

GALE
Digital
Scholar
Lab

数字学术实验室
操作指南2024



GALE

Digital Scholar Lab

数字学术实验室

介绍

当开展分析时，发现、清理和组织数据以及对历史文本的自然语言处理（NLP）常常是一项令人望而却步的任务，特别是想要生成有意义的分析结果时。**Gale数字学术实验室（DSL）** 扫除了这些障碍，让工作流程直观顺畅，研究者能够将更多的时间用于找到之前未被发现的数据、检验理论、分析结果和获取新的见解。

- **创建内容集**

Gale数字学术实验室允许用户创建自定义内容集，每个内容集最多包含10,000篇文档。用户可以在图书馆现有的Gale原始档案馆藏中检索，顺畅地选择文档并加入到他们的自定义内容集中。

- **分析内容集**

用户可以利用Gale数字学术实验室内置的文本分析和可视化工具分析和探讨数据。数字人文分析方法包括：命名实体识别（Named Entity Recognition）、主题建模（Topic Modelling）、词性标注（Parts of Speech）等。

- **管理和分享**

用户的内容集保存在Gale数字学术实验室中，便于他们为长期项目管理他们的研究。用户可以发表他们的研究结果，完全保留所有的知识产权，也可随意分享他们的分析结果。



Interface 登录界面

Library Menu: Gale China

English Signed in: Genesis



My Research Build Clean Analyze



可翻译操作界面语言/登录个人微软账户

GALE DIGITAL SCHOLAR LAB

Explore your institution's Gale Primary Sources using digital humanities methodologies to discover new insights. Create your account to begin.

我的研究/创建内容集/清理文本/分析

Get Inspired With Gale Research Showcase

Explore text and data mining projects created in Gale Digital Scholar Lab at Gale Research Showcase. Then submit your own project for inclusion.



Build Your Content Sets

- Use Gale Primary Sources archives
- Upload and use your own text files
- Download content sets for use elsewhere

[Learn more about the Build step »](#)



Clean Texts for Computational Analysis

- Apply stop words to your analysis
- Use flexible options to target specific character removal
- Reuse configurations across Content Sets

[Learn more about the Clean step »](#)



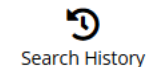
Analyze Content in Powerful New Ways

- Visualize data from up to 10,000 documents at a time
- Explore individual documents overlaid with analysis data
- Download the raw data and your visualizations

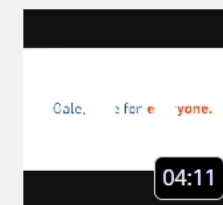
[Learn more about the Analyze step »](#)

数字人文研究指导/学习中心

Build 创建内容集



Build Content Sets by finding documents in your institution's Gale Primary Sources holdings or by uploading your own documents.



Video: Building a Content Set

Additional topics:

- [Search Strategies](#)
- [Understanding Search Results](#)
- [View All Build Help »](#)

Search

Search by keyword



Advanced Search

Refine your search by publication date, source, and more. [View all limiters in Advanced Search »](#)

Logical Operator Examples

AND: feminism AND gloria steinem

OR: comic book OR cartoon OR graphic novel

NOT: war NOT world war II

COMBINATION: cloning AND stem cells NOT sheep

检索输入框/高级检索

通过检索，添加Gale
历史一次文献资源

Upload

通过上传txt格式文本或手动编辑创建内容集
添加元数据便于文献管理

Drag and Drop .txt or .csv Files here or [Browse](#)

+ New Content Set

AF
AF China
AF-china title
AF-chinese-分析
AF-Chinese-分析2
AF-france

Get Upload Template

Create a Text Document

Manage All Uploads

10 MB at a time

Manage Uploads 管理上传文献

GALE DIGITAL SCHOLAR LAB

My Research Build Clean Analyze

Search by keyword Advanced Search

Add Note Add To Content Set Delete Search History

Manage All Uploads

Add your own documents, with metadata, for use in the Digital Scholar Lab.

[Improve your analysis by adding metadata](#)

Video: Building a Content Set

Additional topics:

- Search Strategies
- Understanding Search Results
- [View All Build Help »](#)

Upload Create a Text Document Edit Metadata Change Columns

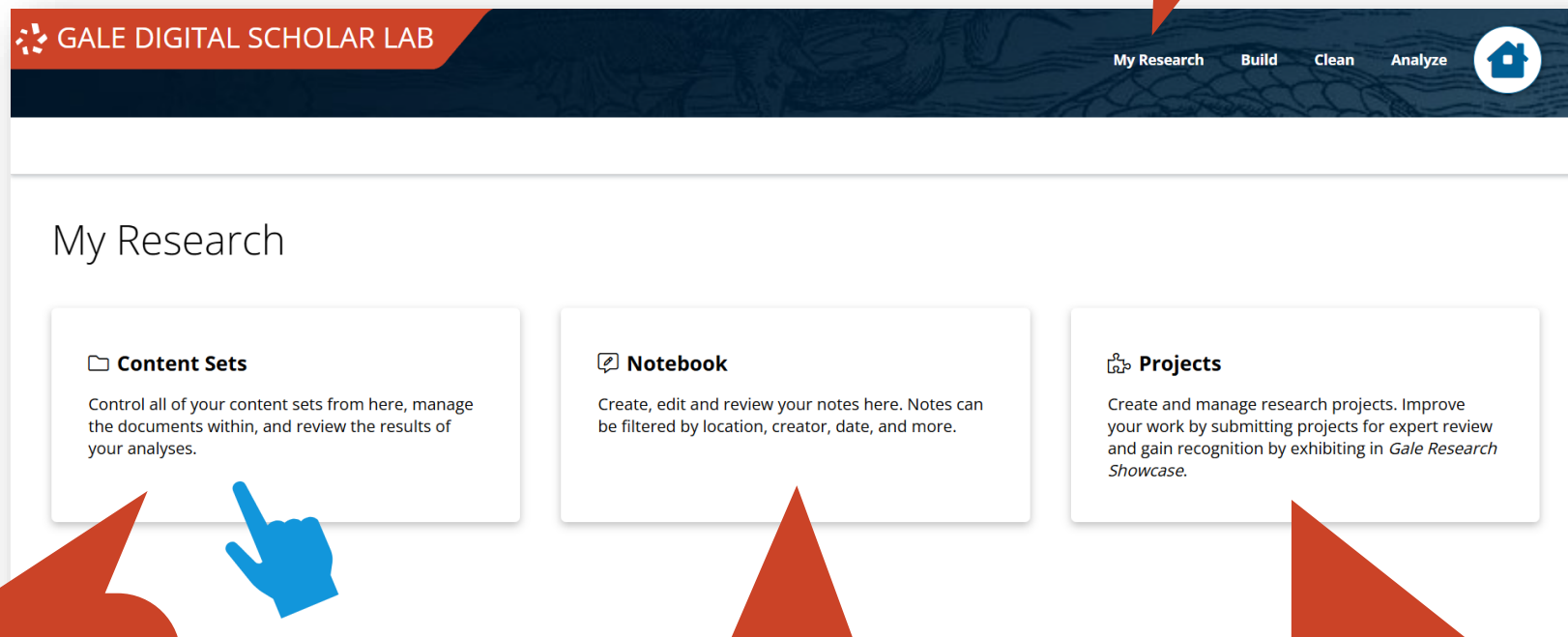
<input type="checkbox"/>	Date Uploaded	Title	Owner	Published	Author	Publisher	Publication Title	Document
<input type="checkbox"/>	1/4/2023	Effectiveness of inactivated and Ad5-nCoV 19 vaccines against BA.2 variant infection, severe illness		10/20/2022	Zhuoying Huang, Shuangfei Xu, Jiechen Liu, Linlin Wu, Jing Qiu and Nan Wang	BioMed Central Ltd.	BMC Medicine(Vol. 20, Issue 1)	Journal
<input type="checkbox"/>	1/4/2023	Gam-COVID-Vac, EpiVacCorona, and CoviVac effectiveness against lung injury during Delta and Omicron		10/10/2022	Anton Barchuk, Anna Bulina, Mikhail Cherkashin, Natalia Berezina, Tatyana Rakova and Darya Kuplevats	BioMed Central Ltd.	Respiratory Research(Vol. 23, Issue 1)	Journal

管理上传文献

管理显示项目与元数据
切换排序等

My Research 我的研究

我的研究



管理内容集

从这里控制所有内容集，管理其中的文档，并查看分析结果。

管理笔记本

在这里创建、编辑和查看您的笔记。笔记可按地点、创建者、日期等进行筛选。

创建和管理研究项目

创建和管理研究项目。通过提交项目供专家评审来改进您的工作，并通过在 Gale 研究展板上展示来获得认可。

Content Sets 内容集



- Add Note
- All Projects
- New Content Set
- New Folder
- Move
- Merge
- Duplicate
- Delete

My Content Sets

Control all of your content sets from here, manage the documents within, and review the results of your analyses.

Add Content

Manage All Uploads

添加笔记/研究项目/创建内容集/创建文件夹/移动内容集/合并内容集/复制内容集



Video: What is a Content Set?

Additional topics:

- Downloading Content Sets
- Sample Projects
- View All Content Set Help »

<input type="checkbox"/>	Type	Name	Build	Analyze
<input type="checkbox"/>	Folder	AF	Content Sets: 13	
<input type="checkbox"/>	Folder	law	Content Sets: 4	
<input type="checkbox"/>	Document	1865-1918 Electricity - Banner of Light - Essays and Articles	Documents: 1,296	Tools: 4 Runs: 7
<input type="checkbox"/>	Folder	war	Content Sets: 5	

内容集文件夹/内容集

内容集或文件夹详情

Content Sets 内容集

The screenshot displays the Gale Digital Scholar Lab interface for a content set titled "1865-1918 Electricity - Banner of Light - Essays and Articles". The interface is divided into two main sections: BUILD and ANALYZE.

BUILD Section:

- Documents (1296):** Gale Documents (1296)
- Timeframe:** 1891 - 1891
- Authors (297):** G. L. Ditson (13), W. J. Colville (13), Lilian Whiting (12)
- Archives Used (1):** American Historical Periodicals from the American Antiquarian Society (1296)
- Source Libraries (1):** American Antiquarian Society (1296)
- Document Type (2):** Essay (855), Article (441)

ANALYZE Section:

- Analyses in Progress:** No analyses currently in progress.
- Recent Successful Runs:**
 - Named Entity Recognition:** Setup: 1 - Mon Feb 19 00:02:19 EST 2024. Entities Found.
 - Ngrams:** Setup: Min2 Max5 Thres4 Electricity - No P No N No Spec - Tue Oct 11 00:38:47 EDT 2022. Includes Word Cloud and Term Frequency.
 - Sentiment Analysis:** Setup: Electricity No P No Num No Spec - Tue Oct 11 00:37:15 EDT 2022. Includes Sentiment Scores and Sentiment Over Time.
 - Topic Modeling:** Setup: 10Words 20Topics Electricity - Tue Oct 11 00:17:22 EDT 2022. Includes Topics and Topic Proportion.

Navigation and utility buttons include: Back, Add To Project, Add Note, Download, Delete, Add Content, and View Results.

内容集概述与详情

内容集详情

内容集管理

管理内容集分析工具

最新分析结果

This screenshot shows the search results page for the content set "1865-1918 Electricity - Banner of Light - Essays and Articles". The results are sorted by Document Title.

Results:

- The 42d Anniversary:** OCR Confidence: 45%. Publication: Banner of Light, April 12, 1890. Archive: American Historical Periodicals from the American Antiquarian Society. Source library: American Antiquarian Society. Context Type: Periodical. Document Type: Essay.
- The 44th Anniversary:** OCR Confidence: 61%. Publication: Banner of Light, April 9, 1892. Archive: American Historical Periodicals from the American Antiquarian Society.

Filter Your Results:

Archives	Content type	Document type
Publication Date	Publication title	Publication Sections
Publication country or territory	Publication state/province	Publication city
Languages	Author - items by	Source library
Illustrated works	OCR confidence range	Search Within
Document Title		

GALE DIGITAL SCHOLAR LAB

My Research Build Clean Analyze

Add Note New Configuration Save As Test Configuration

Clean

Select Cleaning Configuration to Edit
Default Cleaning Configuration

Stop words
Set the words you want the Analysis Tools to ignore.
Choose a Starter List

- a
- about
- above
- across
- after
- afterwards
- again
- against
- all
- almost
- alone
- along
- already
- also
- although
- always
- am
- among
- amongst
- amongst
- amount
- an
- and
- another
- any
- anyhow
- anyone
- anything
- anyway
- anywhere
- are
- around
- as
- at
- back
- be
- became
- because

Ignore stop words case

Clear List

Text Correction
Options for automatic text correction that will be applied before each Analysis

- Turn on all options
- Text Modification**
 - All lower case
- Characters**
 - Remove all extended ASCII characters
 - Remove all number characters
- Special Characters**
 - Remove all special characters
 - [Set specific characters >](#)
- Punctuation**
 - Remove all punctuation
 - [Set specific punctuation >](#)
- Spacing**
 - Remove all tabs
 - Remove all line breaks
 - Reduce multiple spaces to one space (ex: "hello there" becomes "hello there")
- Document Sections**
 - Remove body text
 - Remove all non-body content
 - [Set specific sections >](#)
- Replacements**

Replace this...	With this...
<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>

Add a Row

Configuration Notes
Space to make notes or describe the purpose of this configuration.

默认清理配置，载入词表，去除字符

Video: Cleaning a Content Set
Additional topics:

- Creating a Clean Configuration
- Applying During Analysis
- View All Clean Help >

GALE

ABOUT | HELP | DICTIONARY | CONTACT US | TERMS OF USE | PRIVACY POLICY | ACCESSIBILITY | END SESSION

Gale, here for everyone.

选择清理配置

停用词表

清理配置管理

清理配置笔记

文本纠正

Clean 清理

Text Correction

Options for automatic text correction that will be applied before each Analysis

Turn on all options

▼ Text Modification

All lower case

▼ Characters

Remove all extended ASCII characters

Remove all number characters

▼ Special Characters

Remove all special characters

[Set specific characters >](#)

▼ Punctuation

Remove all punctuation

[Set specific punctuation >](#)

▼ Spacing

Remove all tabs

Remove all line breaks

Reduce multiple spaces to one space (ex: "hello there" becomes "hello there")

▼ Document Sections

Remove body text

Remove all non-body content

[Set specific sections >](#)

▼ Replacements

Replace this...

With this...

[Add a Row](#)

更正

文档字符

特殊字符

标点

空格

文档拆分

替换

Set specific characters ▼

Equal sign (=)

Slash (/)

Percent (%)

Euro sign (€)

Tilde (~)

Angle brackets <>

Plus sign (+)

Hash (#)

Underscore (_)

Backtick (`)

Backslash (\)

Yen sign (¥)

Pipe (|)

At (@)

Dollar sign (\$)

Pound sterling (£)

Carot (^)

Set specific punctuation ▼

Period (.)

Comma (,)

Parentheses ()

Asterisk (*)

Hyphen (-)

Ampersand (&)

Question mark (?)

Colon (:)

Square brackets []

Apostrophe (')

En dash (-)

Ellipsis (...)

Exclamation point (!)

Semicolon (;)

Braces {}

Quotation marks ("")

Em dash (—)

Set specific sections ▼

Front Matter

Table of Contents

Title page

Back Matter

Preface

Index

Analyze 分析

Library Menu: Gale China English Signed in: Genesis

GALE DIGITAL SCHOLAR LAB My Research Build Clean Analyze


Add Note Add Tool Remove Run Selected

Analyze

Content Set
1865-1918 Electricity - Banner of Light - Essays and /~

Select All


Named Entity Recognition



Entities Found

Run Details
1 (Mon Feb 19 00:02:19 EST 2024) [View](#) [New Setup](#) Run Time: 3:50


Ngrams



Word Cloud Term Frequency

Run Details
Min2 Max5 Thres4 Electricity - No P No N I [View](#) [New Setup](#) Run Time: 1:32


Sentiment Analysis



Sentiment Scores Sentiment Over Time


Run Details
Electricity No P No Num No Spec (Tue Oct [View](#) [New Setup](#) Run Time: 0:58

Topic Modeling



Topics Topic Proportion

Run Details
10Words 20Topics Electricity (Tue Oct 11 0 [View](#) [New Setup](#) Run Time: 3:50



GALE ABOUT | HELP | DICTIONARY | CONTACT US | TERMS OF USE | PRIVACY POLICY | ACCESSIBILITY | END SESSION

Gale, here for everyone.

切换内容集

管理分析工具

切换分析结果/查看与配置

Analysis Tools 分析工具


SELECT ANALYSIS TOOLS ✕


<p>Add</p> <p>Document Clustering Toolkit: SciKit Learn Groups documents by similarity based on term frequency. Learn more about Document Clustering »</p>	<p>Add</p> <p>Parts of Speech Toolkit: spaCy Identifies the parts of speech within each document in your content set. Learn more about Parts of Speech »</p>
<p>Added</p> <p>Named Entity Recognition Toolkit: spaCy Identifies entities (nouns and numbers) across your content set. Learn more about Named Entity Recognition »</p>	<p>Added</p> <p>Ngrams Toolkit: Gale Digital Humanities Toolkit Analyzes the frequency of terms and phrases in your content set. Learn more about Ngrams »</p>
<p>Added</p> <p>Sentiment Analysis Toolkit: Gale Digital Humanities Toolkit Analyzes the polarity of terms to determine positive / negative sentiment. Learn more about Sentiment Analysis »</p>	<p>Added</p> <p>Topic Modeling Toolkit: Mallet Analyzes the terms across documents to determine the "topics" that occur. Learn more about Topic Modeling »</p>


Done

- 聚类 (Clustering)
- 词性标注 (Parts of Speech)
- N元语法 (Ngrams)
- 命名实体识别 (Named Entity Recognition)
- 情感分析 (Sentiment Analysis)
- 主题建模 (Topic Modeling)

Document Clustering 文档聚类

 Legend

 Tool Setup

 Run History

Tool Setup

Run Name

Number of Clusters

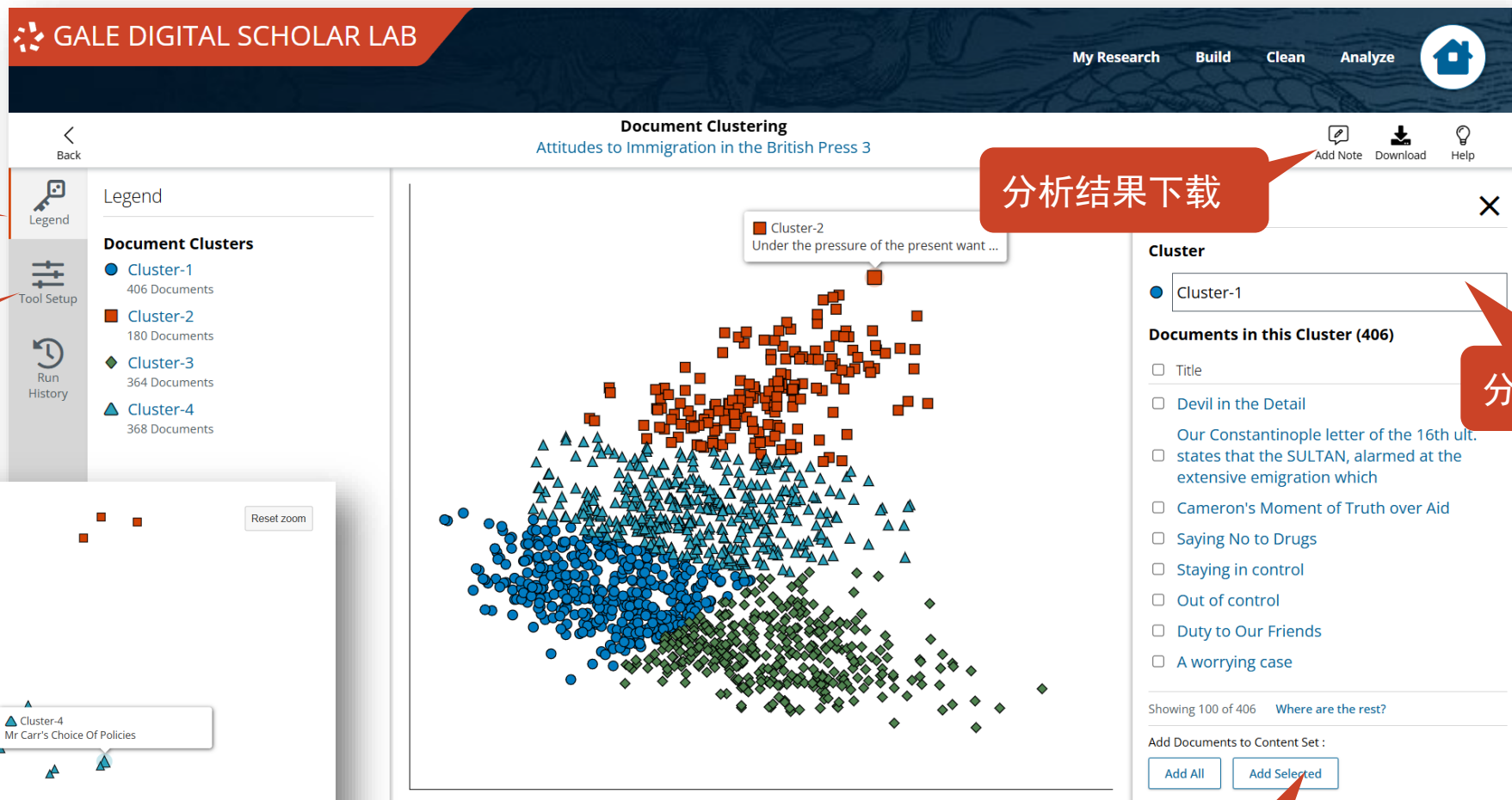
Cleaning Configuration



文档聚类

使用统计指标和方法分析内容集中的文档，围绕特定特征或属性对它们进行分组。该算法将所有点聚集到所需数量的簇中，并为属于每个簇的点着色。每个簇代表文档组，这些文档组彼此之间的相似性比与内容集中其他文档的相似性更高。

可视化分析结果



可视化工具

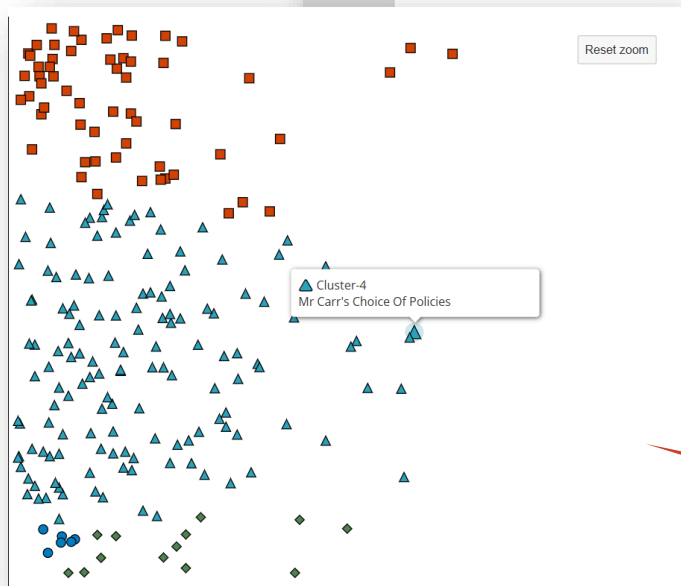
分析设置

分析结果下载

分组命名

放大显示

数据详情



Ngrams N-元语法

Tool Setup

Run Name

Ngram Size ⓘ

Single Range

Default: 1 Min: 1 Max: 6

Ngrams Occurrence Threshold ⓘ

Default: 2 Min: 1

Number of Ngrams Returned ⓘ

Max: 1000

Cleaning Configuration

Default Cleaning Configuration ▾

Run



N-元语法

提供了可视化内容集文档中某些单词或标记的频率的方法。

N 表示单词数，单元词有 1 个词，二元词有 2 个词，三元词有 3 个词，等等。

可视化分析结果

可视化切换

可视化工具

分析设置

GALE DIGITAL SCHOLAR LAB

My Research Build Clean Analyze

1865-1918 Electricity - Banner of Light - Essays and Articles

Legend

View

Top 100 Ngrams

Ngram Search

Enter terms to match ngram

trance speaker

banner light

inspirational speaker

Ngram	Count	Tokens
Content Set	978	1,330,273
LECTURERS' APPOINTMENTS AND ADDRESSES	25	1,176
LECTURER'S APPOINTMENTS AND ADDRESSES	24	1,175

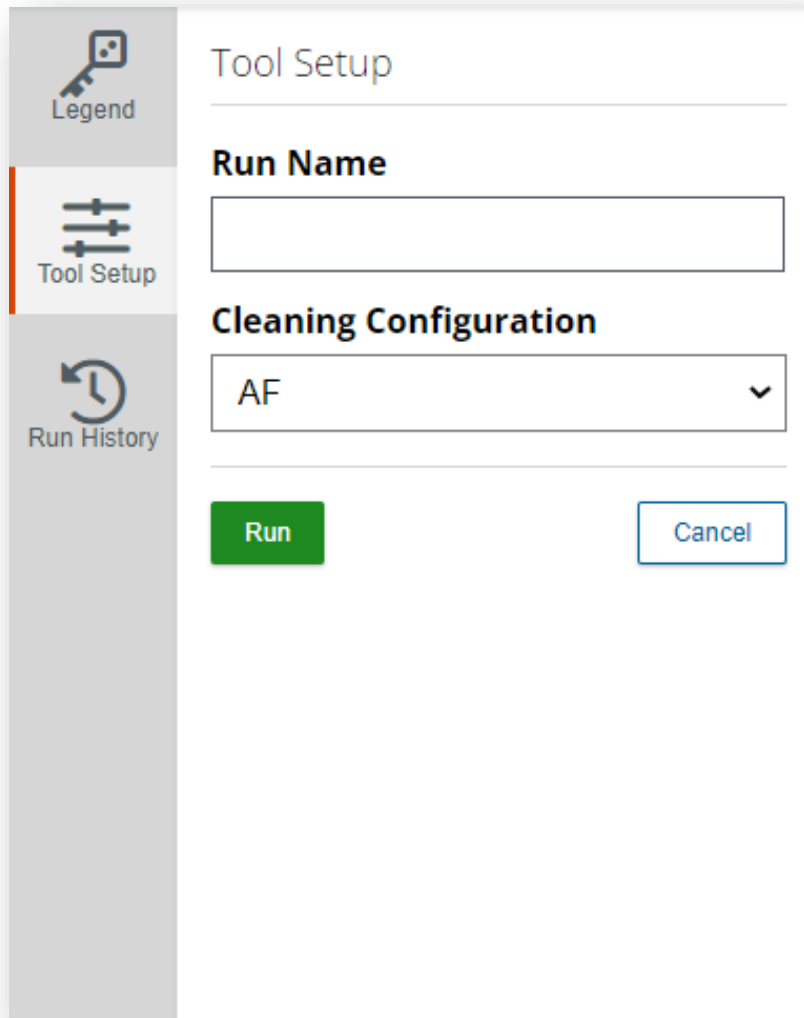
分析结果下载

1865-1918 Electricity - Banner of Light - Essays and Articles

Ngram	Count	Tokens
Content Set	414	1,330,273
The Anniversary	15	3,505
A General View of Spiritualism	15	2,442
The 44th Anniversary	12	2,661
The 42d Anniversary	10	4,404
Thirtv-Ninth Anniversary		

数据详情

Parts of Speech词性标注



Tool Setup

Legend

Tool Setup

Run History

Run Name

Cleaning Configuration

AF

Run

Cancel

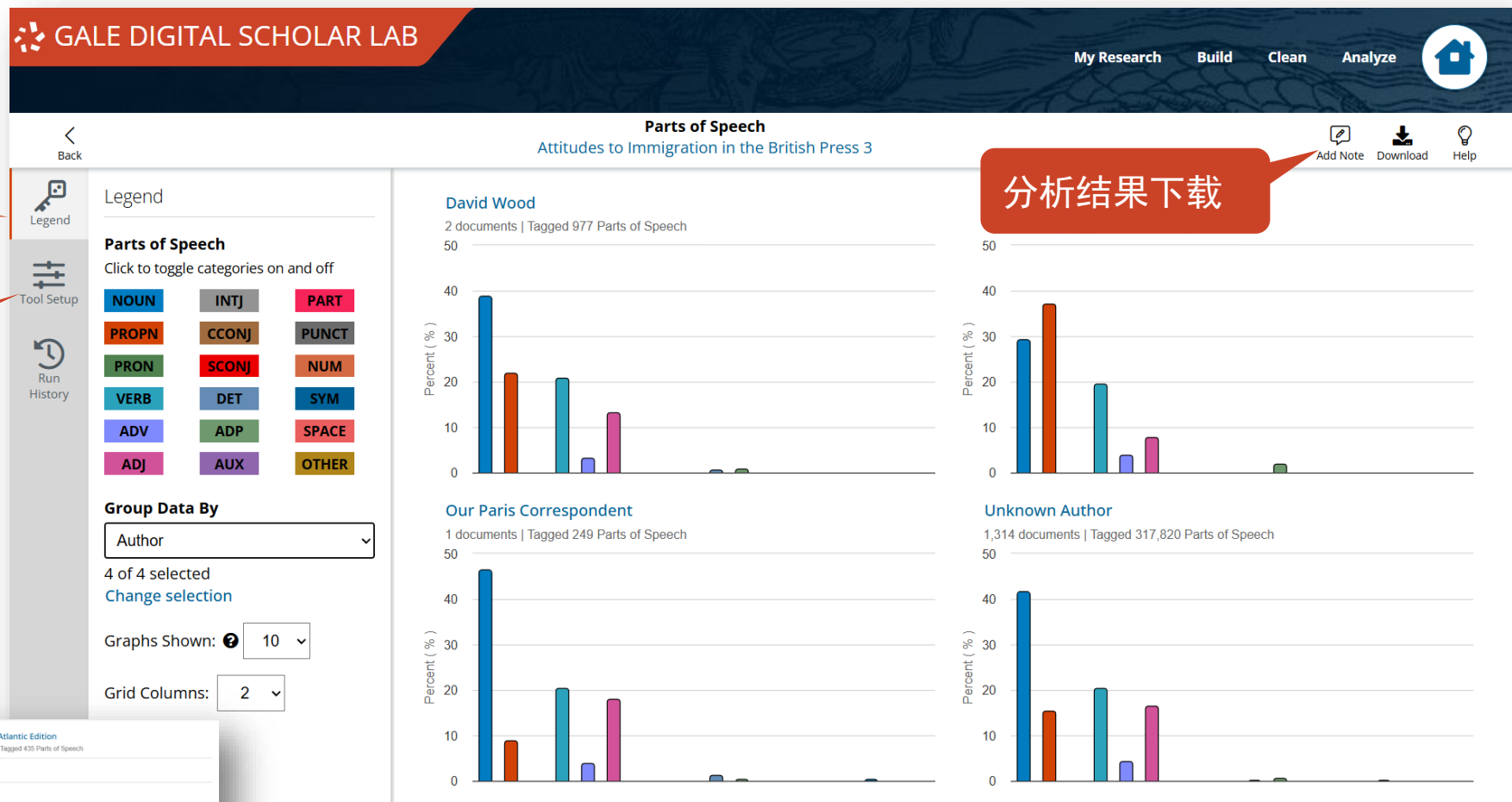


词性标注

它为用户提供了查看内容集中每个文档中短语构造方式的构建块。此工具可有效创建内容集的词典索引或词表。

词类包括：连词、代词、专有名词、符号、基数、形容词、连词、空格、限定词、副词、标点符号、助词、动词、感叹词、名词等。

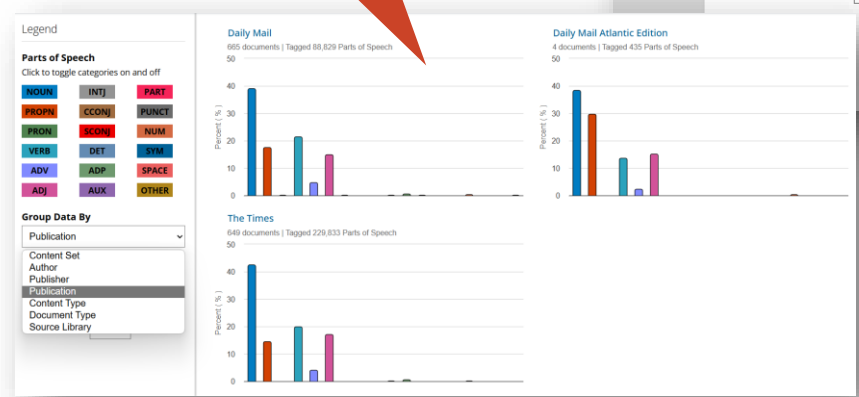
可视化分析结果



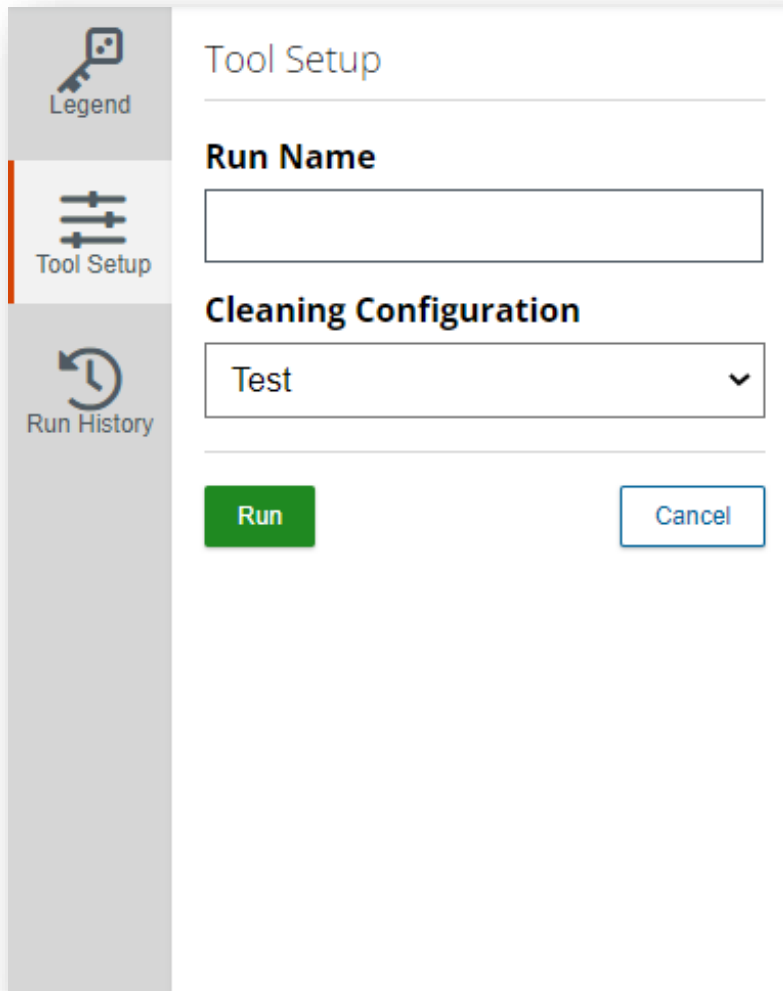
可视化工具

分析设置

切换显示



Named Entity Recognition 命名实体识别



The screenshot shows a software interface for Named Entity Recognition. On the left is a vertical sidebar with three icons: a key for 'Legend', sliders for 'Tool Setup', and a clock for 'Run History'. The main panel is titled 'Tool Setup' and contains a 'Run Name' text input field, a 'Cleaning Configuration' dropdown menu set to 'Test', and two buttons: a green 'Run' button and a blue 'Cancel' button.



命名实体识别

命名实体识别 (NER) 是一种自然语言处理方法，旨在将内容集中的每个术语识别并分类为特定的实体类别或“类”。

包括：数字、日期、事件、地方、地缘政治实体、语言、法律、地理、钱、文化团体、位置、组织、百分比、人、产品、测量、时间、艺术品等

可视化分析结果

GALE DIGITAL SCHOLAR LAB

My Research Build Clean Analyze

Named Entity Recognition
1865-1918 Electricity - Banner of Light - Essays and Articles

Legend

View
● Top 200 Entities by Count
○ Entity Search

Search for entities

Entity categories
 Category

- Date
- Time
- Geography
- Geo-Political Entity
- Place
- Artwork
- Event
- Law
- Product
- Person
- Measurement

Entity	Category	Documents	Count
Spiritualism	Organization	445	1462
Boston	Geo-Political Entity	390	1347
Spiritualists	Cultural Group	357	891
New York	Geo-Political Entity	350	770
Nature	Organization	218	474
Washington	Geo-Political Entity	194	469
second	Position	300	459
Jesus	Person	184	458
Chicago	Geo-Political Entity	158	427
Christian	Cultural Group	217	412
Spiritualist	Organization	205	354
Address	Organization	72	314
Bible	Artwork	158	292

Entity
Boston Geo-Political Entity

Entity Details
● Documents ○ Related Entities
Identified 1347 times across 390 documents

Title	Count
Message Department	5
Meetings in Boston	1
Original Essay	1
LECTURERS' APPOINTMENTS AND ADDRESSES	5
Movements of Platform Lecturers	3
Pamphlets Received	1
Is There a Science of Prophecy?	.

Showing 100 of 390 Where are the rest?

Add Documents to Content Set:
Add All Add Selected

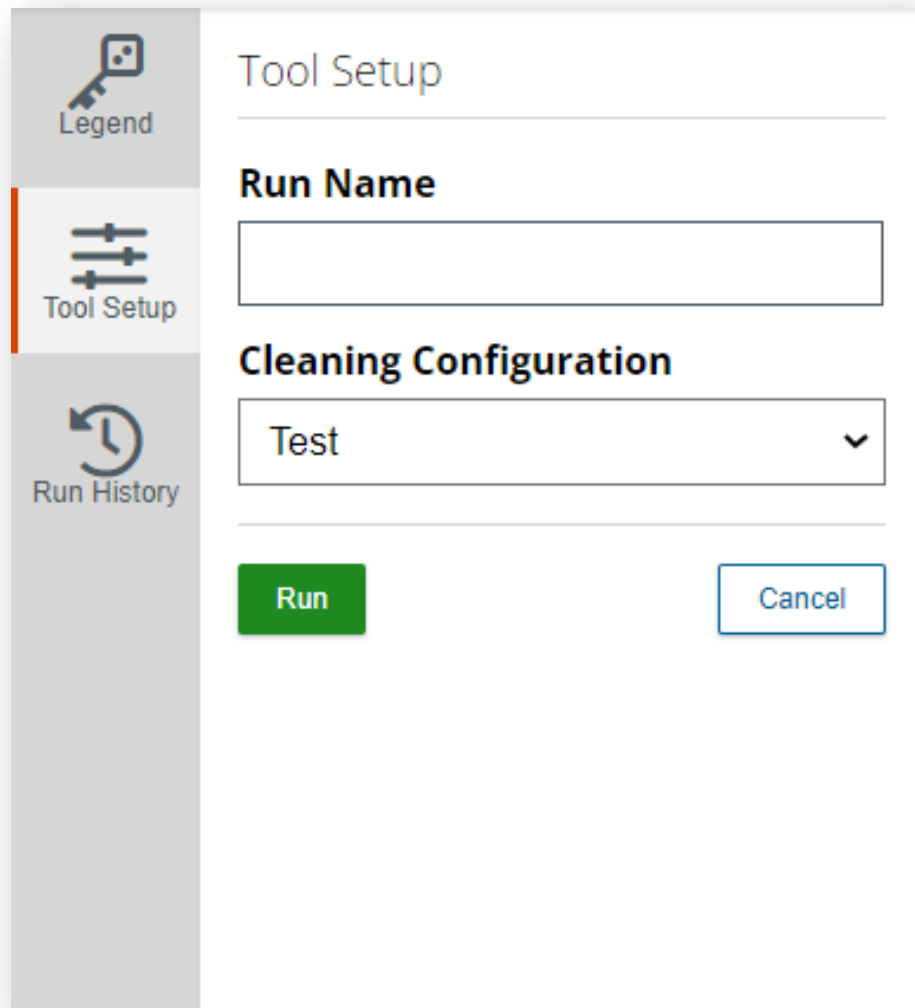
分析设置

可视化工具
实体分类

分析结果下载

数据详情

Sentiment Analysis情感分析



The screenshot shows a software interface for sentiment analysis. On the left is a vertical sidebar with three icons: a key for 'Legend', a gear for 'Tool Setup', and a clock for 'Run History'. The main area is titled 'Tool Setup' and contains a 'Run Name' text box, a 'Cleaning Configuration' dropdown menu set to 'Test', a green 'Run' button, and a blue 'Cancel' button.



情感分析

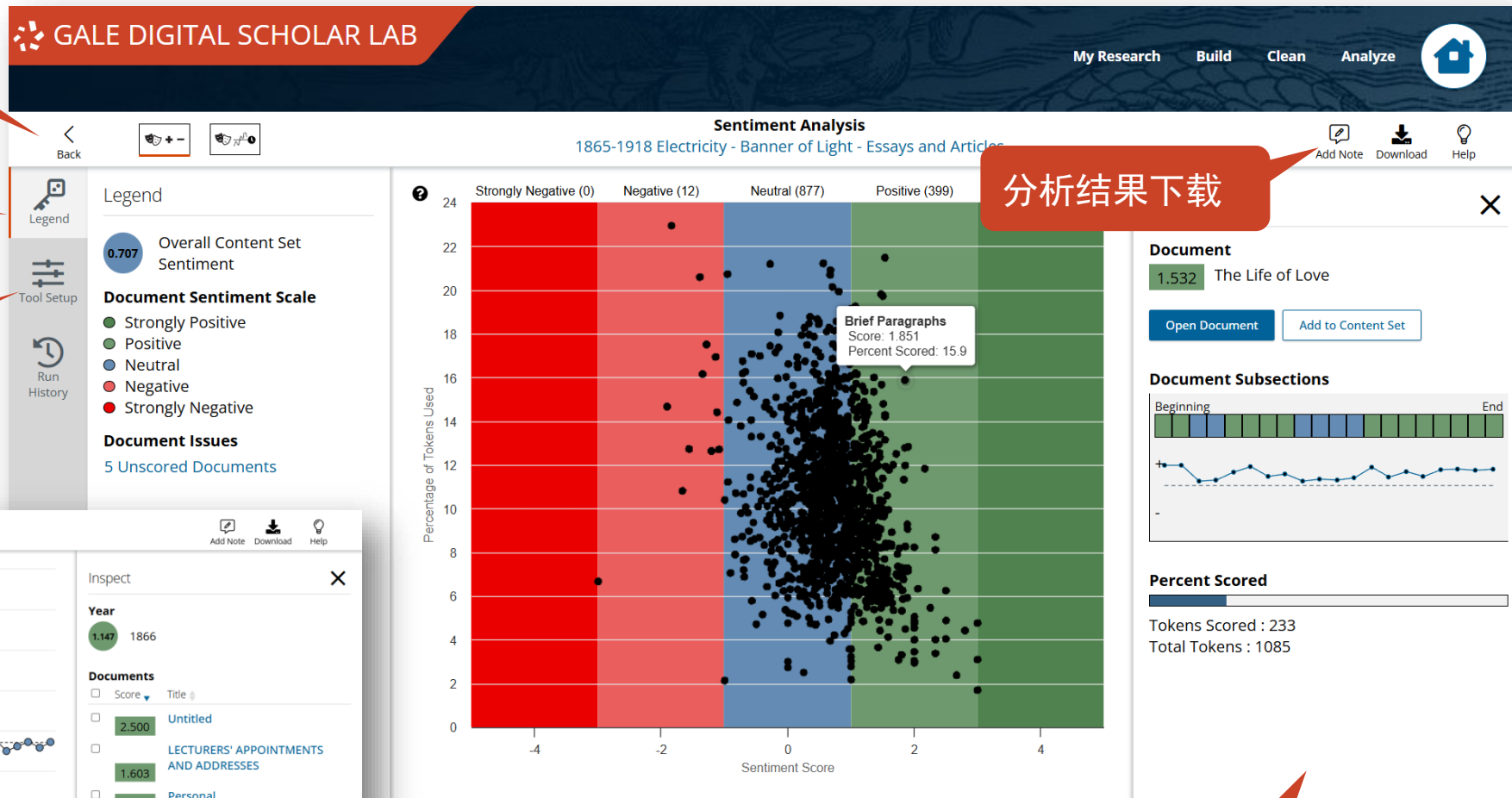
情感分析使用户能够直观地了解其内容集中的积极和消极情感意义。情感意义分析为用户创造了机会，让他们探索作者如何以及为何以积极或消极的方式与主题或事物相关。

可视化分析结果

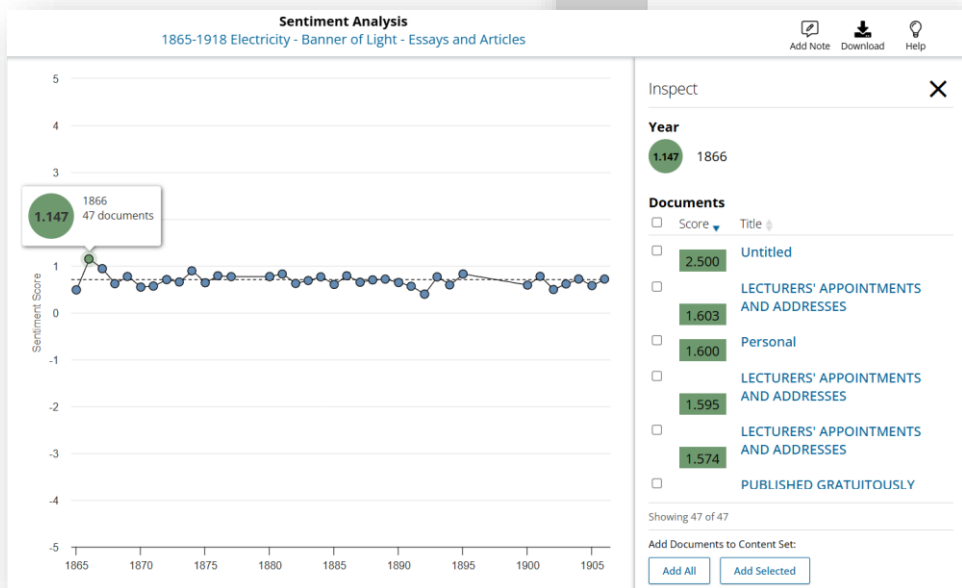
可视化切换

可视化工具

分析设置





分析结果下载



数据详情

Topic Modeling主题建模

 Tool Setup

 Run History

Run Name

Number of Topics ⓘ

Default: 10 Min: 1 Max: 99
Words per Topic ⓘ

Default: 10 Min: 1 Max: 99
Number of Iterations ⓘ

Default: 1000 Min: 1 Max: 9999
Cleaning Configuration



主题建模

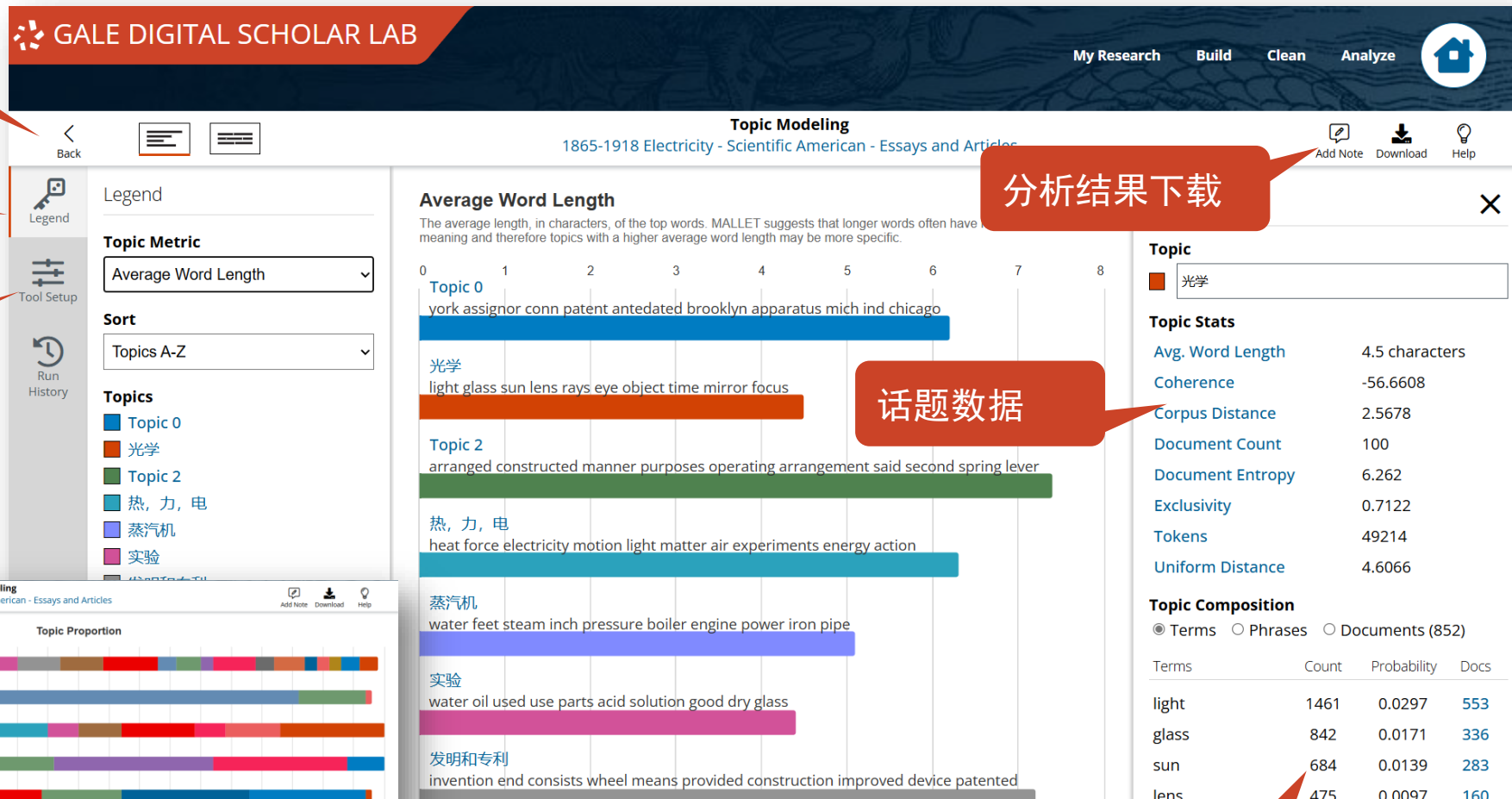
主题建模是一种文本挖掘方法，它允许用户通过分析 MALLETT 分配的高级主题来辨别其内容集中的情况，既它允许用户在内容集中的文档中发现统计上更有可能彼此靠近出现的词组。

可视化分析结果

可视化切换

可视化工具

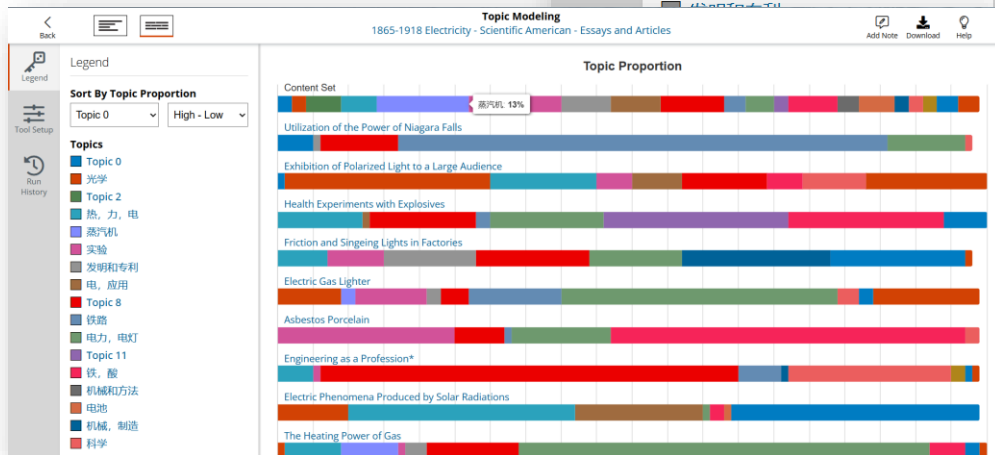
分析设置



分析结果下载

话题数据

数据详情



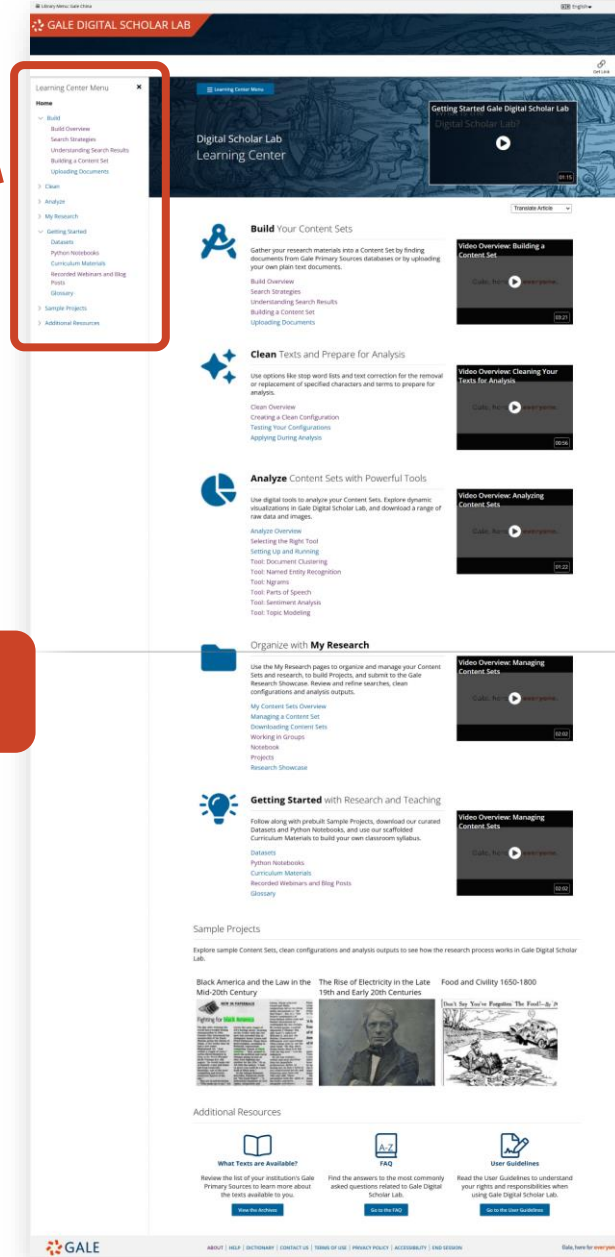
Learning Center 学习中心

Learning Center Menu

Home

- > Build
- > Clean
- > Analyze
- ✓ My Research
 - My Content Sets Overview
 - Managing a Content Set
 - Downloading Content Sets
 - Working in Groups
 - Notebook
 - Projects
 - Research Showcase
- ✓ Getting Started
 - Datasets
 - Python Notebooks
 - Curriculum Materials
 - Recorded Webinars and Blog Posts
 - Glossary
- ✓ Sample Projects
 - Black America and the Law in the Mid-20th Century
 - The Rise of Electricity in the Late 19th and Early 20th Centuries
 - Food and Civility 1650-1800
- > Additional Resources

学习中心目录



- 建立内容集
- 清理文本并准备分析
- 使用功能强大的工具分析内容集
- 整理我的研究
- 开始研究和教学
- 项目案例



更多信息欢迎访问：

www.gale.com

扫描二维码关注Gale官方微信

