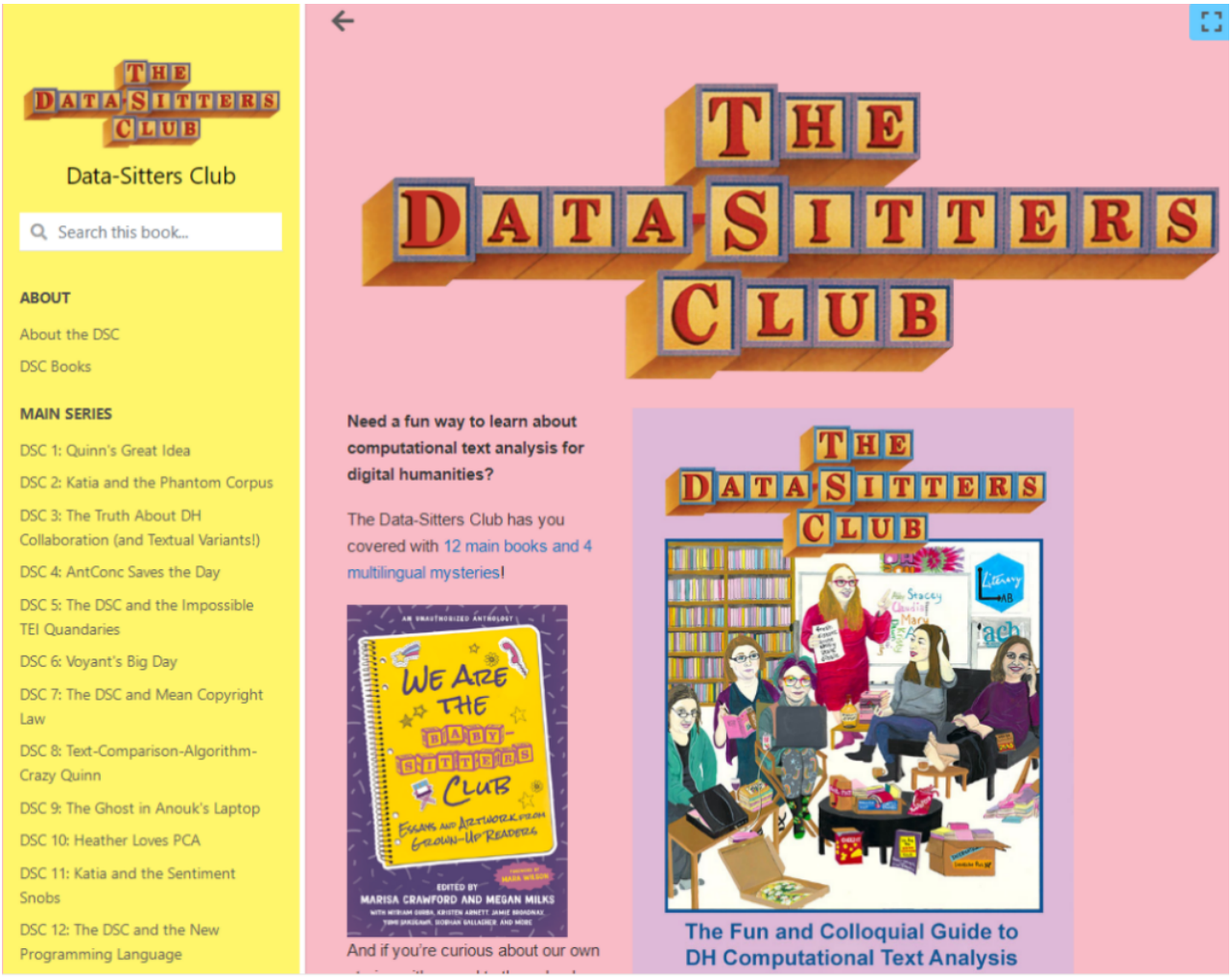


图1 The Data-Sitters Club主页面

“数据保护者俱乐部”项目（The Data-Sitters Club，项目网址：https://datasittersclub.github.io/site/books.html），是一个致力于采用数字工具进行文本分析的项目。该项目将丛书“The Baby-Sitters Club”整合建立电子文本数据库，并在其中展示了数据分析与可视化具体过程，对文本的语言风格、情感倾向等进行了详细的分析，为文学创作和学术研究提供了新思路。

二 项目背景

01

研究背景

斯坦福大学文学、文化和语言学部的学术技术专家奎因·多姆布罗夫斯基（Quinn Dombrowski）在拉斯维加斯度假时，开始怀念安·马丁（Ann M. Martin）撰写的关于20世纪90年代美国中上阶层郊区少女时代的标志性系列丛书“The Baby-Sitters Club”。

由于关于该系列丛书的学术研究稀少，奎因便想将数字人文文本分析工具和方法应用于该系列丛书（简称“BSC”）的语料库，详细描述如何进行文本分析并说明所有的步骤。于是，奎恩选择在DH（数字人文）Twitter上发布信息，最终找到了一些志同道合的朋友，一起创办了“数据保护者俱乐部”（The Data-Sitters Club）。

02

团队介绍

李•斯卡勒鲁普•贝塞特

乔治敦大学比较文学博士，主要研究魁北克文学的翻译。从事数字教育学的工作。

蒂亚•鲍尔斯

不列颠哥伦比亚大学副教授，主要研究十九世纪的俄罗斯文学。

玛丽亚•萨奇科•塞西雷

巴德学院的文学副教授，同时也是巴德实验人文学科跨学科课程和中心的创始主任，主要研究儿童文学。

奎因•东布罗夫斯基

斯坦福大学文学、文化和语言学部的学术技术专家。“数据保护者俱乐部”是她第一个也是唯一的英语文本分析项目。

阿努克•朗

（爱丁堡大学）

在悉尼读六年级时接触到“The Baby-Sitters Club”；在英国生活了20多年。

鲁皮卡•里萨姆

塞勒姆州立大学中等和高等教育及英语专业的副教授，喜欢数字人文学科。

马克•阿尔吉-休伊特

斯坦福大学的英语助理教授，斯坦福文学实验室主任。

伊莎贝尔•格里博蒙特（利物浦大学）

在比利时长大，是比利时/法国数据保护者。

此外，还有一些特邀数据保护者，如伊莉莎·贝谢罗-邦达尔、希瑟·弗罗利希、安妮·k·拉马尔、马修·萨格和杰夫·塔尔森。

三 项目实施过程

项目包括两个系列，分别是主要系列（Main Series）和多语言的奥秘系列（Multilingual Mysteries）。主要系列（Main Series）丛书有12册，多语言的奥秘系列（Multilingual Mysteries）丛书有4册，网站上分别用DSC1—DSC12和 DSCM1—DSCM4来表示。

01

MAIN SERIES

DSC#1

本章介绍了数据保护者俱乐部（The Data-Sitters Club）的成立过程。

DSC#2

由于网上搜集到的BSC书籍和通过扫描和光学字符识别（OCR）数字化的最新版本书籍有很多差异（如下图），俱乐部成员建立了这两个书籍的数据集。本章展示并介绍了该数据集。

1	Book	Change	Type	Context
189	006 Kristy's Big Day	remove "on the roll of film in the Polaroid"	Technology	There were two shots left on the roll of film in the Polaroid.
190	006 Kristy's Big Day	when the film was developed --> we looked at the image	Technology	When the film was developed, we saw that Tony was slumped over
191	006 Kristy's Big Day	130 --> 225	Money (inflation)	Watson and Mom took the members of the Baby-sitters Club aside and handed us each an envelope containing a check for one hundred and thirty dollars.
192	006 Kristy's Big Day	120 --> 200	Money (inflation)	"That's one hundred and twenty for a job very well done," Mom told us.
193	006 Kristy's Big Day	10 --> 25	Money (inflation)	"And a ten-dollar bonus," added Watson
194	007 Claudia and Mean Jeanine	vice president --> vice-president	Language (orthography)	I'm the vice president of the club.
195	007 Claudia and Mean Jeanine	junk food --> junk-food	Language (orthography)	I'm a junk food addict.
196	007 Claudia and Mean Jeanine	Remove perm references + "which had been long and fluffy"	Fashion	"I love it! Did you get it permed again?" "Yup. But after this I might let the perm grow out." Stacey sat on my bed. Her blonde hair, which had been long and fluffy, was now cut to just above her shoulders.
197	007 Claudia and Mean Jeanine	"Nineteen dollars even," she announced. She said this --> She announced the total	Money (removal)	"Nineteen dollars even," she announced. She said this after one pretty quick glance at the bills and change, which is why she's our treasurer.

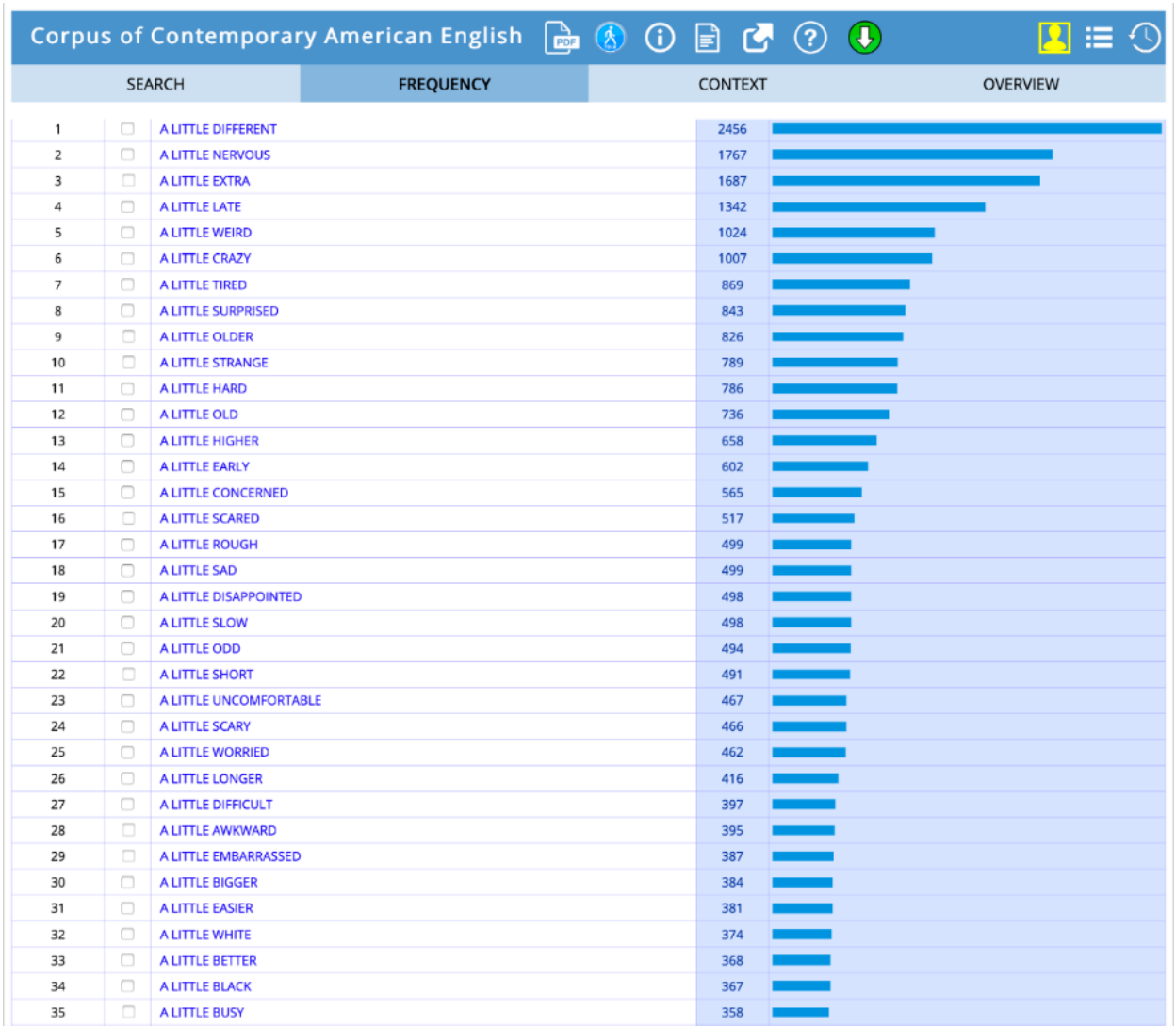
表 1 BSC电子书籍与OCR版本集差异展示

DSC#3

本章介绍了项目使用的文本校对方法。识别文本的两个版本之间的差异需要计算方法，而文本校对是一种最直观易懂的计算文本分析方法，如利用Diff Match Patch，Juxta和CollateX等工具。

DSC#4

本章利用AntConc软件对“a little”后常常出现的形容词做了词频分析，发现不仅这些形容词中的大多数有负面含义，而且“a little”和形容词之间常常利用其他方法来推迟负面判断出现的时刻（如语气词well）。



1	A LITTLE DIFFERENT	2456
2	A LITTLE NERVOUS	1767
3	A LITTLE EXTRA	1687
4	A LITTLE LATE	1342
5	A LITTLE WEIRD	1024
6	A LITTLE CRAZY	1007
7	A LITTLE TIRED	869
8	A LITTLE SURPRISED	843
9	A LITTLE OLDER	826
10	A LITTLE STRANGE	789
11	A LITTLE HARD	786
12	A LITTLE OLD	736
13	A LITTLE HIGHER	658
14	A LITTLE EARLY	602
15	A LITTLE CONCERNED	565
16	A LITTLE SCARED	517
17	A LITTLE ROUGH	499
18	A LITTLE SAD	499
19	A LITTLE DISAPPOINTED	498
20	A LITTLE SLOW	498
21	A LITTLE ODD	494
22	A LITTLE SHORT	491
23	A LITTLE UNCOMFORTABLE	467
24	A LITTLE SCARY	466
25	A LITTLE WORRIED	462
26	A LITTLE LONGER	416
27	A LITTLE DIFFICULT	397
28	A LITTLE AWKWARD	395
29	A LITTLE EMBARRASSED	387
30	A LITTLE BIGGER	384
31	A LITTLE EASIER	381
32	A LITTLE WHITE	374
33	A LITTLE BETTER	368
34	A LITTLE BLACK	367
35	A LITTLE BUSY	358

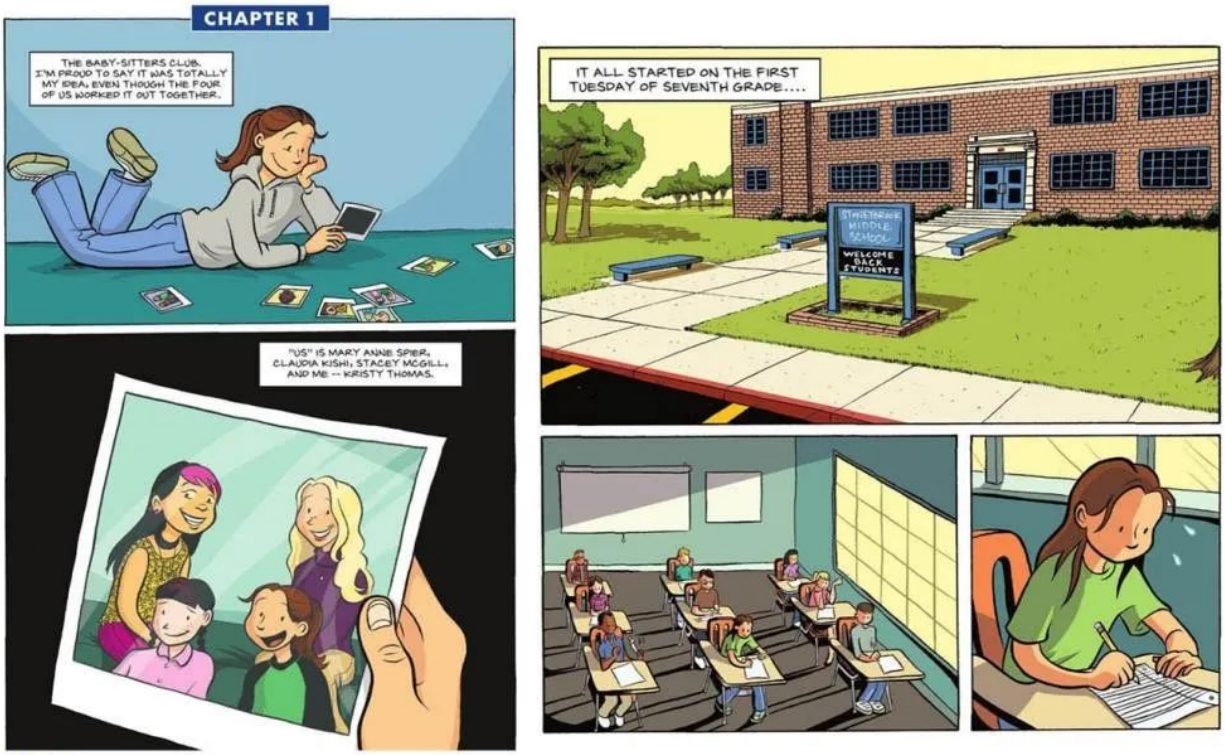
图 2 “a little”后出现的形容词的词频分析

can and set the slime on the table . " I — I feel	a little ... a little ... faint , " Dawn replied weakly . And with that , she
he 's a great baby-sitter , but that the meeting was	a little ... awkward . Then just see what he says . " " I think that '
, Laurel and Patsy , belong to the Krushers . Jake's always been	a little ... coordination-challenged . He had been improving , though , before the team
of her seat and started hugging me . Sometimes Melissa can be	a little ... excitable . But I didn't mind . I was psyched . Ms.
set the slime on the table . " I — I feel a little ...	a little ... faint , " Dawn replied weakly . And with that , she flopped over
I ca n't really picture her baby-sitting . She seems	a little ... I do n't know ... a little young for her
. These are cute . " " I guess , " agreed Sunny . " But I want something	a little ... I don't know ... wilder . Like these . " " Those ? " I said . "
can't I find someone for myself ? " " Maybe your standards are	a little ... I don't know ... too high , " Stacey suggested delicately . I
of the play , but from what I did see , it seems	a little ... I don't know ... a little insulting . " " What about artistic
, don't take this wrong , but you know , they do act	a little ... " " Immature , " I finished . " Well , yeah . " " Don't worry , " I said . "
to know about his hair . " " Well , yeah , but isn't this	a little ... mean ? " Sanji's smile disappeared . " Don't do anything to
could see a few of the Renwick 's ritzy customers looking	a little ... shocked . We did n't care . It was a cool ,
my skirt looks so great on you , after all . It's	a little ... tight . " Everyone gasped . " Are you implying that I'm fat ? "
driveway . Did you notice my dad 's reaction ? Not negative , but	a little ... uncomfortable . He gets that way when I mention anything to
changing . Ever since Sunny's mom got cancer , Sunny has seemed	a little ... wild or something . She takes risks . She's daring . And

图 3 具体展示a little”后出现的形容词

DSC#5

在本章中，奎因和Lee讨论了处理漫画小说的方法，并提到了漫画标记语言CBML。CBML是文本编码计划（Text Encoding Initiative，TEI）协议的变体，最初是为标记不同种类的文本而开发的，它添加了专门为漫画书设计的元素、属性和词汇。




```
<text>
    <body>
      <div type="chapter" n="01" xml:id="c01c01">
        <div type="narration" xml:id="cs01">
          <p>The Baby-Sitters Club. I'm proud to say it was totally my idea, even though the four of us worked it out together.</p>
        <p>"Us" is Mary Anne Spier, Claudia Kishi, Stacey McGill, and me -- Kristy Thomas.</p>
        </div>
        <div type="scene" subtype="mr_redmont_classroom" xml:id="cs02">
          <p>It all started on the first Tuesday of seventh grade...</p>
```

图 4 CBML处理漫画小说的结果展示

DSC#6

本章详细介绍了Voyant Tools这一基于网络的数字人文文本分析工具。上传文本后，该工具会创建一系列的可视化效果，从一个新的角度对文本进行解释和分析。数据可视化的方法包括直线图、图表、词云、气泡图等。

图 5 应用Voyant Tools中多样的数据可视化方法

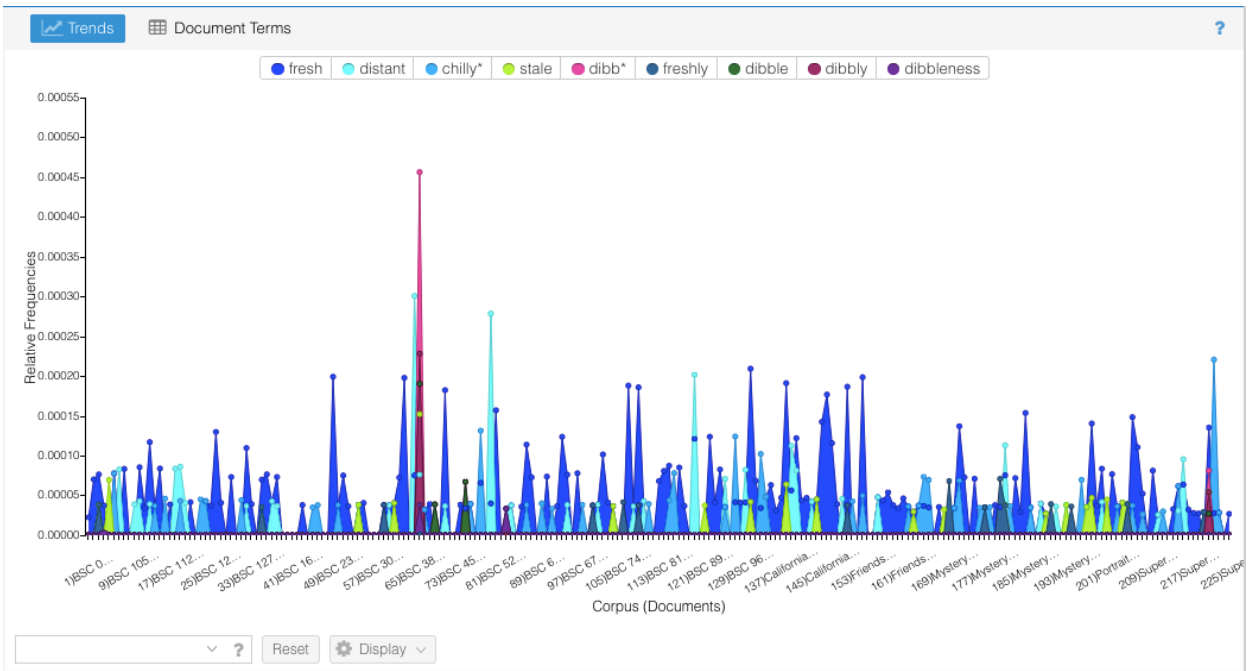


图 6 应用Voyant Tools工具进行文风统计

DSC #7

本章主要讲述版权、权限和责任道德等的相关问题。文章中以“The Data-Sitters Club”为例讲解各种问题，如在网站上显示所有版本的书籍的完全封面是否完全合法、在网站加入书中包含的一些图形小说面版是否存在版权问题、在使用语料库时可能会遇到的问题等。

图 7 部分书籍封面

DSC #8

本章主要介绍了比较文本的方法和内容，以原著中在结构上重复性高的第二章作为切入点，使用欧氏距离、余弦相似性等方法和Tableau软件与其他章节作对比计算。分析发现，第二章在读者眼中虽然重复性高，但在文本比较法下的结果与读者感受是相反的。

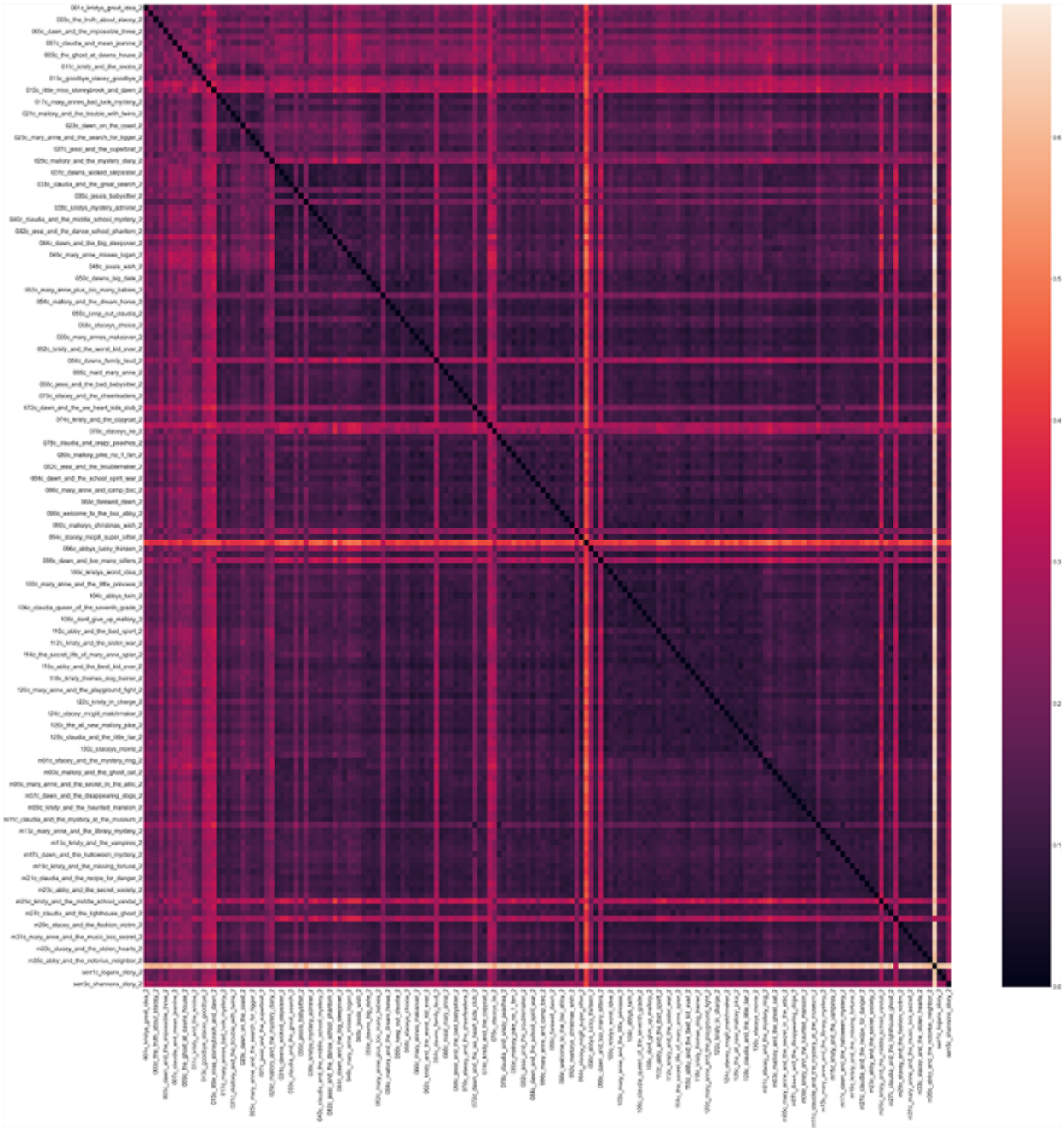
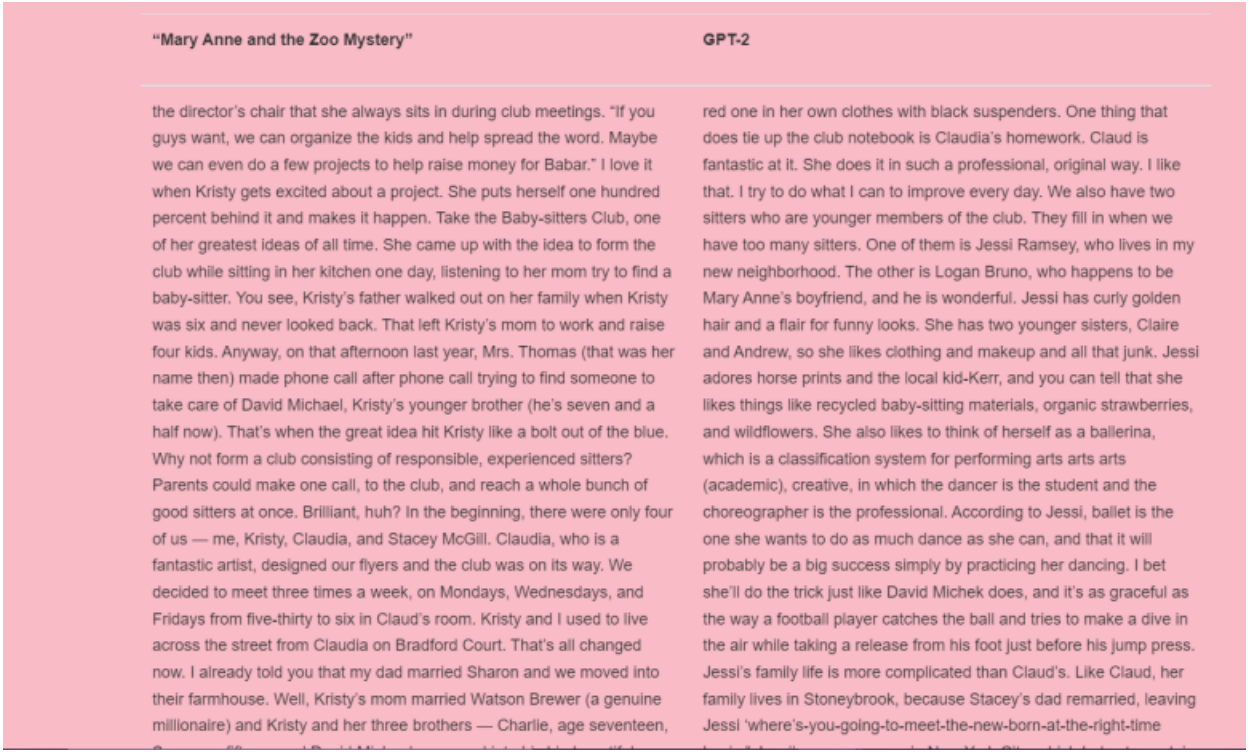


图 8 第二章与其他章节的文本相似度矩阵

DSC #9

在本章中，Anouk Lang用GPT-2与AI Dungeon等安装包，使用不同的语料库，以神经网络生成对话文本，然后使用机器学习不断训练GPT-2，希望能生成符合原著角色性格的对话以及文本。他也尝试把文学、叙事和语言学运用到机器学习中，令其更“人性化”，但遗憾的是该尝试并未成功。



“Mary Anne and the Zoo Mystery”	GPT-2
the director’s chair that she always sits in during club meetings. “If you guys want, we can organize the kids and help spread the word. Maybe we can even do a few projects to help raise money for Babar.” I love it when Kristy gets excited about a project. She puts herself one hundred percent behind it and makes it happen. Take the Baby-sitters Club, one of her greatest ideas of all time. She came up with the idea to form the club while sitting in her kitchen one day, listening to her mom try to find a baby-sitter. You see, Kristy’s father walked out on her family when Kristy was six and never looked back. That left Kristy’s mom to work and raise four kids. Anyway, on that afternoon last year, Mrs. Thomas (that was her name then) made phone call after phone call trying to find someone to take care of David Michael, Kristy’s younger brother (he’s seven and a half now). That’s when the great idea hit Kristy like a bolt out of the blue. Why not form a club consisting of responsible, experienced sitters? Parents could make one call, to the club, and reach a whole bunch of good sitters at once. Brilliant, huh? In the beginning, there were only four of us — me, Kristy, Claudia, and Stacey McGill. Claudia, who is a fantastic artist, designed our flyers and the club was on its way. We decided to meet three times a week, on Mondays, Wednesdays, and Fridays from five-thirty to six in Claud’s room. Kristy and I used to live across the street from Claudia on Bradford Court. That’s all changed now. I already told you that my dad married Sharon and we moved into their farmhouse. Well, Kristy’s mom married Watson Brewer (a genuine millionaire) and Kristy and her three brothers — Charlie, age seventeen,	red one in her own clothes with black suspenders. One thing that does tie up the club notebook is Claudia’s homework. Claud is fantastic at it. She does it in such a professional, original way. I like that. I try to do what I can to improve every day. We also have two sitters who are younger members of the club. They fill in when we have too many sitters. One of them is Jessi Ramsey, who lives in my new neighborhood. The other is Logan Bruno, who happens to be Mary Anne’s boyfriend, and he is wonderful. Jessi has curly golden hair and a flair for funny looks. She has two younger sisters, Claire and Andrew, so she likes clothing and makeup and all that junk. Jessi adores horse prints and the local kid-Kerr, and you can tell that she likes things like recycled baby-sitting materials, organic strawberries, and wildflowers. She also likes to think of herself as a ballerina, which is a classification system for performing arts arts arts (academic), creative, in which the dancer is the student and the choreographer is the professional. According to Jessi, ballet is the one she wants to do as much dance as she can, and that it will probably be a big success simply by practicing her dancing. I bet she’ll do the trick just like David Michek does, and it’s as graceful as the way a football player catches the ball and tries to make a dive in the air while taking a release from his foot just before his jump press. Jessi’s family life is more complicated than Claud’s. Like Claud, her family lives in Stoneybrook, because Stacey’s dad remarried, leaving Jessi ‘where’s-you-going-to-meet-the-new-born-at-the-right-time

图 9 用GPT-2生成的与语料库最相似的文本

DSC #10

本章先是讲解了主成分分析（PCA）的定义及其原理，并用YAMS语料库与“The Baby-Sitters Club”进行对比，发现一些不太显眼的名词反而推动了情节的发展，强化了情节间的关系。

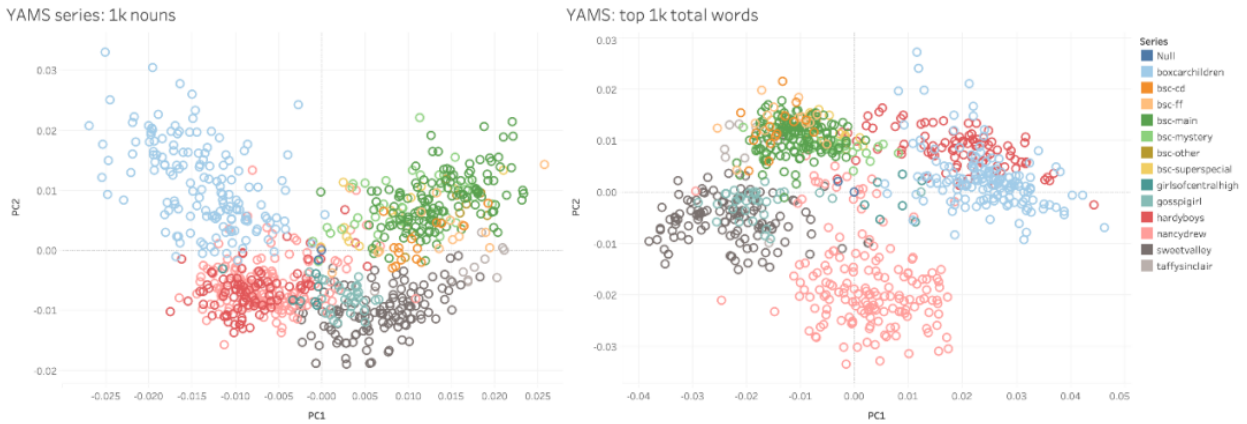


图 10 利用PCA分析YAMS语料库与“The Baby-Sitters Club”的异同

其后，本章对典型性代码进行观察和可视化，并把运行思路和结果带入到PCA的运行制作中，发现语料库中词汇多样性的缺乏是因为计算机在处理内容时，把能使代表情节更突出的词语删除，因此人们在看相关语料库时会觉得单调，且联想不起具体的情节。

图 11 语料库中词汇

DSC #11

本章讲述情感分析的技术方法、评价方式、在文学文本使用的效果、对原著的分析结果。在分析中作者发现，“The Baby-Sitters Club”的积极和消极情绪均比他们预期中的要少。由于作者用于计算和分析的句子数量较少，所以分析结果不一定是正确的，他们也从而开始思考“数据的基本适用性”问题。

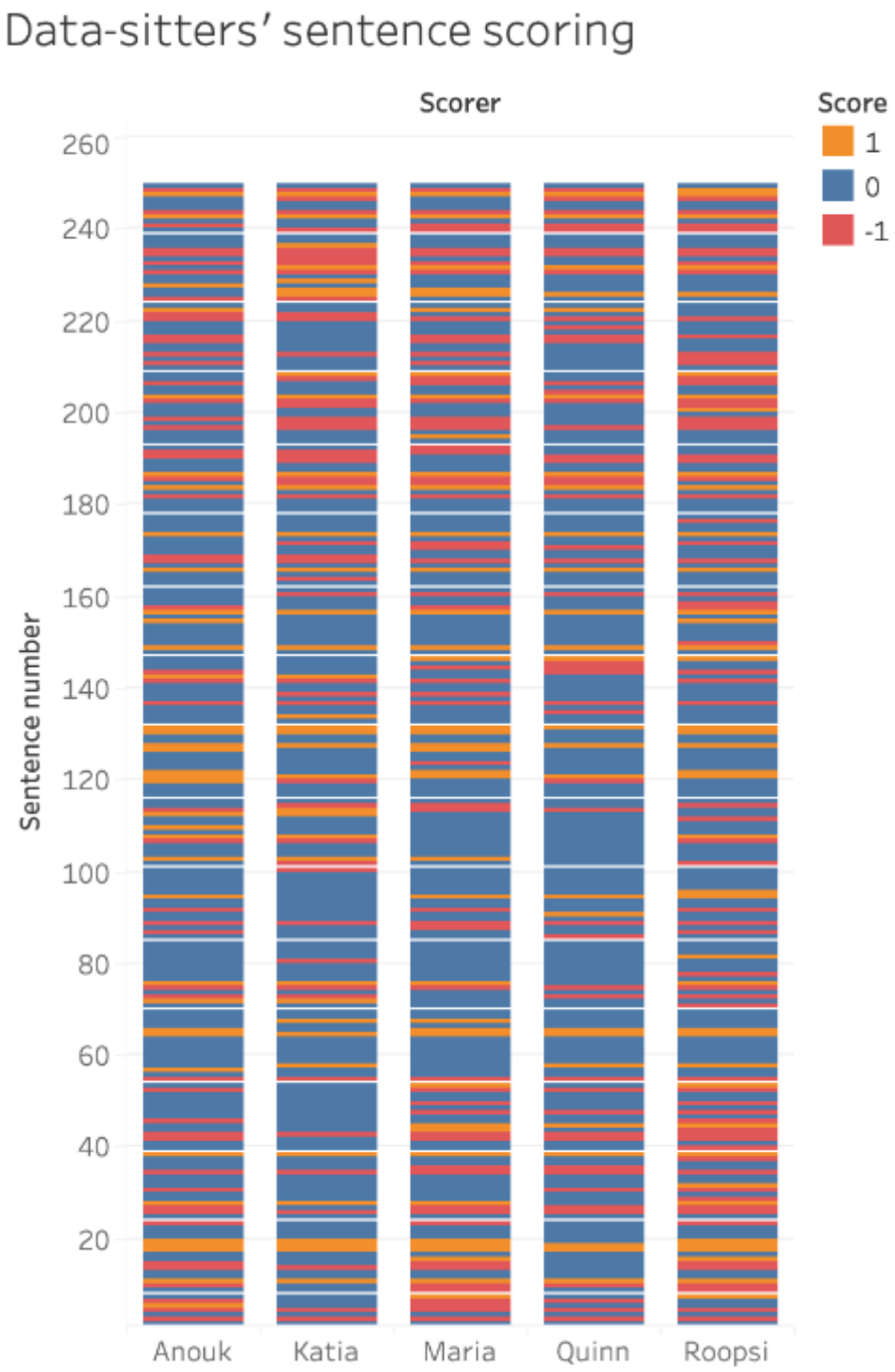


图 12 Baby-Sitters Club 的五位不同作者的前250个句子的情绪得分

（其中1代表正面陈述，0代表中立陈述，-1代表负面陈述）

02

多语言的奥秘DSCM

这是本书的第二个模块，给出了本系列书籍多语言译本两个主要的数据可视化成果。

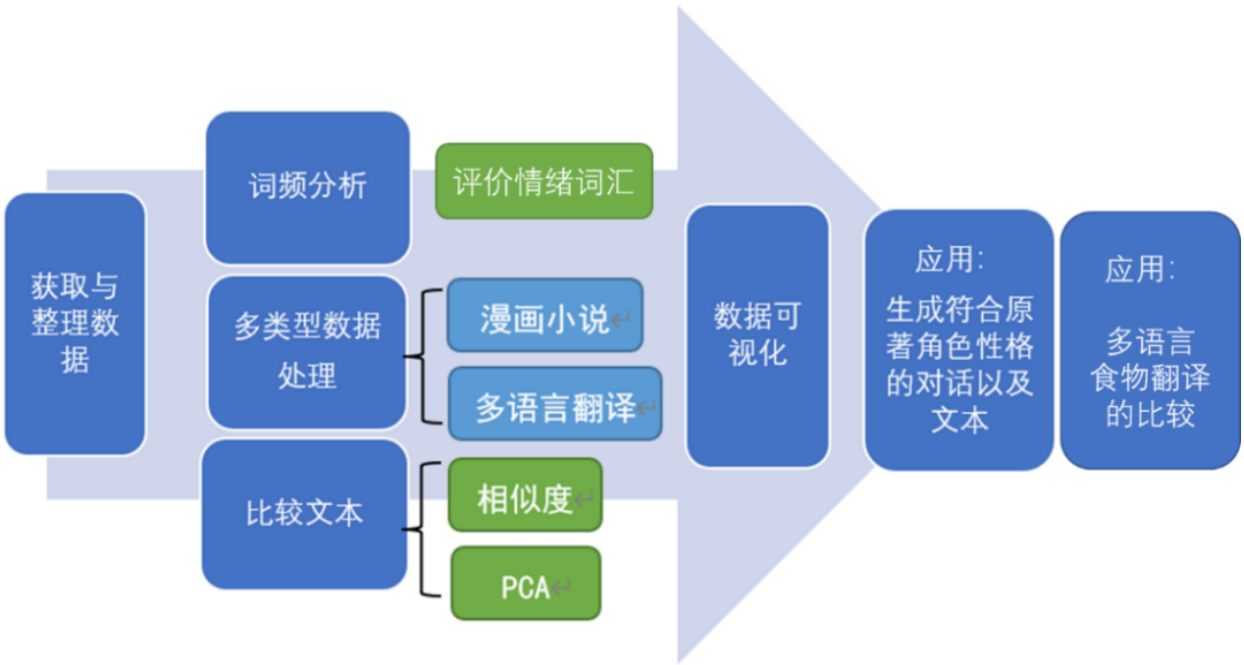
首先，为了解不同语言译者翻译本书的一致性程度，作者运用命名实体识别（NER）抽取非结构化实体，总结出不同语言译本的翻译特点：（1）在人名翻译上存在文化偏向，如法译非常喜爱

“spoiled brat”，同时部分译本对姓氏翻译存在缺漏；（2）在翻译风格上，比利时译本进行全面的欧洲化改编；魁北克译本增大差异；法国译本强调文本的美国性。

其次，作者利用Bleualign识别文中包含食物单词的所有句子对，得到“多语言系列书籍食谱”的可视化成果：（1）一种食品的使用多种翻译策略，如Twinkies注心蛋糕在比利时版中变得十分多样，译为巧克力、夹心饼干等；（2）无品牌食品经常被改造以适应当地的喜好，软糖变成巧克力软糖，PB&J三明治被分成两个单独的三明治；（3）本地品牌有时翻译为一个在外观、品味甚至偶尔听起来上与原始产品相似的品牌。

对多语言译本中食物翻译的探索，揭示了不同语言环境中的儿童出版业对美国文化的态度。

四 技术方法和工具



本项目体量庞大，作者使用了多种NLP技术进行文本分析并实现相关应用。项目可归类为四大模块：数据获取与整理、文本分析、数据可视化、相关应用实现。

01

数据获取与整理

通过网络爬虫获取多语言译本，并用Bleualign将翻译与英语对齐，并清洗元数据。

02

文本分析

a. 词频分析：利用Python分析词语与情绪类别的关系，评价情绪词汇在文本中的使用效果。

b. 多类型数据处理：①处理漫画小说形式数据，采用TEI变体中的漫画标记语言（CBML）处理漫画小说中的图像；②处理多语言译本，运用实体命名识别（NER）信息识别并抽取非结构化实体，并查找典型词语的上下文文本。

c. 比较文本：①使用欧氏距离、余弦相似性等方法和Tableau软件评估原著结构层面上的文本相似度；②用YAMS语料库与“The Baby-Sitters Club”进行主成分分析（PCA）对比。

03

数据可视化

a. 采用Voyant Tools文本分析工具，创建一系列的可视化效果。功能主要有：①统计词频；②显示单词在文本上下文中如何出现；③查看单词在文本语料库中出现的方式。

b. 以多种可视化类型呈现：多种直线图、图表、词云、气泡图等。

04

相关应用实现

利用GPT-2与AI Dungeon等安装包及不同的语料库，以神经网络生成对话文本，然后使用机器学习不断训练GPT-2，生成符合原著角色性格的对话以及文本。

五 项目成果

(1) 建立电子文本数据库：较为全面地整合并数字化20世纪90年代美国中上阶层郊区少女时代的标志性系列丛书“The Baby-Sitters Club”，建立电子文本数据库，便于采用数字化的分析方式对文本进行风格与语料分析。

(2) 全面而新颖的文本分析：对该标志性系列文本进行学术研究。采用多元的文本分析工具对文本语言风格、翻译差异、情感倾向、主题内容建立数字化模型并可视化，兼具学术研究价值与创新美观性。

(3) 建立The Data-Sitters Club开源网站：展示数据分析与可视化具体过程，并生动描绘困难点的教程细节。为文本分析学者与爱好者提供数据渠道与开展思路。
网址：https://datasittersclub.github.io/site/

六 项目特色与评价

01

项目特色

(1) 网站艺术风格鲜明：网站配色及风格较独特，为国外九十年代的艺术风格，复古且独特的配色能引起读者的兴趣。另外，其呈现方式与书籍《The Baby-Sitters Club》相似，网站就像故事集，具有连续性和趣味性。

(2) 数据内容丰富：网站使用的数据，数量十分庞大且丰富，除了有文本内容还有图片内容等。除此之外，其分析方法和呈现方法较多，基本每个“章节”介绍的工具、数据内容和可视化方式都不同，内容十分丰富。

(3) 文字风格生动有趣：网站文字风格有别于一般的数据网站，其描写生动有趣，作者像撰写日志、故事般去记录项目内容。在浏览网站时能感受到作者是和自己分享其经历和成果，十分愉悦。

(4) 网站结构清晰：该网站结构简单，共四个一级栏目，每个一级栏目下只有二级栏目，方便读者搜索。另外，每个栏目的标题清晰易懂且有趣、网页排版合理，读者能获得良好的阅读体验。

02

项目价值及意义

(1) 建立了“The Baby-Sitters Club”语料库

与“The Baby-Sitters Club”相关的书籍有200多本，印刷品超过1.76亿册，还有电影一部、电影影集一系列等，数据量十分庞大，具备很多值得分析和研究的内容，建立语料库能为学者研究当中的文化内涵、性别角色的描绘、文学语言等提供很大便利，发挥重要作用。

(2) 尝试采用机器学习进行文学创作

“The Baby-Sitters Club”中有部分系列书籍是由代笔作家撰写的，其中的语言风格特点不一，通过研究对比其中的差异能剖析作品中的文学特点，在网站中也有依据其特点训练机器写作的尝试，为文学创作提供了一种可能性。

(3) 是少有的以女性经历为中心的数字人文文本分析

既有数字人文文本分析大多以男性为中心，团队中的一位研究者（玛丽亚·萨奇科·塞西雷）对原著中美国少女的设置十分感兴趣，即原著中不同的女性角色在社会上有各自的定位。以女性体验为中心以及对女权主义集体进行分析的方式，在数字人文文本分析中都是特别的，所以该项目在性别层面角度上看也是相对独特的。

03

项目评价

该项目网站内容丰富且艺术风格鲜明、文字风格生动有趣、结构和排版清晰，使读者拥有良好的阅读体验。项目也有一些不足之处，例如：

(1) 某些章节的文字内容过多，而且专业名词和所使用技术难度较大，没有基础的人可能看不懂或者没有耐心阅读。我们认为可以缩短一些篇幅或者把一个章节拆分一下；或者在进入正文内容前，可以先对所运用的技术和专业名词进行简单的介绍，更方便读者阅读。

(2) 某些章节中，作者运用很大的篇幅描写处理、分析数据和可视化等的过程，我们认为这些过程可以用视频或图表等方式呈现，让读者更直观地了解，也可以增加趣味性和可读性。

排版：郝李臻