WEEK 6b Qualitative or Quantitative? Considering which tool to choose through the lens of sample projects

---

Sample Project: The Rise of Electricity in the late 19th & Early 20th Centuries

Synopsis

The harnessing and production of electricity is one of the defining watersheds of the late 19th century. This source of power, which acts and looks unlike anything other than the supernatural, opened numerous scientific and economic doors, while terrifying and amazing peoples and societies with its potential. What electricity might do was unclear - it could power the new machines and technologies of the industrial revolution, but could it also reshape and repower, or transform, the human body and self, too? As a substance electricity shared much with 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.

- Banner of Light one of many spiritualist journals / serials - runs from 1857 to 1907
- Scientific American, a leading science journal runs from 1845 to the present

Core Research Question:

- How do two serial publications treat the topic of Electricity in the late 19th & early 20th century?
- How can the same topic be compared between two serial publications, or for that matter, two distinct content sets?

More Precise Questions:

1.What themes are evident in the Banner of Light and Scientific American when discussing the topic of ‘electricity’?

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 19th century?

3.What sentiments appear in discussions of electricity?

4.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 19th and early 20th centuries?

Thinking about Methodology & Specific Tools

- Topic Modeling - we can use this tool to see if there are any themes or topics which cut across a collection of texts
- nGram - we can use this tool to track different kinds of phrases or terms which might occur together, and the number of times a phrase appears
- Sentiment Analysis - we can use this tool to examine whether the contents of the document were overall positive or negative according to the AFINN dictionary.

Building the Content Set

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’

Limits & Parameters

Content Type ("Article" or “Essay”)

Archive - American Historical Periodicals

Publication Date (1865-1918)

Keywords (in individual rows)

Entire Document: Electricity

Statistics & Info

Banner of Light

Content Set Name: 1865-1918 Banner of Light - Essays and Articles

Content Set ID: 1580766975821

Number of Documents: 1296

Scientific American

Content Set Name: 1865-1918 Scientific American - Essays and Articles

Content Set ID: 1580767055855

Number of Documents: 3356

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.

Specific Questions

Question 3 suggests that the sentiment analysis tool might be an option. But the sentiment analysis tool provides a metric using the AFINN dictionary - it doesn't necessarily tell researchers what specific sentiments may be present within a text. In this respect, Topic Modeling and the nGram tool allow us to explore the kinds of phrases and words which 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.

Cleaning the Content Set

Specific research questions:

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. The Banner of Light was published in Boston, Mass. for most of its existence, and referenced other spiritualist centers in the United States. Since our 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, Mass., Washington, Philadelphia, etc. 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.

Running Tools

Selecting particular views for each tool was extremely straightforward. We selected an approach that reflected the size of our content sets: we looked for more things and raised the bar for what made the cut for the results. Both were for very simple reasons: Topic Modeling as a tool statistically discerns what words are more likely to appear near to one another. More topics, and more words lowers the threshold of what is 'significant', meaning we get a finer grained picture of what the statistical analysis could suggest. In very similar documents, like those appearing in serial publications like Banner of Light and Scientific American, the chance is 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 of sifting through some of these 'known' similarities, and can work in tandem with stop word lists to help 'drill down' into a large content set. For nGrams, we took a similar approach to thinking about potential 'noise' - we want to see what turns up. But the highest count in a result doesn't always translate into the most meaningful or interesting. There's a balance between number and noise.

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. We opted for 15 word topics, and 20 topics.

Sentiment Analysis

This tool has no settings other than selection of the cleaning configuration.

nGrams

Like Topic Modeling, it seemed worthwhile to go beyond the default settings given the size of our content set. We raised the threshold for the number of times an nGram had to appear to be considered useful, and set it at 4. Equally, we wanted to find collocates rather than just single words, so we set the minimum nGram size 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”.

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. Banner of Light’s topics 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

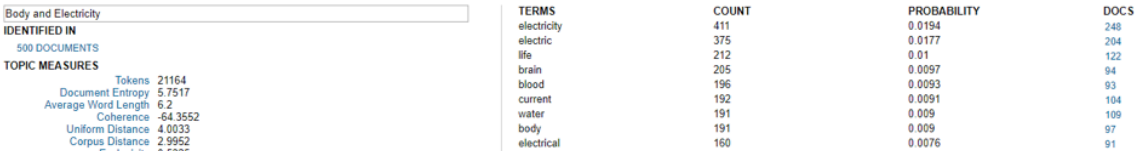
- spiritualism,phenomena,science,facts,scientific,spirits,subject,fact,truth,spiritual
- medical,healing,medicine,disease,health,practice,physicians,law,treatment,physician
- spirit,spirits,spiritual,earth,medium,life,body,power,form,conditions

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

For this project, we 'named' our Topics in the results. This is purely optional, and it's 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. But once this is done, we have a better sense of what the Topic Modeling comparison view can do to help us understand specific metrics and the topics created by the tool.

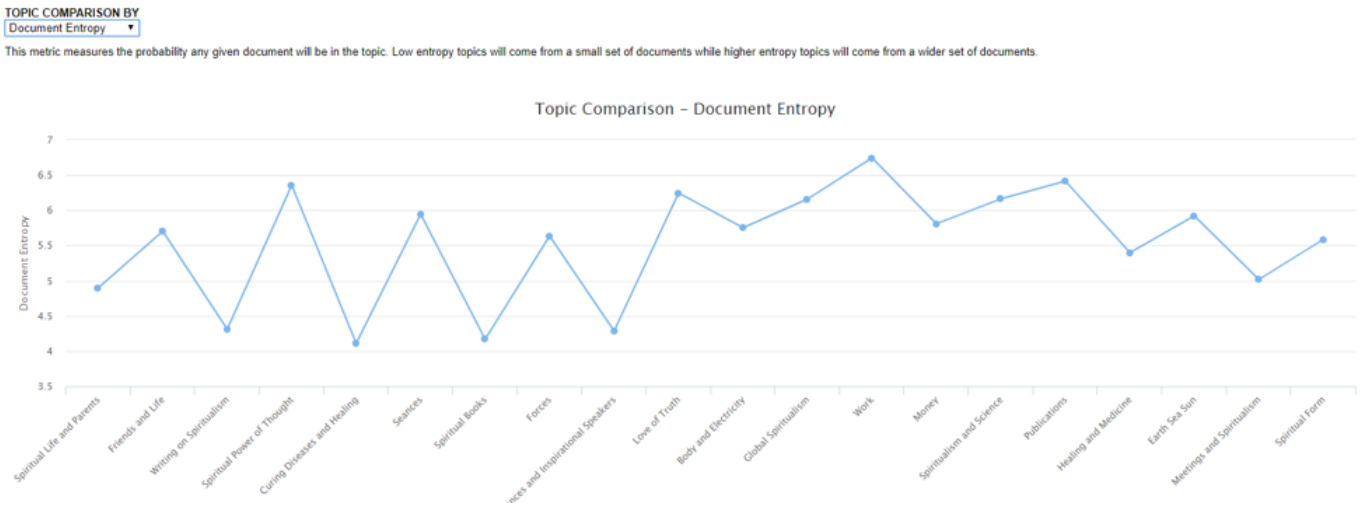
Before moving to the topic comparison view, it's worth looking at a single topic. Here we see on the left a summary of the topic measures, and how many documents where the tool has identified the topic. On the right, the four columns each contain information on the number of words you've selected for your topics, and the number of documents in which the word appears.



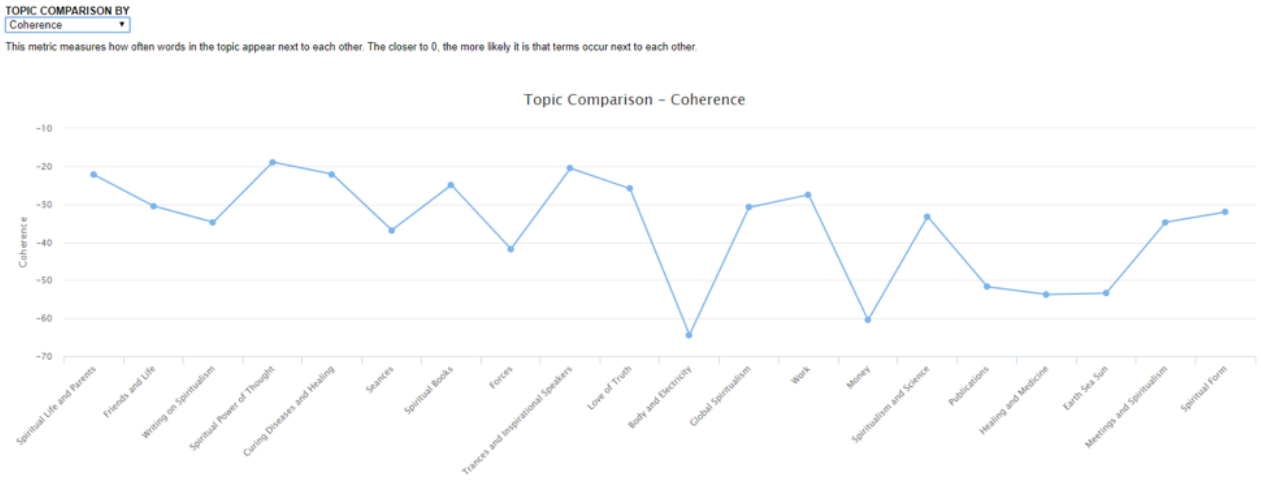
Examining the Topic Modeling comparison view we find specific metrics which can help us work through what kinds of subjects our Content Sets contain. The most important thing - unsurprisingly - is that the documents also discuss topics that have nothing to do with Electricity. That said, electricity does appear alongside other related words such as ‘force’, ‘energy’, ‘light’,’spirit’, etc.

Banner of Light

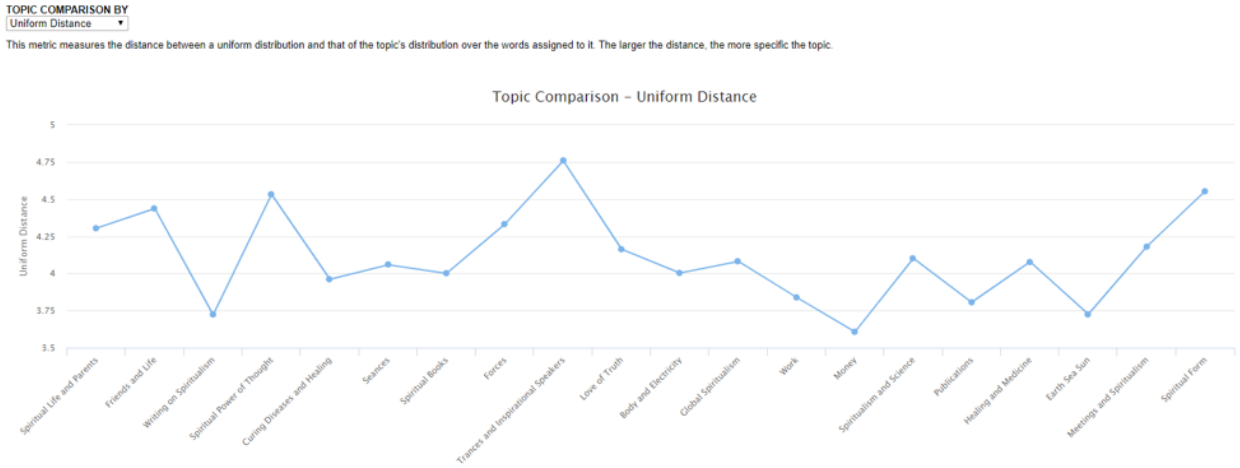
Document Entropy offers a way of thinking about the prevalence of topics within a content set as indicates how extensive a topic is across all the entire set. In the example above, the topic we’ve named as ‘work’, is the most prevalent. The next is ‘publications’, and the third ‘spiritual power of thought’. The results suggest that electricity really doesn’t appear as a distinct topic across the entire content set. Despite its presence in various topics, it’s not pervasive.



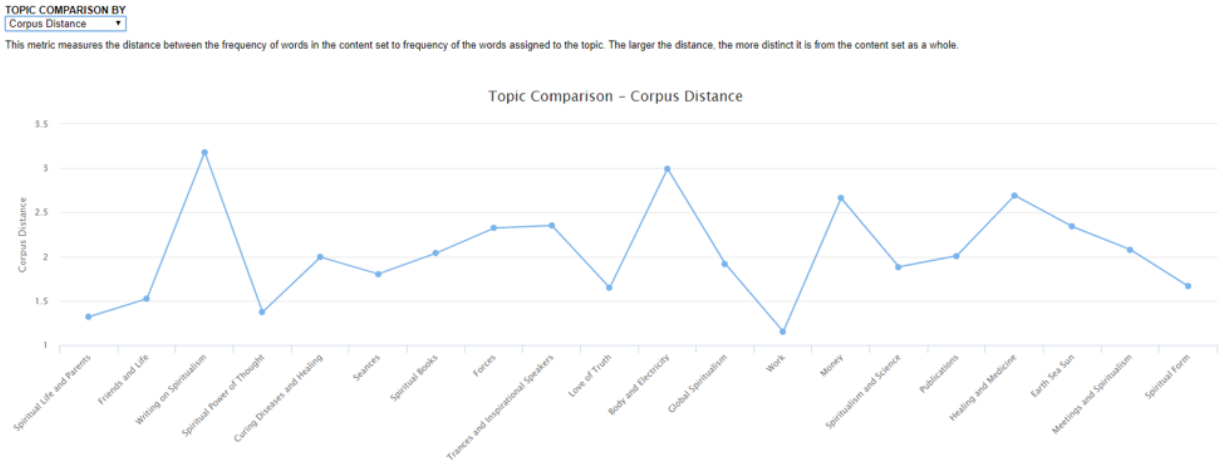
Coherence provides a metric that suggests how closely knit the words in the topic are within the texts. Since a topic is composed of words that have a greater statistical chance of appear near each other, this shows how great that proximity actually is for a topic. Notably out of all the topics, ‘body and electricity’ has the lowest coherence, indicating that the words that the tool suggest are a topic, are in fact spread out further from one another in contrast to those which make up the topics ‘spiritual power of thought’ or ‘trances and inspirational speakers’.



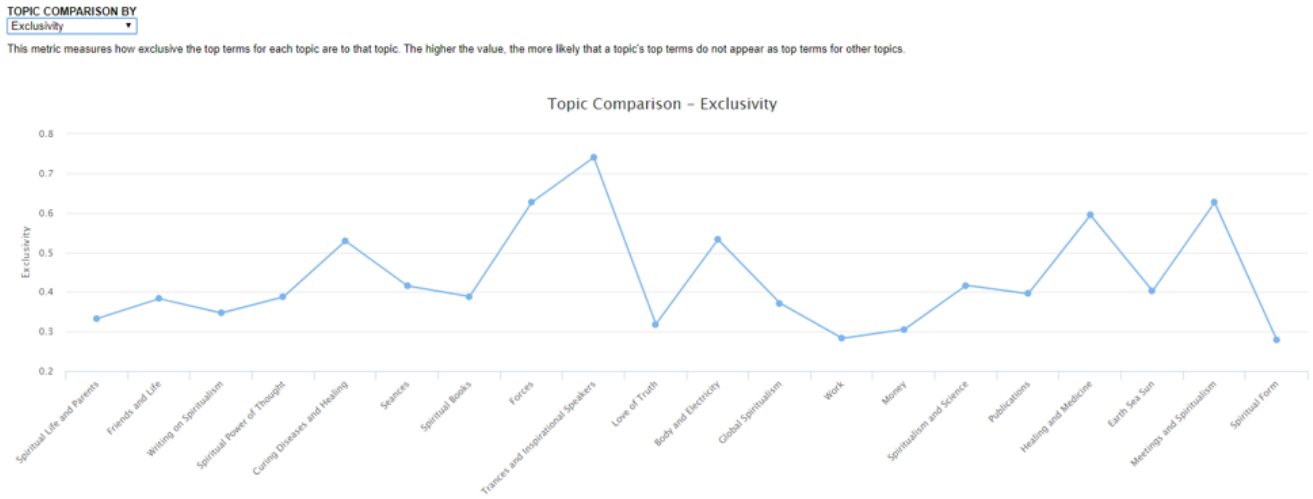
Uniform distance essentially compares a topic to the ways in which words are distributed within texts. Conceptually, it's related to the coherence metric, but instead of comparing the distance of words within a topic, it compares those to how all words are distributed, and in relation to the topic words themselves. This metric helps discern how specific a topic might be - in this case 'trances and inspirational speakers' is the highest, with 'spiritual form' 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 'body and electricity' comes in a very pronounced second, suggesting that this topic really is quite distinct from the rest of the language within a content set.



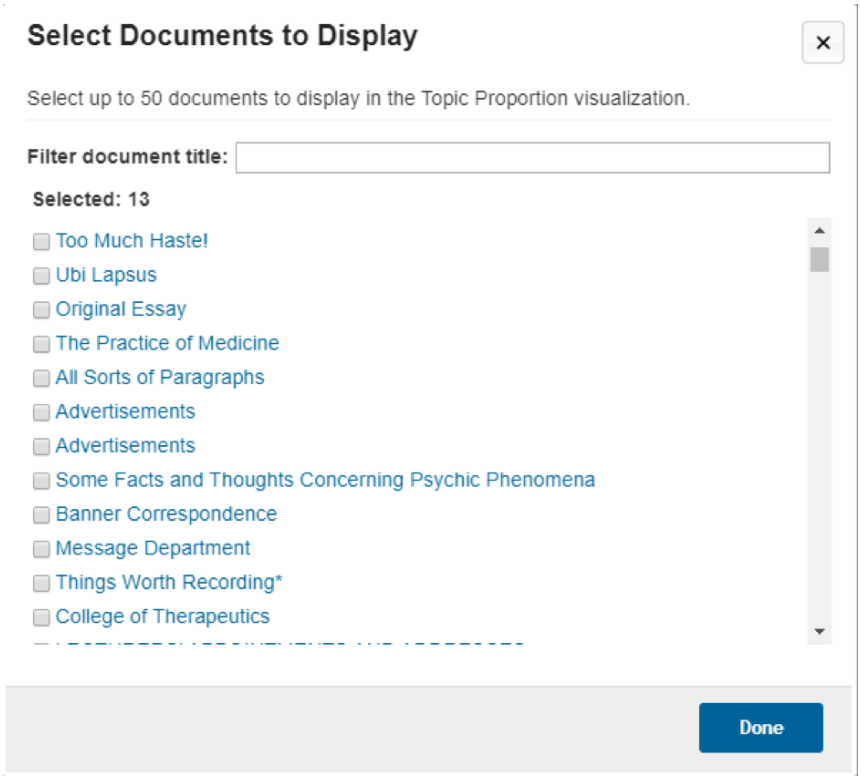
Exclusivity shows us another way of thinking about uniqueness or distinctiveness of a topic in relation to the broader content set. Since each topic is made up of a group of statistically proximate words, there's a good chance those words overlap with one another. The degree to which they don't - ie that they are distinct to a specific topic - 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. Here we see 'trances' yet again, but also 'forces', 'healing and medicine', and 'meetings and spiritualism'. The words that make up 'body and electricity' are 5th in terms of exclusivity.



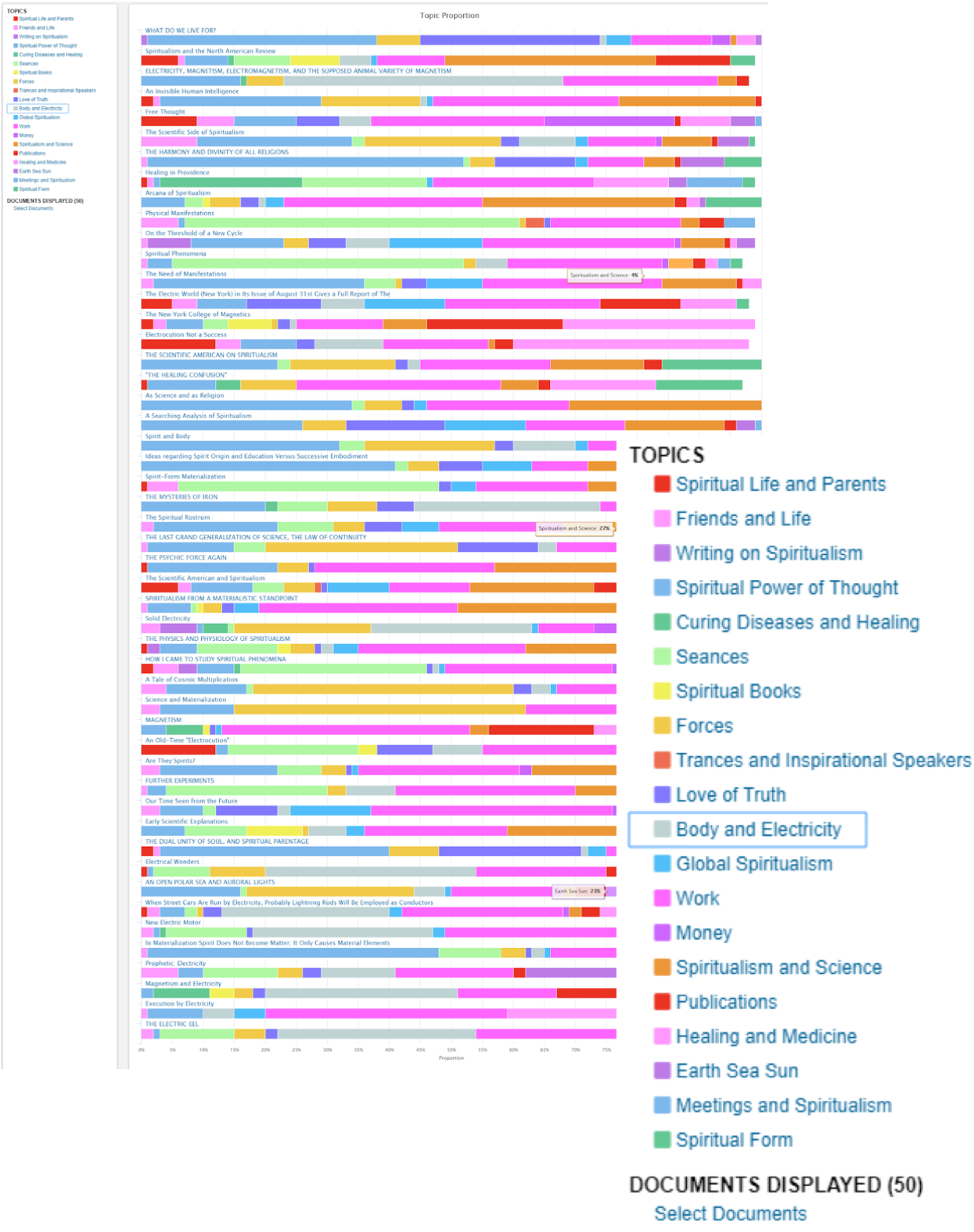
Configuring the Topic Proportion

The Topic Proportion view of the Topic Modeling tool, permits us 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. But you can select your own and refresh the visualization.



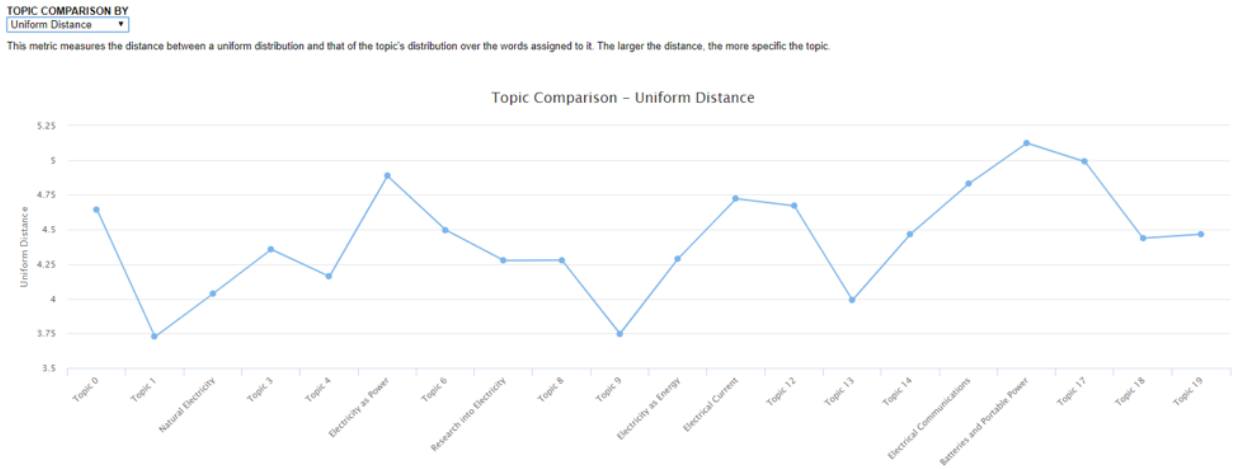
Along with the Topic Comparisons, it helps to create names for your topics when using this view. If you click on a specific topic, the visualization will shift to display only the percentage of each text where the topic appears.


TOPICS
Spiritual Life and Parents
Friends and Life
Writing on Spiritualism
Spiritual Power of Thought
Curing Diseases and Healing
Seances
Spiritual Books
Forces
Trances and Inspirational Speakers
Love of Truth
Body and Electricity
Global Spiritualism
Work
Money
Spiritualism and Science
Publications
Healing and Medicine
Earth Sea Sun
Meetings and Spiritualism
Spiritual Form
DOCUMENTS DISPLAYED (50)
Select Documents
Topic Proportion
WHAT DO WE LIVE FOR?
Spiritualism and the North American Review
ELECTRICITY, MAGNETISM, ELECTROMAGNETISM, AND THE SUPPOSED ANIMAL VARIETY OF MAGNETISM
An Invisible Human Intelligence
Free Thought
The Scientific Side of Spiritualism
THE HARMONY AND DIVINITY OF ALL RELIGIONS
Healing in Providence
Arcana of Spiritualism
Physical Manifestations
On the Threshold of a New Cycle
Spiritual Phenomena
The Need of Manifestations
Spiritualism and Science: 4%
The Electric World (New York) in Its Issue of August 31st Gives a Full Report of The
The New York College of Magnetics
Electrocution Not a Success
THE SCIENTIFIC AMERICAN ON SPIRITUALISM
"THE HEALING CONFUSION"
As Science and as Religion
A Searching Analysis of Spiritualism
Spirit and Body
Ideas regarding Spirit Origin and Education Versus Successive Embodiment
Spirit-Form Materialization
THE MYSTERIES OF IRON
The Spiritual Rostrum
Spiritualism and Science: 27%
THE LAST GRAND GENERALIZATION OF SCIENCE, THE LAW OF CONTINUITY
THE PSYCHIC FORCE AGAIN
The Scientific American and Spiritualism
SPIRITUALISM FROM A MATERIALISTIC STANDPOINT
Solid Electricity
THE PHYSICS AND PHYSIOLOGY OF SPIRITUALISM
HOW I CAME TO STUDY SPIRITUAL PHENOMENA
A Tale of Cosmic Multiplication
Science and Materialization
MAGNETISM
An Old-Time "Electrocution"
Are They Spirits?
FURTHER EXPERIMENTS
Our Time Seen from the Future
Early Scientific Explanations
THE DUAL UNITY OF SOUL, AND SPIRITUAL PARENTAGE
Electrical Wonders
AN OPEN POLAR SEA AND AURORAL LIGHTS
Earth Sea Sun: 23%
When Street Cars Are Run by Electricity, Probably Lightning Rods Will Be Employed as Conductors
New Electric Motor
In Materialization Spirit Does Not Become Matter; It Only Causes Material Elements
Prophetic: Electricity
Magnetism and Electricity
Execution by Electricity
THE ELECTRIC EEL
0%
5%
10%
15%
20%
25%
30%
35%
40%
45%
50%
55%
60%
65%
70%
75%
Proportion
TOPICS
Spiritual Life and Parents
Friends and Life
Writing on Spiritualism
Spiritual Power of Thought
Curing Diseases and Healing
Seances
Spiritual Books
Forces
Trances and Inspirational Speakers
Love of Truth
Body and Electricity
Global Spiritualism
Work
Money
Spiritualism and Science
Publications
Healing and Medicine
Earth Sea Sun
Meetings and Spiritualism
Spiritual Form
DOCUMENTS DISPLAYED (50)
Select Documents

Scientific American

With this overview of Topic Modeling, we'll just provide some highlights here from the Scientific American results to compare against the Banner of Light results. There aren't any obvious overlaps.

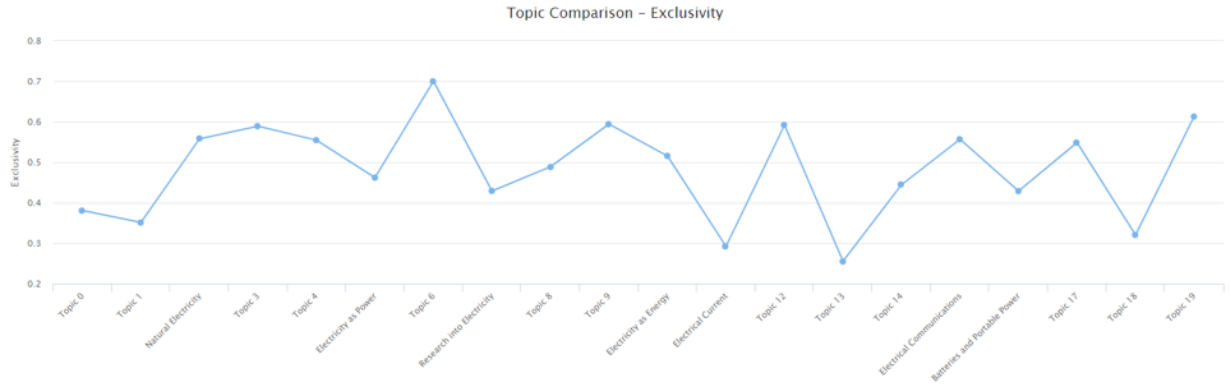
The uniform distance measure does indicate that the most specific topics with electricity are 'Electrical Communications', 'Batteries and Portable Power', 'Electrical Current', and 'Electricity as Power'.

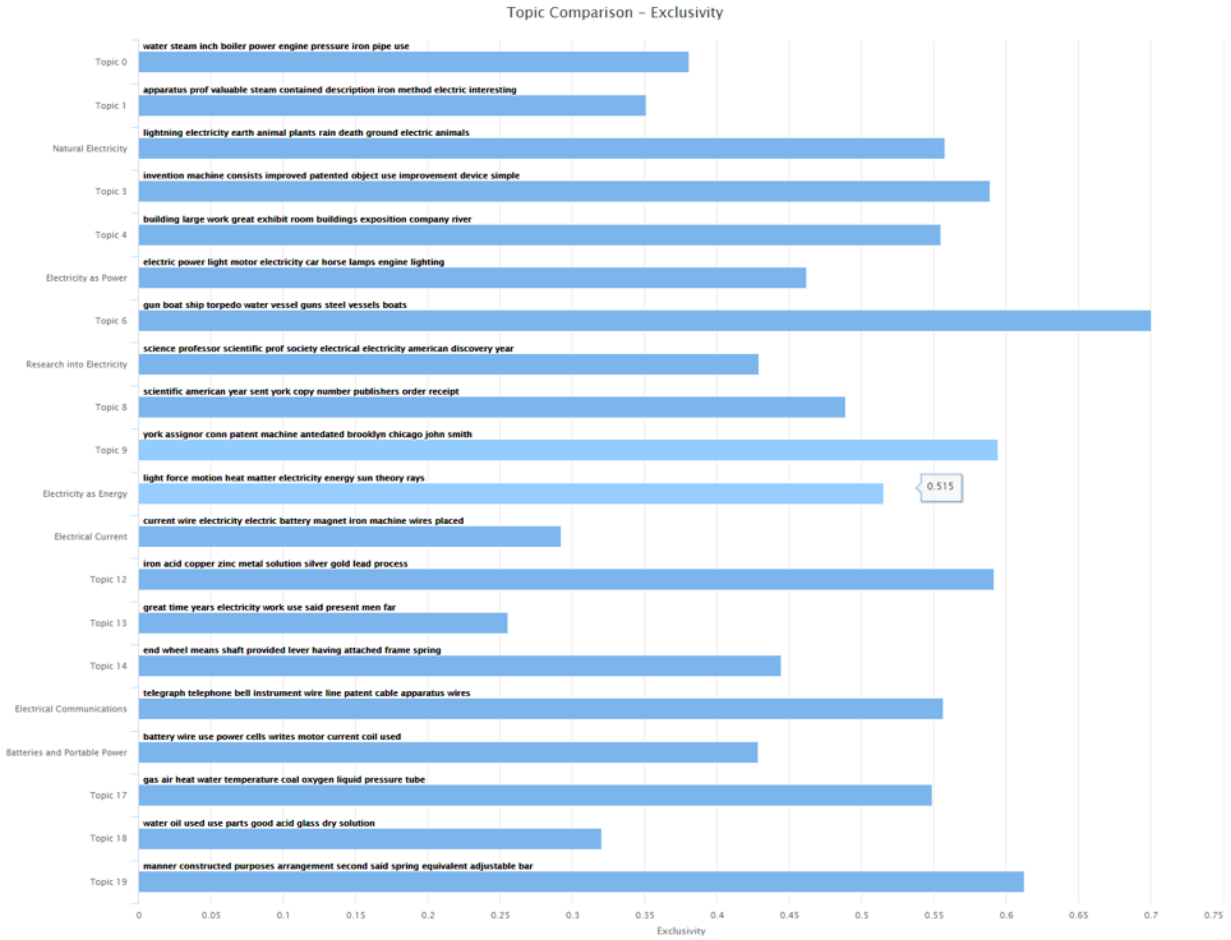


Exclusivity seems somewhat inconclusive as well. In the end the measures for these topics indicate a wide array of topics, and though there are some clear exclusive topics such as Topic 6 - involving guns and boats, there's also a healthy mix of topics in this content set.

TOPIC COMPARISON BY
Exclusivity

This metric measures how exclusive the top terms for each topic are to that topic. The higher the value, the more likely that a topic's top terms do not appear as top terms for other topics.

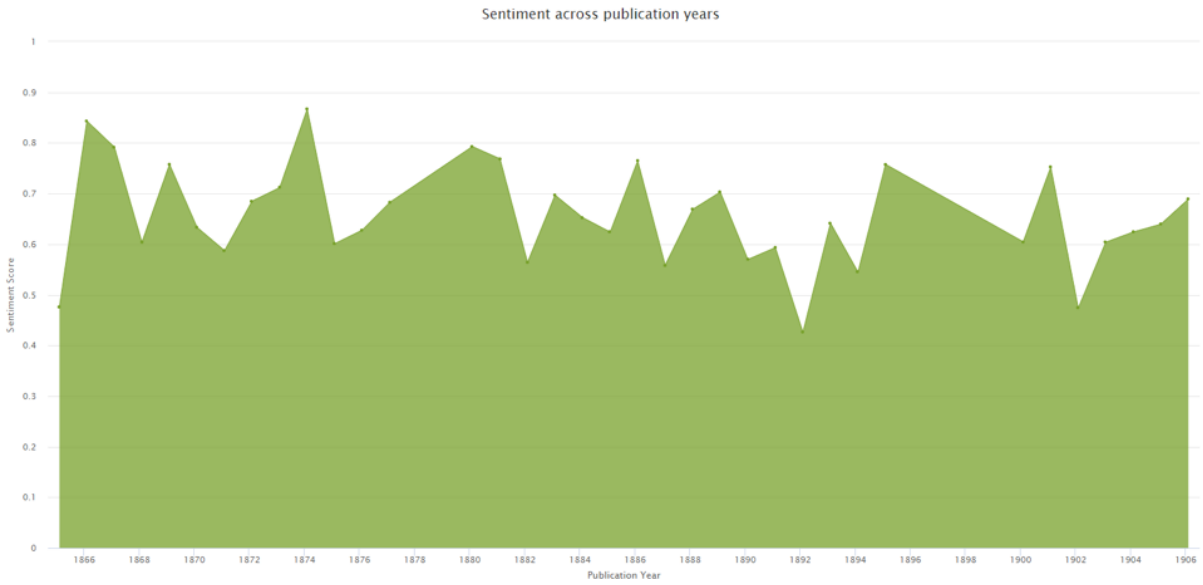




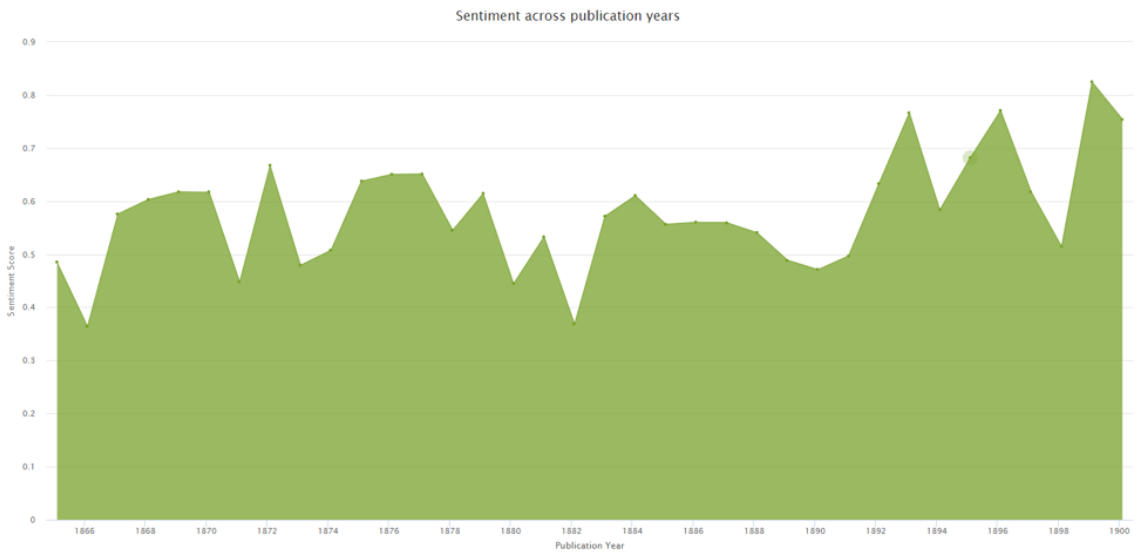
Sentiment Analysis

Perhaps unsurprisingly, the results for sentiment analysis for both content sets is consistently positive for both periodicals. This could suggest that electricity was an overwhelmingly positive topic, but we must remember that this tool provides a measure across the entire document - and that in serial publications with highly varied and mixed content, there's no clear means of ascertaining if such positive sentiment is in fact due to electricity, or perhaps rather, something else. They could also be a matter of the topics of both journals - spiritual well-being and growth, and scientific advancement and progress. Both of these broad areas of interest tend to focus on either personal or social / scientific / economic development, which tend (usually) to have more positive vocabularies than negative. The idea of betterment and progress, which was such a part of the culture of invention and discovery showcased by Scientific American, might explain why the highest values are seen in the 1880s and 1890s in the Scientific American content set. Consequently, though the results might look like something overwhelmingly positive, more precise content sets are likely required to drill down into these results.

Banner of Light



Scientific American

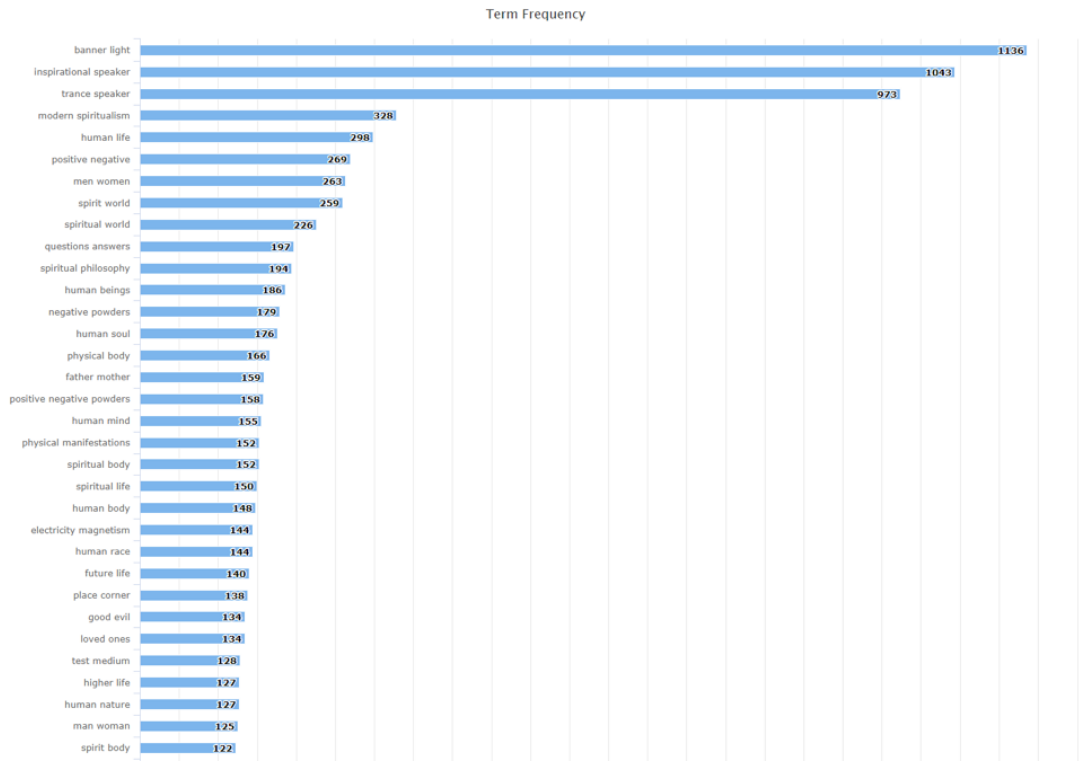


nGrams

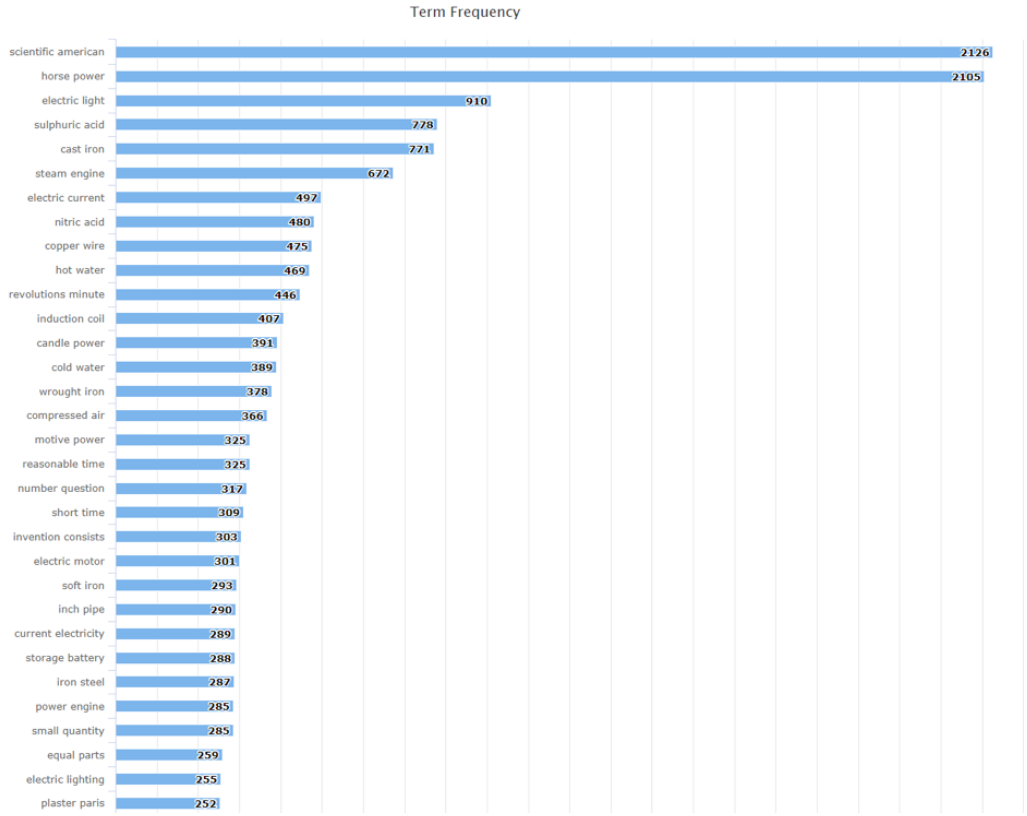
Configuration: Min 2 Max 5, Threshold 5

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 nGram results is not merely that electricity isn't very prevalent, but that in the Banner of Light results, the word which appears with electricity is magnetism, not something more apparently to do with spirituality as we might expect. Equally, in Scientific American, the word light appears before current - 'light', as you can see above, is an important word in the Banner of Light topics. But it's also one of the most important scientific inventions involving electricity - the light bulb. With the Banner of Light nGram results, we find some hints, finally of some overlap between the two periodicals, and a new research question: what's the link between electricity and magnetism across the two publications?

Research Outcomes

The results are inconclusive, but we do find some commonalities. The overwhelming positivity of the Sentiment Analysis results, and the low appearance of electricity in the nGram results seem 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’?

    A. It’s 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 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 19th century?

    A. At first glance, it doesn’t appear so. However, familiarity with the period, and the history of science and religion, 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 nGram 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. We can see this in the topic 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. What sentiments appear in discussions of electricity?

    A. They’re overwhelmingly positive - as we discussed above, this requires some reflection as there might be other factors at play here.

4. 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 19th and early 20th centuries?

    A. Not that we can see at the moment.

New Questions

5. Question 2 above presents a possible new line of inquiry - Electricity and Spirit, perhaps. Or health, or body, or force. This will require more precise Content Sets to explore.

6. Perhaps the most fascinating discovery was a piece 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 - as the document is part of our content set.

Reflections on Method

Content Set Building

The content set building for this comparative project consisted of finding appropriate temporal boundaries for two serial publications - we opted for 1865 as this is a significant year in American history - the end of the Civil War. 1918, as the end of the First World War or World War I, is also an important cultural moment. In between these two conflicts, the place of electricity within Americans society moved from a fairly limited notion through industrial and scientific development, into practical applications. At the same time, what it meant, and what it was, was a topic of intense cultural fascination and discussion. We can see both of these concerns 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 - the temporal boundaries, and the word used: ‘electricity’.

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 need their own stop word lists, in order to remove the ‘noise’ - words arising from serial publications and advertisements, as well as unuseful information, like placenames for Banner of Light, and scientific experimentation words for Scientific American.

Both content sets, however, were also muddied by the presence of documents that often appear in serial publications - advertisements, notes and queries, letters from readers, set or repeating editorial sections etc. 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:

All Content (137)

Search Terms: *Entire Document* ("electricity") *And Entire Document* ("health") *Not Document Title* ("banner of light") *Not Document Title* ("Healing Media") *Not Document Title* ("Untitled") *Not Document Title* ("BUSINESS MATTERS") *Not Document Title* ("Multiple Essay Items") *Not Document Title* ("Newsy Notes and Pithy Points") *Not Document Title* ("Answers to Questions") *Not Document Title* ("The Spiritual Bostrum") *Not Document Title* ("advertisements") *Not Document Title* ("message department") *Not Document Title* ("meetings in boston") *Not Document Title* ("lecturer's appointments and addresses") *Not Document Title* ("the spiritual rostrum") *Not Document Title* ("All Sorts of Paragraphs") *Not Document Title* ("The Rostrum") *Not Document Title* ("Brief Paragraphs") LIMITS: *Publication Title* ("Banner of light" X) *And Document Type* ("Essay" X) *And Archive* (American Historical Periodicals from the American Antiquarian Society X) *And Publication Date* (1865 - 1918 X)

The new content sets, and their results, are linked below. It's worth comparing them against the original sets - can you find any further clarification to the original research questions?

Banner of Light - Essays and Articles (with titles excluded)

https://go.gale.com/ps/textAnalysisTools?method=updateTools&userGroupName=gdc_all&prodId=DSLAB&authType=Google&contentSetName=1581035076845

Scientific American - Essays and Articles (with titles excluded)

https://go.gale.com/ps/textAnalysisTools?method=updateTools&userGroupName=gdc_all&prodId=DSLAB&authType=Google&contentSetName=1581036274641

Understanding Outcomes

Revising Questions

As discussed in the guide, it's not only normal to revise your 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 which could lay beyond the scope of your 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 did not 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 determining, following examination of each document, whether or not it was appropriate to include in a Content Set focused on the specific parameters of the project.
- Iteration is not restricted to cleaning - it's 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.

Beyond the Lab

Presentations

All of the tool outputs can be downloaded as images to use in powerpoints, or embedded in webpages or other ways to present your work.

New Visualizations

It's also possible to download the data which power the visualizations as comma delimited (CSV) or javascript object notation (JSON) files, allowing you to create and format your own visualizations. If you have the skills, it's possible to collate or create new visualizations that may combine outputs from similar visualizations into one, allowing you to compare and contrast in new ways that the DSL tool does not. 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. If you're a programmer, this is the ideal place to start to explore the data created by the DSL 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 might be breaking out the scores for topics by publication date, and plotting them along the time scale, as in the Topic Modeling Martha Ballards Diary project (http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/).

And it could be done with a mind to moments in the development of electricity itself - using resources like the Electricity Timeline (http://resources.schoolscience.co.uk/britishenergy/14-16/index.html).

Refining the Content Sets

Understanding the limitations of the DSL allows us to consider what can be done to both to build content sets, and 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 like electric, electrical - but those like current, force, spark, power, motion, spirit - suggest that perhaps using the Topic Modeling tool might offer a rich method of building more precise content sets when closer reading isn't possible.

Downloading the Content Sets

As powerful as the DSL's tools are, they offer fairly standardized configurations, and are not customizable at the moment. Downloading your content sets not only allows use of other tools, but also permits custom editing and cleaning of the documents. The DSL's cleaning configurations are not as powerful as more extensive, iterative techniques that make several passes over documents to refine and clear up problems arising from OCR digitization. Downloading also allows you, as a researcher, to find problem words and tokens that can complicate or mess up your results in the DSL. Download your content set and experiment, and use what you find to help refine your DSL projects.

Reading

Helle Porsdam, 'Digital Humanities: On Finding the Proper Balance between Qualitative and Quantitative Ways of Doing Research in the Humanities', Digital Humanities Quarterly 7.3 (2013)

Bernhard Rieder and Theo Röhle, 'Digital Methods: Five Challenges', in Berry, David M.(ed), *Understanding Digital Humanities* (2012).