Sample Project 2: The Rise of Electricity in the Late 19th and Early 20th Centuries

Overview

The harnessing and production of electricity is one of the defining watersheds of the late nineteenth century. This source of power opened numerous scientific and economic doors, while terrifying and amazing peoples and societies with its seemingly supernatural behavior and potential. The possibilities for the application of electricity were as undefined as they were limitless; electricity could power the new machines and technologies of the Industrial Revolution, but many wondered if it could also reshape and repower, or transform, the human body and self, too. As a substance, electricity shared many of the same qualities as light and spirit, which were two crucial paradigms for how human beings understood themselves and their place in the natural and spiritual worlds they inhabited. That electricity could be coursing through human bodies was a truly astounding idea. And so electricity itself is a topic that bridges science and spirituality, industry and invention, as well as fantasy and reality. After all, electricity played a role in Frankenstein as much as it did the lightbulb. This project compares how electricity features in two serial publications between the end of the Civil War in 1865 and the end of the First World War in 1918.

1. Ideate

What are the core research questions?

How do two serial publications treat the topic of electricity in the late nineteenth and early twentieth centuries?

How can the same topic be compared between two serial publications, or for that matter, two distinct content sets?

What are other more precise, relevant questions?

What themes are evident in the *Banner of Light* and *Scientific American* when discussing the topic of electricity?

Do the *Banner of Light* and *Scientific American* share any topics or similar points of view in their treatment of electricity in the late nineteenth century

What sentiments appear in discussions of electricity?

Are there any other distinguishing features surrounding electricity in these journals that may reflect on contemporary views of industrialization, invention, and their effects on men and women in the late nineteenth and early twentieth centuries?

2. Build

Steps

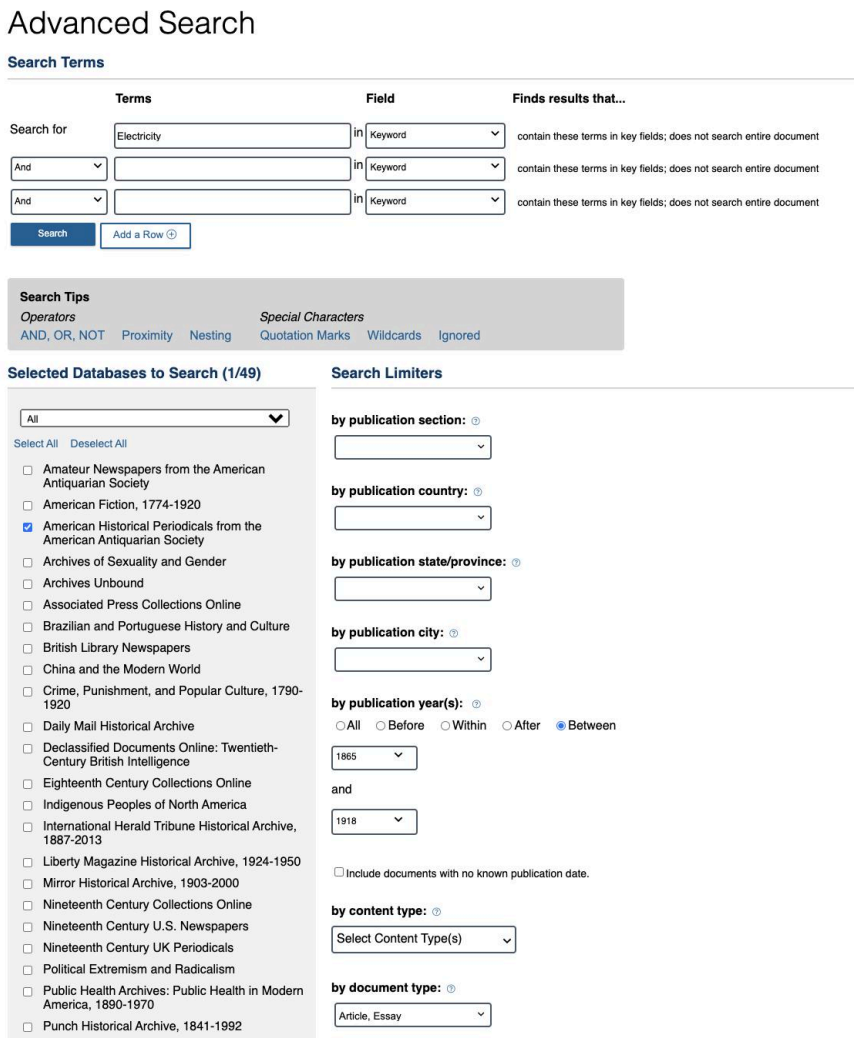
2.1 Searching

The initial content sets were constructed with the same variables in mind but distinct serial publications. The archive used for both was the American Historical Periodicals. The first

serial content set is derived from *Banner of Light* as the publication title, and the second *Scientific American*.

Banner of Light is one of many spiritualist journals/serials, running from 1857 to 1907.

Scientific American, a leading science journal, began publication in 1845 and continues to the present.



- Search Terms: Electricity
- Selected Databases to Search: American Historical Periodicals from the American Antiquarian Society
- Search Limiters - by publication year(s): Between 1865 - 1918
- Search Limiters - by publication title: *Banner of Light*
- Search Terms: Electricity
- Selected Databases to Search: American Historical Periodicals from the American Antiquarian Society
- Search Limiters - by publication year(s): Between 1865 - 1918

- Search Limiters - by publication title: *Scientific American*

Thinking about Methodology

Topic Modeling: This tool allows the researcher to see if there are any themes or topics that cut across a collection of texts.

Ngrams: This tool tracks different kinds of phrases or terms that might occur together, as well as the number of times a phrase appears

Sentiment Analysis: This tool examines whether the contents of the documents were positive or negative, overall, according to the AFINN dictionary.

2.2.Specific Tools

None of the tools required specific content sets. It was easier to create Cleaning Configurations that removed all punctuation, set all characters to lowercase, and also removed extended ASCII characters. Numbers were also removed.

2.3 Specific Questions

The question “What sentiments appear in discussions of electricity?” suggests a possible usage for the sentiment analysis tool. It is important to keep in mind, however, that the sentiment analysis tool provides a metric using the AFINN dictionary; it does not necessarily tell researchers what specific sentiments may be present within a text. In this respect, Topic Modeling and the Ngrams tool allow researchers to explore the kinds of phrases and words that could support and help contextualize and explain the results of the sentiment analysis tool. This is an example of understanding how the results of one tool can buttress or reinforce, or even help explain, the results of another.

3. Clean

The content sets were both cleaned for punctuation, as well as numbers and special characters as above. However, each required the creation of their own specific stop word lists. *Banner of Light* was published in Boston, Massachusetts, for most of its existence, and referenced other spiritualist centers in the United States. Since the research questions are

focused on electricity, rather than places, it made sense to include state abbreviations as well as common cities and place names such as Boston, Washington, and Philadelphia.

Also, since each content set is derived from a periodical with advertisements, columns, and sections, it made sense to remove common words associated with prices, publication sections, like pages, etc. These were far too common when running Topic Modeling and Ngrams; putting them on the stop word lists helped to clean up the "noise" in these results.

How did Clean Configurations change according to the research questions and analysis?

Topic Modeling

It seemed best to cast a wider net in part to see what kinds of words appeared in the models created by the MALLET software that powers the tool. Requesting more words than the default and double the topics produces finer grained topics in reflection of the size of the content set. The sample opted for 15 word topics and 20 topics.

Sentiment Analysis

This tool has no settings other than selection of the Cleaning Configuration.

Ngrams

As is the case with Topic Modeling, it seemed worthwhile to go beyond the default settings given the size of the content set. The threshold for the number of times an Ngrams had to appear to be considered useful was raised to 4. In order to find collocates rather than just single words, the minimum Ngrams size was set to 2 (biGram), and the maximum size to 5. These settings translate into a search for "Ngrams of between 2 to 5 words that appear in documents at least 4 or more times".

4. Analyze

Selecting particular views for each tool was extremely straightforward. The size of the content sets suggested an approach in which the search maximized the number of results, which could then be further refined. Topic Modeling as a tool

statistically discerns what words are more likely to appear near to one another. More topics and more words lower the threshold of what is significant, meaning the result is a finer grained picture of what the statistical analysis could suggest. In very similar documents, as would be the case for those appearing in serial publications such as *Banner of Light* and *Scientific American*, it is likely that there will be similar words related to advertisements, questions posed by readers, comments and notices, etc. Selecting more words and more topics is a good way to sift through some of these known similarities, and can work in tandem with stop word lists to help drill down into a large content set. A similar approach to thinking about potential "noise" was taken with the Ngrams in order to maximize results that could then be further refined to produce more meaningful results. When it comes to selecting views and Ngrams, the highest count in a result does not always translate into the most meaningful or interesting. There must be a balance between number and noise.

Understanding Results

This project involved numerous iterations of Cleaning Configurations to obtain initial clear results.

Topic Modeling

There was some clear overlap, as expected, between the two periodicals when it came to Topic Modeling, but also distinct topics and concerns. The topics in *Banner of Light* focused more on the self and the body, and how electricity fit into human existence—not surprisingly, given the spiritualist nature of the publication. *Scientific American* shared some of these concerns but was also very interested in the use of electricity as an invention or means of bettering society. Where the two seemed to overlap is around the idea of betterment, often seen somewhat with words revolving around health, medicine, and discovery.

Banner of Light

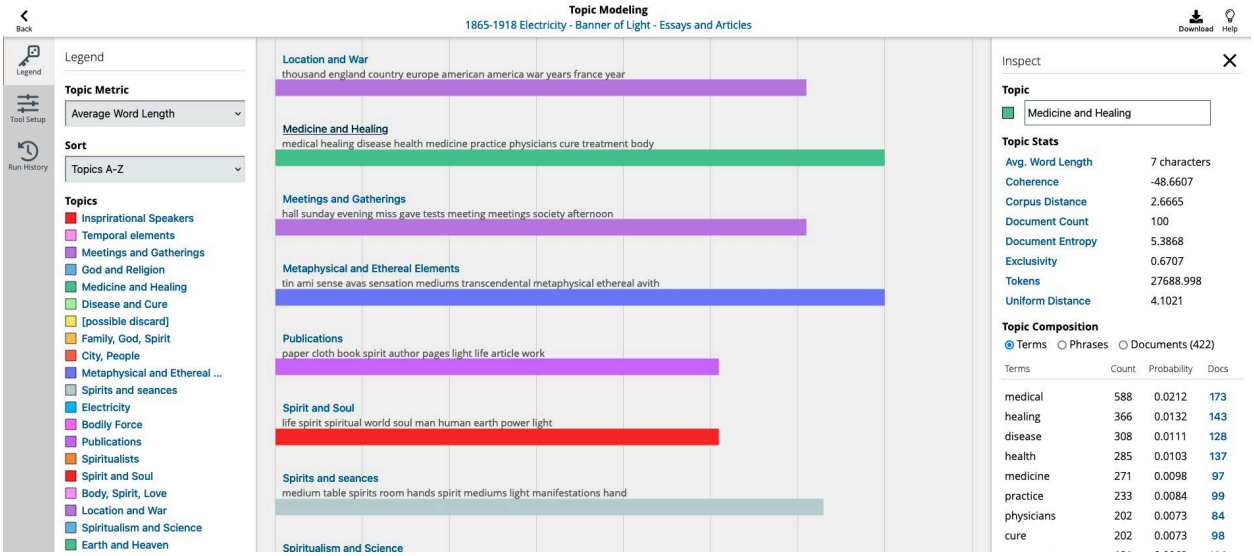
- Life, spirit, man, spiritual, human, mind, nature, power, thought, soul
- powders, positive, diseases, cure, office, cured, negative, sent, disease, healing
- medium, table, room, hand, said, hands, came, spirits, spirit, saw
- cloth, light, spiritual, paper, book, banner, free, place, rich, white
- force, matter, form, motion, light, forms, atoms, heat, substance, forces
- electricity, electric, life, brain, blood, current, water, body, electrical, air
- spiritualism, phenomena, science, facts, scientific, spirits, subject, fact, truth, spiritual
- medical, healing, medicine, disease, health, practice, physicians, law, treatment, physician
- spirit, spirits, spiritual, earth, medium, life, body, power, form, conditions

Scientific American

- water, steam, inch, boiler, power, engine, pressure, iron, pipe, use
- apparatus, prof, valuable, steam, contained, description, iron, method, electric, interesting
- lightning, electricity, earth, animal, plants, rain, death, ground, electric, animals
- electric, power, light, motor, electricity, car, horse, lamps, engine, lighting
- science, professor, scientific, prof, society, electrical, electricity, american, discovery, year
- light, force, motion, heat, matter, electricity, energy, sun, theory, rays
- current, wire, electricity, electric, battery, magnet, iron, machine, wires, placed
- telegraph, telephone, bell, instrument, wire, line, patent, cable, apparatus, wires
- battery, wire, use, power, cells, writes, motor, current, coil, used

Researchers have the option of “naming” their Topics in the results, as was done for this project. It is crucial to note that researchers have to come up with their own interpretations of what the lists above might represent as a coherent “topic”, rather than as a statistically

created list of words. Doing so creates a better sense of what the Topic Modeling comparison view can do to help the researcher understand specific metrics and the topics created by the tool.

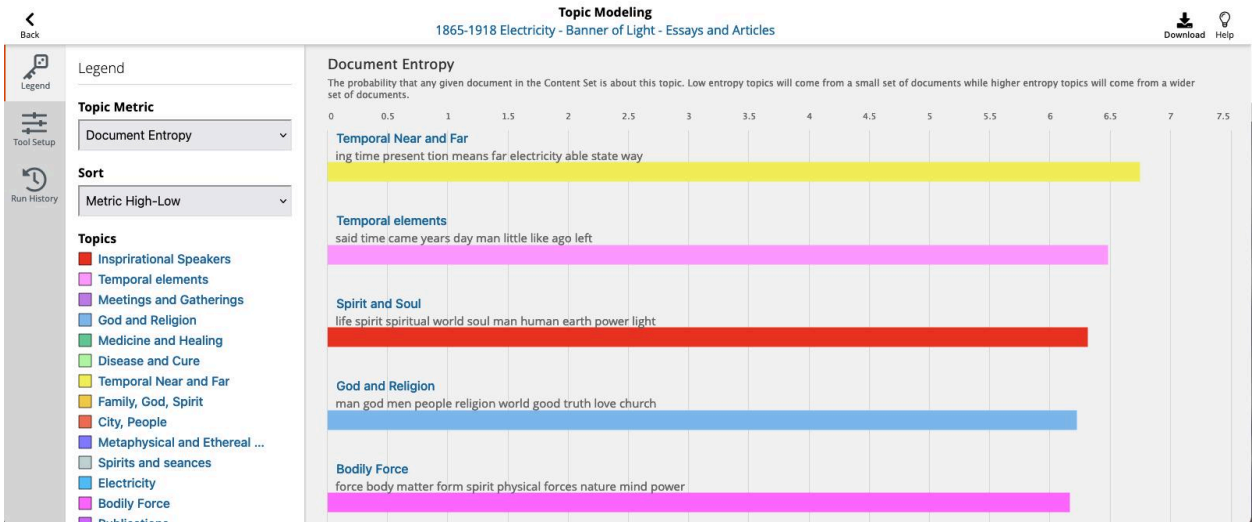


Before moving to the topic comparison view, it is worth looking at a single topic. Clicking on a topic in the “Legend” on the left of the screen, will open the “Inspect” panel. This includes a summary of the Topic Stats, and the Topic Composition, which can be explored by terms, phrases or documents.

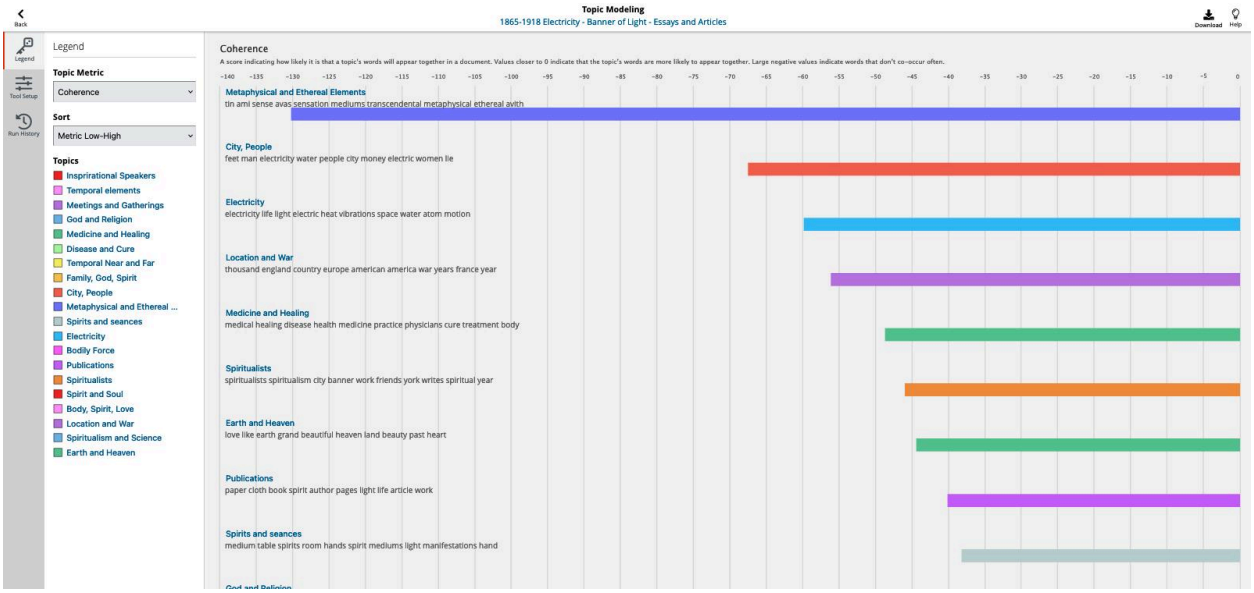
The Topic Modeling comparison view provides specific metrics that can help the researcher work through the various subjects the content sets contain. While the topic will contain terms which pertain directly to the topic being studied, the documents will often also discuss topics that have nothing to do with electricity.

Banner of Light

Document Entropy offers a way of thinking about the prevalence of topics within a content set as indicated by how extensively topics appear across the entire set. In the example, the named topic “Temporal Near and Far” is the most prevalent. The next is “Temporal Elements”, then “Spirit and Soul” and “God and Religion”. The results suggest that electricity really does not appear as a distinct topic across the entire content set. Despite its presence in various topics, it is not pervasive.

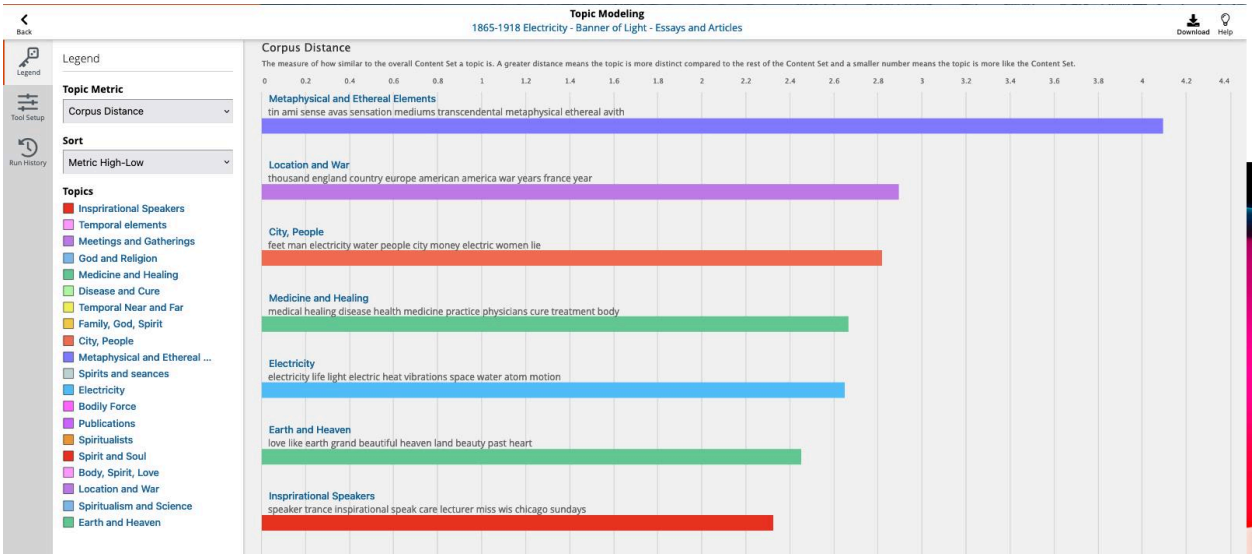


Coherence provides a metric that suggests how closely related the words in the topic are within the texts. Words closer to O in the barchart indicate that they are more likely to appear together. Since a topic is composed of words that have a greater statistical chance of appearing near each other, this shows how great that proximity actually is for a topic. Notably out of all the topics, “metaphysics and ethereal elements” has the lowest coherence, indicating that the words that the tool suggests as a topic are in fact spread out further from one another in contrast to those that make up the topics “disease and cure” or “spirit and soul”.

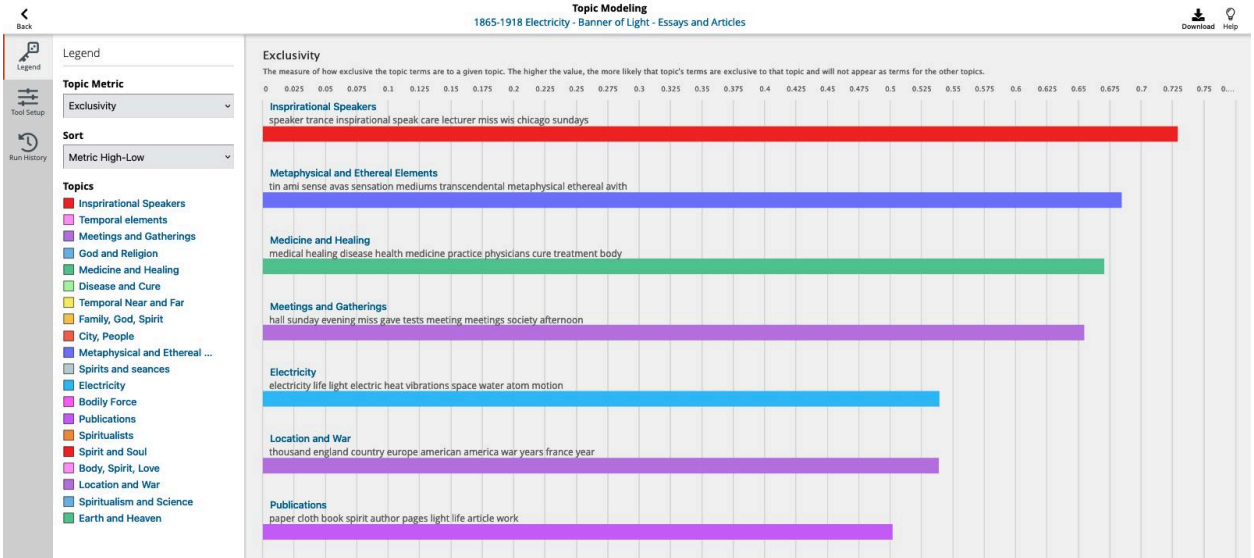


Uniform distance essentially compares a topic to the ways in which words are distributed within texts. Conceptually, it is related to the coherence metric, but instead of comparing the distance of words within a topic, it compares those to how all words are distributed, and in relation to the topic words themselves. This metric helps discern how specific a topic might be; in this case “inspirational speakers” is the highest, with “spirit and soul” coming in second.

Corpus distance offers a metric for ascertaining how unique the words that make up a topic might be within a content set. The frequency measure shows how distinct words within a topic are from the entire content set. In this case “metaphysical and ethereal elements” comes in a pronounced first, suggesting that this topic really is quite distinct from the rest of the language within the content set.

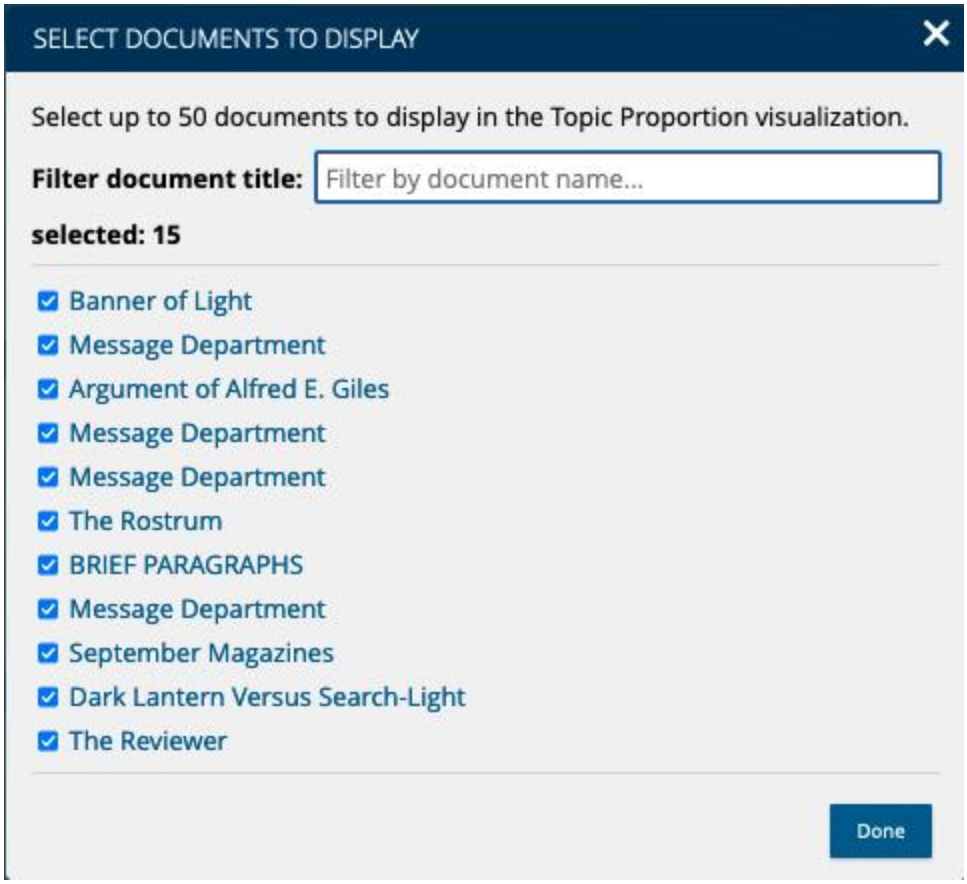


Exclusivity demonstrates another way of thinking about the uniqueness or distinctiveness of a topic in relation to the broader content set. Since each topic is made up of a group of statistically proximate words, there is a good chance those words overlap with one another. The degree to which they do notâ€”i.e., that they are distinct to a specific topicâ€”suggests that the vocabulary that composes a topic is itself limited to that topic, making it more “exclusive” rather than woven into and possibly appearing within other topics. For example, “trance” appears again in the “inspirational speakers” content set, with “metaphysical and ethereal elements” coming second, and “medicine and healing” third.

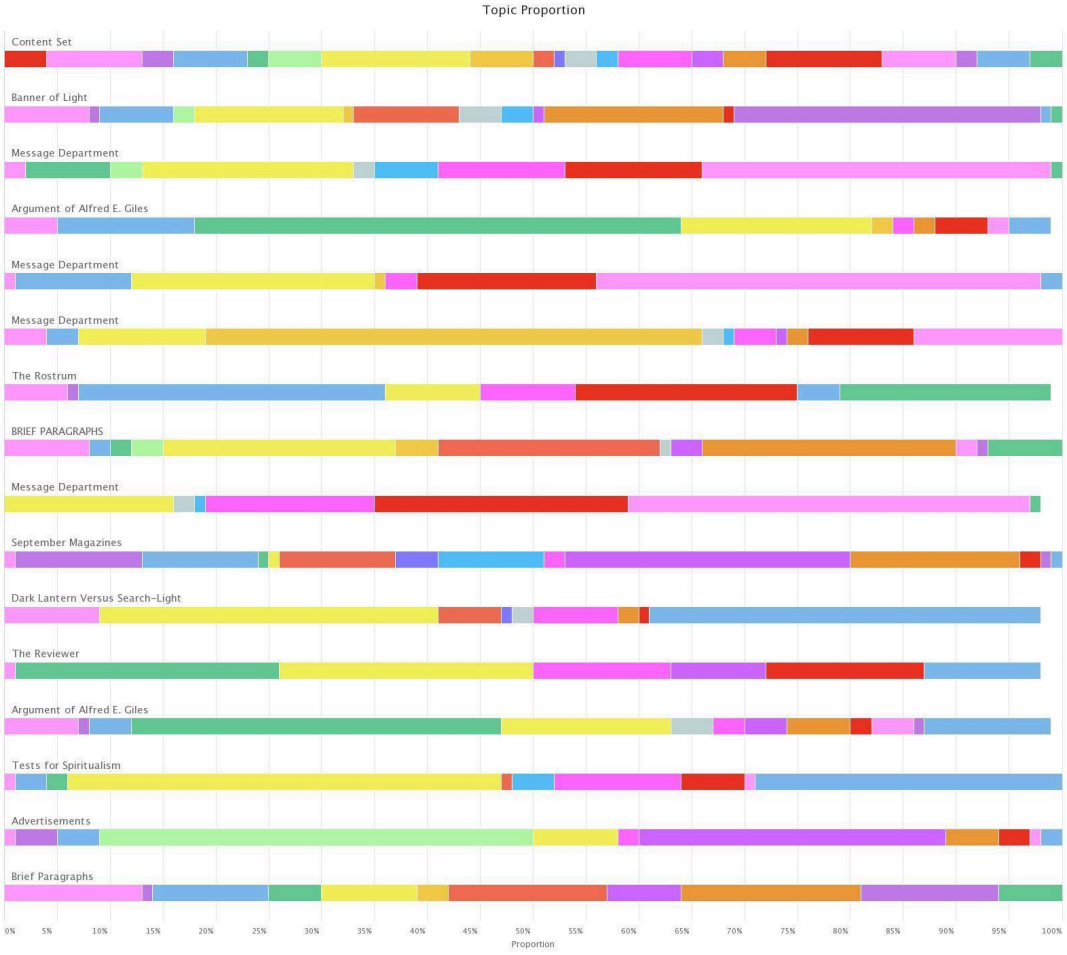


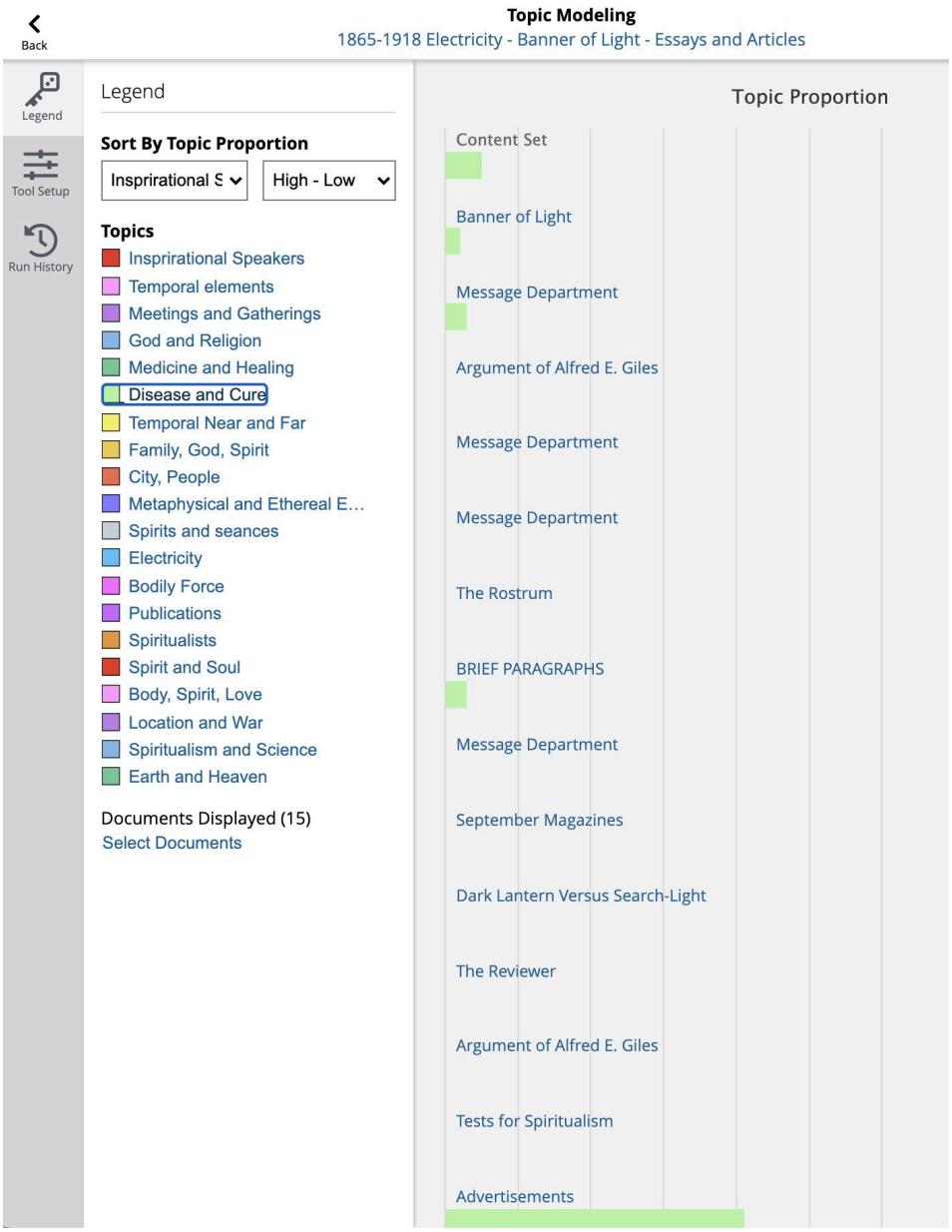
Configuring the Topic Proportion

The Topic Proportion view of the Topic Modeling tool permits the researcher to compare how much of the texts within a subset of a content set are allocated to each topic. This visualization provides a quick means of seeing how prevalent topics are against one another, rather than always relying on numbers. The visualization is limited to 50 documents, which are initially randomly selected from the content set. Researchers do, however, have the option of selecting their own documents and refreshing the visualization.



As is the case with the Topic Comparisons, it is recommended that the researcher create names for the topics when using this view. Clicking on a specific topic will cause the visualization to display only the percentage of each text where the topic appears.


Topic Proportion
Content Set
Banner of Light
Message Department
Argument of Alfred E. Giles
Message Department
Message Department
The Rostrum
BRIEF PARAGRAPHS
Message Department
September Magazines
Dark Lantern Versus Search-Light
The Reviewer
Argument of Alfred E. Giles
Tests for Spiritualism
Advertisements
Brief Paragraphs
0%
5%
10%
15%
20%
25%
30%
35%
40%
45%
50%
55%
60%
65%
70%
75%
80%
85%
90%
95%
100%
Proportion



Scientific American

As part of this overview of Topic Modeling, here are some highlights from the *Scientific American* results to compare against the *Banner of Light* results. There are not any obvious overlaps.

The uniform distance measure does indicate that the most specific topics associated with electricity are “Electrical Communications”, “Batteries and Portable Power”, “Electrical Current”, and “Electricity as Power”.

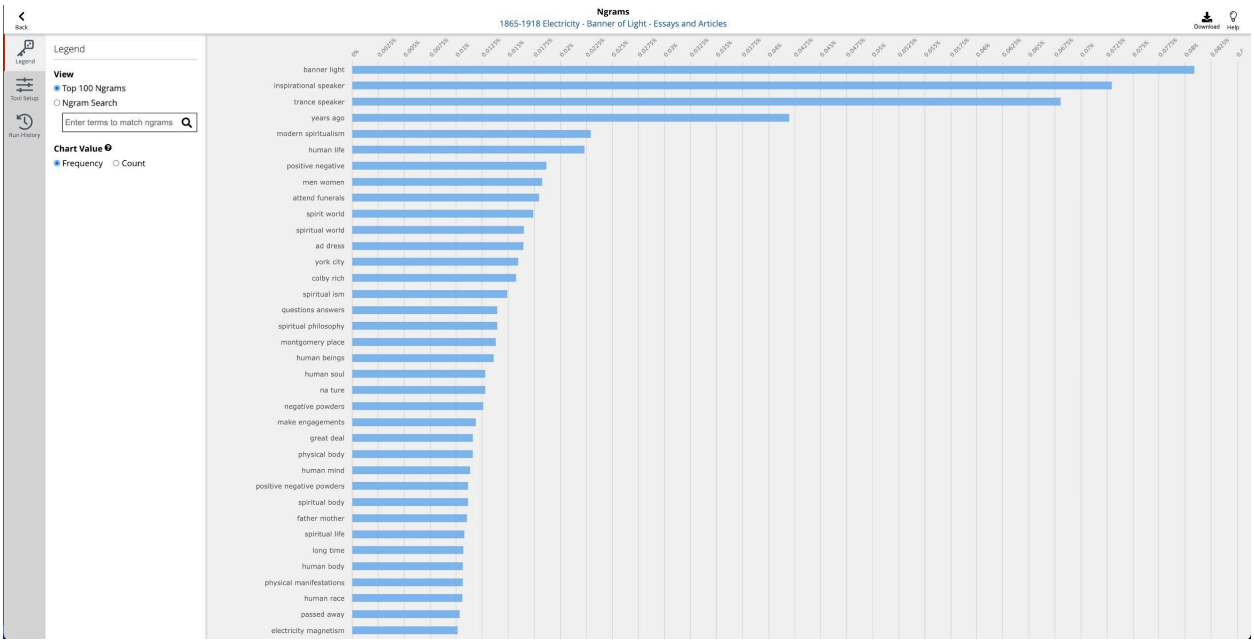
Exclusivity seems somewhat inconclusive as well. In the end, the measures for these topics indicate a wide array of themes, and though there are some clearly exclusive topics, such as Topic 6 involving guns and boats, there is also a healthy mix of topics in this content set.

Ngrams

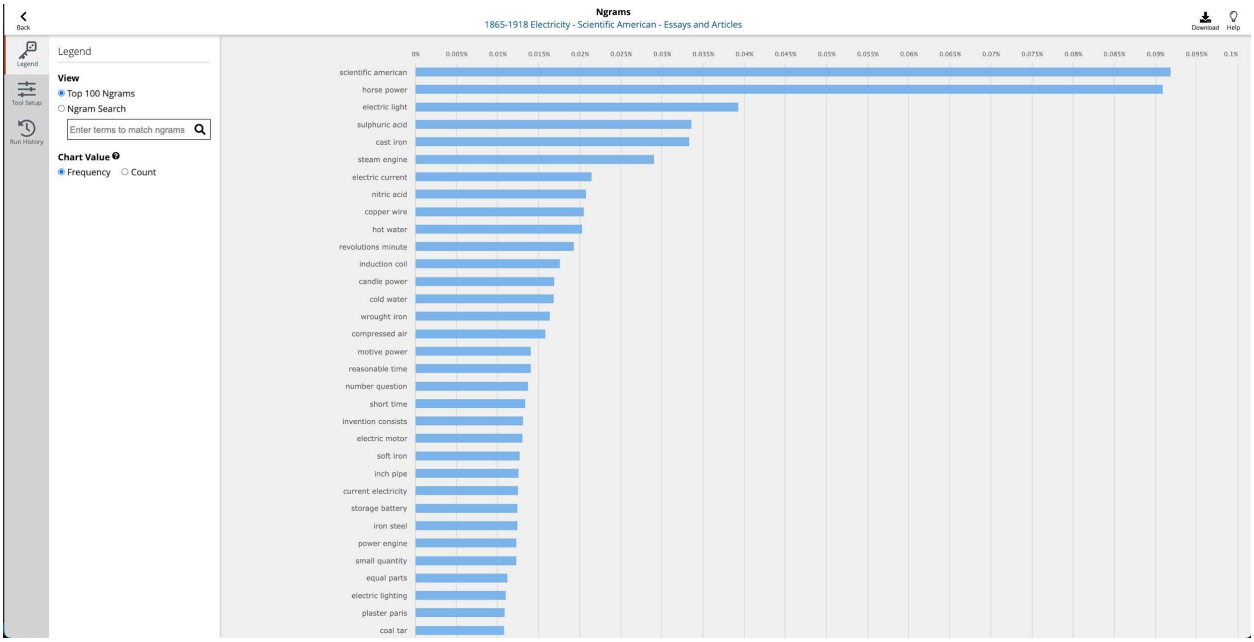
Configuration: Min 2 Max 5, Threshold 5

Cleaning Configuration: Electricity - Banner or Scientific American, no Punctuation, No Numbers, No Special Characters

Banner of Light



Scientific American



The most interesting thing about the Ngrams results is not merely that electricity is not very prevalent, but that in the *Banner of Light* results the word that appears with electricity is magnetism, not something more obviously to do with spirituality as might be expected. Equally, in *Scientific American*, the word “light” appears before “current”; “light”, as can be seen above, is an important word in the *Banner of Light* topics. But it is also one of the most important scientific inventions involving electricity: the light bulb. With the *Banner of Light* Ngrams results, there are finally some hints of overlap between the two periodicals, and a new research question: what is the link between electricity and magnetism across the two publications?

5. Interpret

Read more about ways you can expand this project with iteration, research questions, and analysis.

Research Outcomes

The results are inconclusive, but there are some commonalities. For example, the low appearance of “electricity” in the Ngrams results seems to indicate that more refinement of the content sets are needed.

Original Questions

1. What themes are evident in the *Banner of Light* and *Scientific American* when discussing the topic of electricity? It is clear that the two publications have fairly distinct interests. Where the *Banner of Light* mentions or treats topics related to electricity, it has to do with the body as well as spiritual “force”. In contrast, *Scientific American* is very much concerned with the practical application of discoveries to new inventions.

2. Do the *Banner of Light* and *Scientific American* share any topics or similar points of view in their treatment of electricity in the late nineteenth century? At first glance, it does not appear so. However, familiarity with the period, and the history of science and religion, suggests there are some possible similarities. This is where results of text analysis really require deeper knowledge of a particular field of research. The idea of spiritual “force”, and the Ngrams result in *Banner of Light* for “electricity magnetism”, relate directly to the connection between current and electricity as a force that was also a way of describing the soul and the spirit. This is evident in the topics found in the *Banner of Light* results:

    - force, matter, form, motion, light, forms, atoms, heat, substance, forces
    - electricity, electric, life, brain, blood, current, water, body, electrical, air

3. Are there any other distinguishing features surrounding electricity in these journals that may reflect on contemporary views of industrialization, invention, and their effects on men and women in the late nineteenth and early twentieth centuries? Not that can be seen at the moment.

New Questions

1. Question 2 above presents a possible new line of inquiry on electricity and spirit, health, body, or force. This will require more precise content sets to explore.

2. Perhaps the most fascinating discovery was an article in the *Banner of Light* discussing *Scientific American*. This document could act as the foundation for an entire study on how the two periodicals reflect overlapping and maybe competing concerns of the era. Electricity clearly appears in the discussion since the document is part of the content set.

Thinking Critically About Research

Content Set Building

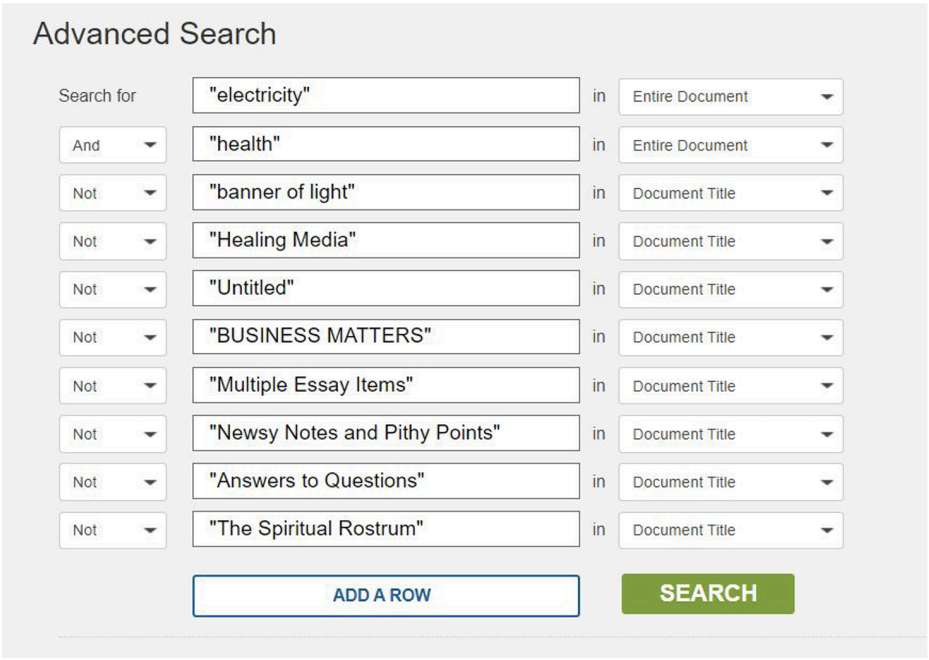
The content-set-building for this comparative project consisted of finding appropriate temporal boundaries for two serial publications. The date range was defined using two significant events in history: the end of the Civil War in 1865 for the start date and the end of the First World War in 1918 as the end date. In between these two conflicts, the place of electricity within American society moved from a fairly limited notion into practical applications through industrial and scientific development. At the same time, electricity was a topic of intense cultural fascination and discussion. Both of these concerns are evident in the serial publications selected for this project. Building the content sets, as a result, only varied in the selection of the periodicals themselves. Everything else remained the same.

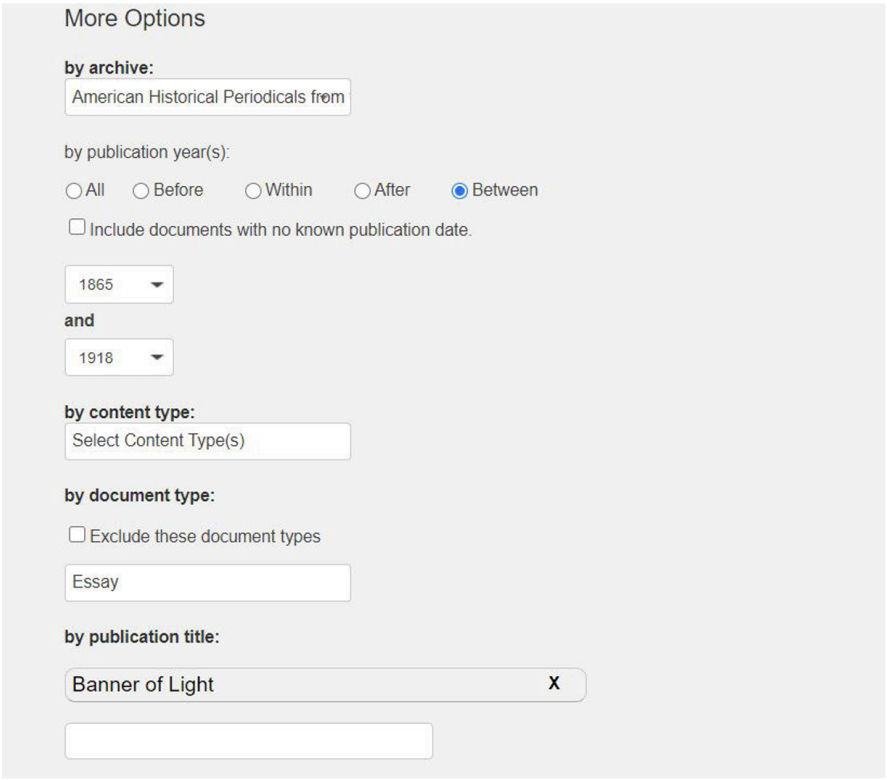
Iteration

As easy as it was to build the initial content sets, each required creation of distinct Cleaning Configurations, as the Ngram and Topic Modeling tools revealed new words that obscured meaningful results. Both content sets needed their own stop word lists in order to remove the ”noise,” or words arising from serial publications and advertisements, as well as unuseful information, such as place names for *Banner of Light* and scientific experimentation words for *Scientific American*.

Both content sets were also muddied by the presence of documents that often appear in serial publications, such as advertisements, notes and queries, letters from readers, and

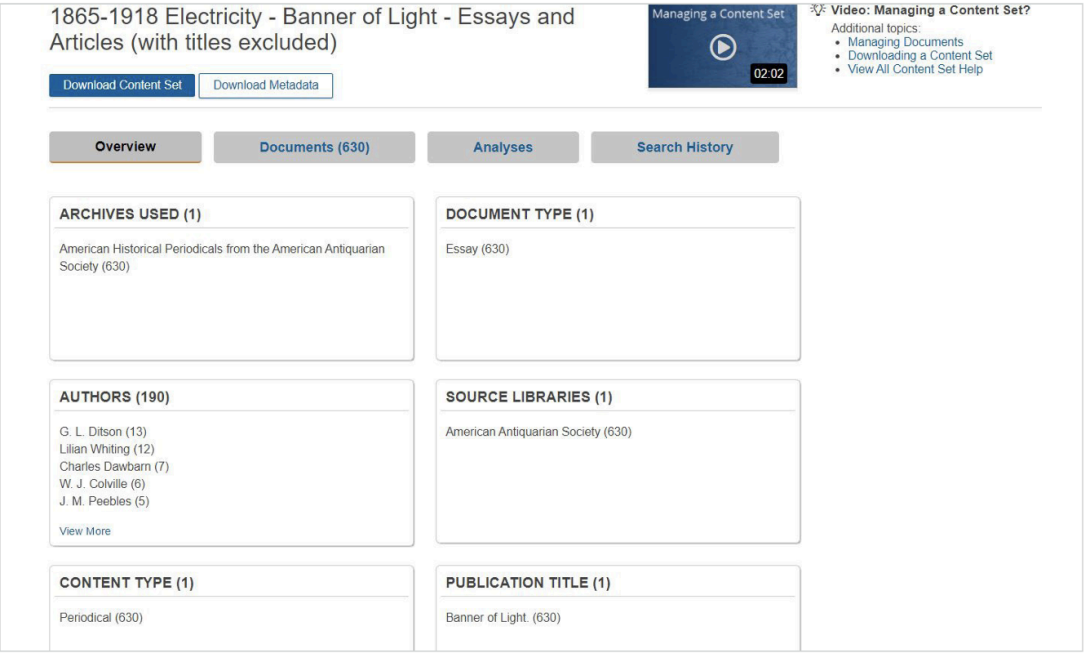
set or repeating editorial sections. The easiest method of scrolling through the titles of the documents to find repeating titles (and thus standard sections of the publications) was to browse the documents in Topic Proportions and make a list. Then the original search parameters were revised (see Search History) and the titles added as rows to the advanced fields with the “not” selection. Here’s an example search for Banner of Light with these repeated document titles excluded:





It’s worth comparing the revised content sets against the original sets to identify any opportunities for further clarification to the original research questions.

Banner of Light – Essays and Articles (with Titles excluded)



Scientific American - Essays and Articles (with titles excluded)

Understanding Outcomes

Revising Questions

As discussed in the guide, it is not only normal to revise the research questions after running analysis tools on a content set, it's an integral part of the research process. Often, analysis will turn up new questions that could lay beyond the scope of the current project. This is how researchers develop new projects and lines of scholarship: by following clues and new questions that come up while pursuing other research.

Limitations

• This project didn't build content sets using actual cases, nor were they built following a close reading of the documents included in the sets. A more precise content set could be built by examining each document to determine whether or not it was appropriate to include in a content set focused on the specific parameters of the project.

• Iteration isn't restricted to cleaning; it is a key part of content-set-building as well. It became clear in this project that the recurring sections of serial publications or periodicals can add considerable noise to a content set. Excluding documents with titles that repeat can substantially change outcomes.

Presentations

All of the tool outputs can be downloaded as images to use in PowerPoints or embedded in web pages or other ways to present one's work.

New Visualizations

It is also possible to download the data that power the visualizations as comma-separated values (CSV) or JavaScript Object Notation (JSON) files, allowing for the creation and formatting of the researcher's own visualizations. With the proper skills, it's possible to collate or create new visualizations that may combine outputs from similar visualizations into one, allowing the researcher to compare and contrast in new ways not available in the Gale Digital Scholar Lab tool. The Topic Modeling tool downloads are especially rich with possibilities for new visualization. The Topic View download is large and contains results for each document and measure for the tool—much more data than the Topic Model visualizations can currently display. Programmers can treat this as the ideal place to start exploring the data created by the Lab using other tools and visualization designs.

The development of electricity, and the comparative nature of this project, could be greatly extended beyond the Lab by plotting downloadable data along timelines. The first approach might involve breaking out the scores for topics by publication date and plotting them along the time scale, as in the Topic Modeling Martha Ballard's Diary project.

This could also be accomplished by considering important moments in the development of electricity itself using resources like the Electricity Timeline.

Refining the Content Sets

Understanding the limitations of Gale Digital Scholar Lab helps inform both the methodology used to build content sets and ways to use the results produced by its tools. As precise as the revised content sets might be, the many words surrounding

electricity—not only permutations of it such as “electric,” “electrical” but also terms such as “current,” “force,” “spark,” “power,” “motion,” “spirit”—suggest that using the Topic Modeling tool might offer a rich method of building more precise content sets when a closer reading isn’t possible.

Downloading the Content Sets

As powerful as Gale Digital Scholar Lab’s tools are, they offer fairly standardized configurations and are not currently customizable. Researchers may opt to download their content sets in order to use other tools and customize the editing and cleaning of the documents. The Lab’s Cleaning Configurations aren’t as powerful as more extensive, iterative techniques that make several passes over documents to refine and clear up problems arising from optical character recognition digitization. Downloading also allows the researcher to find problem words and tokens that can complicate or mess up results in the Lab.