Tracking Archaeology in *The Illustrated London News*

Sarah Ketchley

This digital project was prompted by my interest in the reporting of archaeology in The Illustrated London News, a publication notable for its fine illustrations and contributions by some of the preeminent figures of the day. Gale Primary Sources offers access to the entire run of the newspaper, covering the period 1842–2003. This project essay describes my workflow for the preliminary investigation of the data: initial content set creation, cleaning, analysis, export, and visualization. At the outset, the research questions were necessarily broad:

- Which words were most prevalent in articles reporting about archaeological digs?
- What themes or topics are most prevalent in the dataset?
- What was the overall feeling about this type of reporting? Was it reported favorably?
- Is it possible to identify which archaeologists were directly contributing to the publication and how many contributions they made?

Engaging in the practical process of curation and analysis offers opportunities to refine these questions, and almost inevitably suggests new avenues for future exploration.

Building a Dataset

I built my dataset by searching the *Gale Primary Sources* archive using Advanced Search in *Gale Digital Scholar Lab* for keywords, including "archaeology," which appears for the first time in 1881; "excavations"; and "ruins." I also searched by site, excavator, and civilization; for example, "Layard," "Assyria," "Nineveh," "Sumerian," "Egypt," "tomb," etc. While the results weren't comprehensive, I ended up with an initial content set of 2,513 primary source documents—mostly newspaper articles with some advertisements. This corpus comprised a collection of optical character recognition (OCR) texts generated from original scans of the ILN Historical Archive, accessed via the *Gale Digital Scholar Lab* and *Gale Primary Sources*.

I wanted to examine the content of the documents to see if I could identify recurrent themes or topics, along with the most common words and expressions of positive or negative sentiment in the dataset. I opted to work with Tableau to generate multiple visualizations for display on an interactive dashboard. To create the statistical data underlying these visualizations, I used the Lab to run a series of text-mining analyses with the tools available in the platform using the following workflow:

1. In order to perform statistical analysis on my content set, I needed to **clean** the texts to remove recurrent OCR errors, and also to remove stop words, which are the most common words in the English language and not of interest for my research. I ran a preliminary nGrams analysis to identify prevalent OCR errors: I configured the tool to return only unigrams (single words), and exported the results as a CSV. I was able to identify and weed out some regularly occurring OCR errors reflected in the CSV, which I pasted into the stop word list in the Clean tool. I continued to iterate on this process, removing and correcting the base OCR texts.

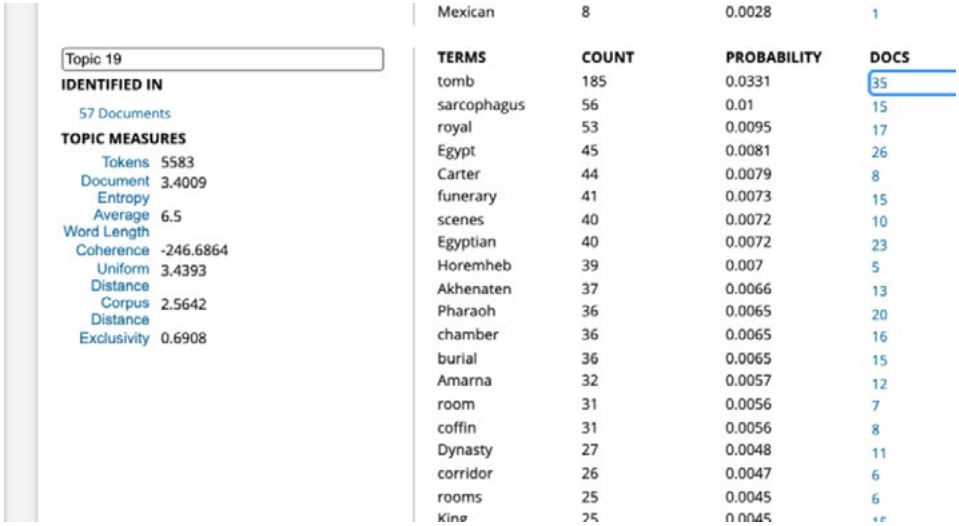
Once I created a clean configuration that removed most of the most common OCR text errors, I ran the collected dataset through my analysis tools.

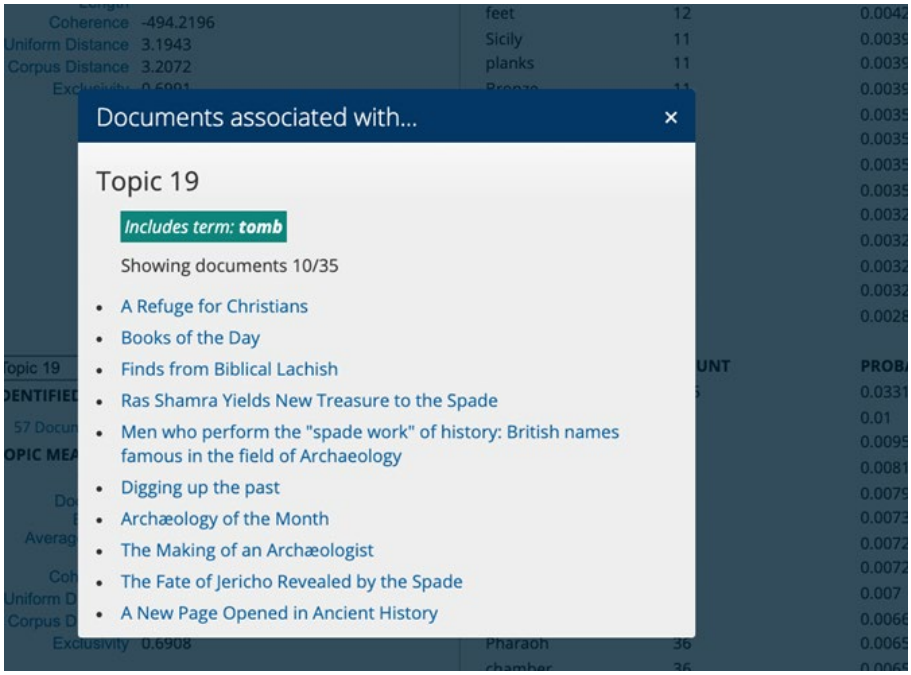
2. **Topic Modeling:** The Lab uses a tool called MALLET to perform LDA topic modeling on the corpus of texts. The algorithm iterates through the "bag of words," or collected textual data, and identifies terms that are topically similar, then groups them together. The configuration I chose to run was 30 topics, each containing 20 topic terms, while also applying the cleaning configuration I'd created. I chose this number of topics because I wanted to move beyond what the algorithm would find most obvious, and instead discern connections that were less apparent in the dataset. I exported the results of this analysis run as a CSV, as well as a second "topic proportion by document" analysis spreadsheet, which I didn't end up using for this run of visualizations but which nonetheless provides a wealth of granular detail about my documents.

	A	B	C	D	E	F	G
1	topic_name	topic_id	topic_allocation_count	topic_allocation_ratio	topic_coherence	topic_corpus_dist	topic_document_entropy
2	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
3	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
4	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
5	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
6	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
7	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
8	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
9	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
10	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
11	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
12	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
13	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
14	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
15	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
16	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
17	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
18	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
19	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
20	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
21	Rome and Roman	0	0.0784	0.006	-399.0384	3.4582	5.5557
22	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
23	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
24	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
25	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
26	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
27	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
28	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
29	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
30	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
31	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
32	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
33	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
34	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
35	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
36	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472
37	London Institutions	1	0.1574	0.2341	-435.9915	2.8971	5.8472

Gale Digital Scholar Lab Topic Modeling CSV Output.

I examined individual articles returned in the results list using the Document Terms output alongside the Documents by Topic pop-up.





3. **Sentiment Analysis:** Gale Digital Scholar Lab visualizes sentiment across time using the AFINN sentiment lexicon, which ranks documents as positive, neutral, or negative on a scale of +5 to -5, based on the words included in the text. I ran this tool on my cleaned dataset and exported the results as a CSV.

	A	B	C	D	E
1	docId	document title	publication year	sentiment score	relative mean score
2	HN3100297680	Archv¶ological Discovery for Years	1933	2.4	3.0687444
3	HN3100523109	Awards	1984	2.3888888	3.0508146
4	HN3100522845	Hotel Guide	1984	2.3333333	2.9611664
5	HN3100520327	Archaeology Awards	1978	2.2162163	2.772179
6	HN3100105681	Archv¶ology	1879	2.1666667	2.6922224
7	HN3100559116	in the Field of Mayan Archv¶ology: One of	1947	2.142857	2.6538014
8	HN3100304840	Animal Sculpture in	1935	2.1363637	2.6433234
9	HN3100283757	and Other	1930	2.1176472	2.613121
10	HN3100183131	The King's Tour	1903	2.1052632	2.5931375
11	HN3100256276	Fashions and Fancies	1923	2.032258	2.4753315
12	HN3100359177	Nimrud Looked like	1951	2	2.4232779
13	HN3100305298	Bronze Age Pottery	1935	2	2.4232779
14	HN3100279345	News	1928	2	2.4232779
15	HN3100538155	Original Monument in Bronze Has Lately	1887	2	2.4232779
16	HN3100294863	Fountain-Head of	1932	2	2.4232779
17	HN3100360321	Civilisation in	1951	1.9180328	2.2910101
18	HN3100553758	of Ancient Sculpture	1926	1.9166666	2.2888057
19	HN3100446399	Afghanistan	1879	1.882353	2.233435
20	HN3100413042	Discoveries at Igbo	1970	1.88	2.2296379
21	HN3100408282	Greek Coinage...	1966	1.875	2.2215695
22	HN3100294412	Rome: Relics of	1932	1.8695652	2.2127998
23	HN3100274467	News	1927	1.8571428	2.192754

I anticipated that visualizing this data would give me a sense of how archaeology and archaeological reporting was presented in a popular publication; ideally, I'd like to compare this with other contemporary newspaper reporting while also taking into account authorship—whether the material was written by an archaeologist or by a staff reporter.

4. **Clustering:** This analysis is carried out using the k-means clustering algorithm. The documents were grouped in 20 clusters, according to the algorithm's ranking of proximity or similarity in document content or other factors, and I again downloaded the CSV output.

	A	B	C
1	cluster	docId	document title
2	Cluster-1	HN3100249693	Meiringen
3	Cluster-1	HN3100502954	Dealing with the Oriental Cylinder Seals, Appearing in Our Last Issue
4	Cluster-1	HN3100429318	Lectures
5	Cluster-1	HN3100120034	to a Fellowship at Lincoln College; and Mr. W. M. Ramsay, M. A., Fellow of
6	Cluster-1	HN3100042114	Nineveh Antiquities
7	Cluster-1	HN3100505169	The New Year Honours
8	Cluster-1	HN3100555365	(illustrated in Our Issue of Nov. 26) Found at Beisan during Excavations by the
9	Cluster-1	HN3100499709	A Window on the World
10	Cluster-1	HN3100135502	CausV¢, upon Mr. Augustus Wollaston Franks, M. A., of Trinity College, C. B., F. R. S., Keeper of British and MediV¶val Antiquities. &c., at the British Museum; and
11	Cluster-1	HN3100073518	A Bronze Money-chest Has Lately Been Discovered in the Excavations at Pompeii
12	Cluster-1	HN3100186253	Roman Military Antiquities in Scotland: Fortifications and Military Pits
13	Cluster-1	HN3100544733	The Ears of a God: Wonderful Prayer-Tablets and Other Discoveries at Memphis
14	Cluster-1	HN3100000160	Saturday Last
15	Cluster-1	HN3100194222	Latest Excavations in the Sinai Peninsula: Monuments from 4600 B. C.
16	Cluster-1	HN3100096092	ArchV¶ology
17	Cluster-1	HN3100522845	Shetland's Own Good Hotel Guide
18	Cluster-1	HN3100259126	Grindel-Wald
19	Cluster-1	HN3100116524	A Buried City of the Exodus
20	Cluster-1	HN3100418510	Swans Art Treasures Tours
21	Cluster-1	HN3100058333	Eleven Inches High, Having upon It the Name of Xerxes in Two Languages?one
22	Cluster-1	HN3100428727	Explore
23	Cluster-1	HN3100396843	TV Schweppsicolor
24	Cluster-1	HN3100087381	Calendar for the Week Ending June 8
25	Cluster-1	HN3100065694	Flags of Which Were Scattered about More Than 200 Skeletons

I found this output more challenging to visualize in a meaningful way, and ultimately opted to drop clustering in favor of topic modeling, which provided a thematic breakdown sufficiently detailed to be of interest.

5. Finally, I downloaded the **metadata** for all the documents, which included author, title, date and place of publication, and document ID. I wanted to see if I could identify whether archaeologists wrote regularly for the newspaper, and the content and context they provided compared with more formal academic publications. This is where the real challenge began!

Working with Dates in Excel

I looked at the date formats and noted that post-1900 dates were written numerically, while pre-1900 were text. This prompted the methodological question of how to clean up these variances, an issue that was compounded by Excel not recognizing dates before January 1, 1900.

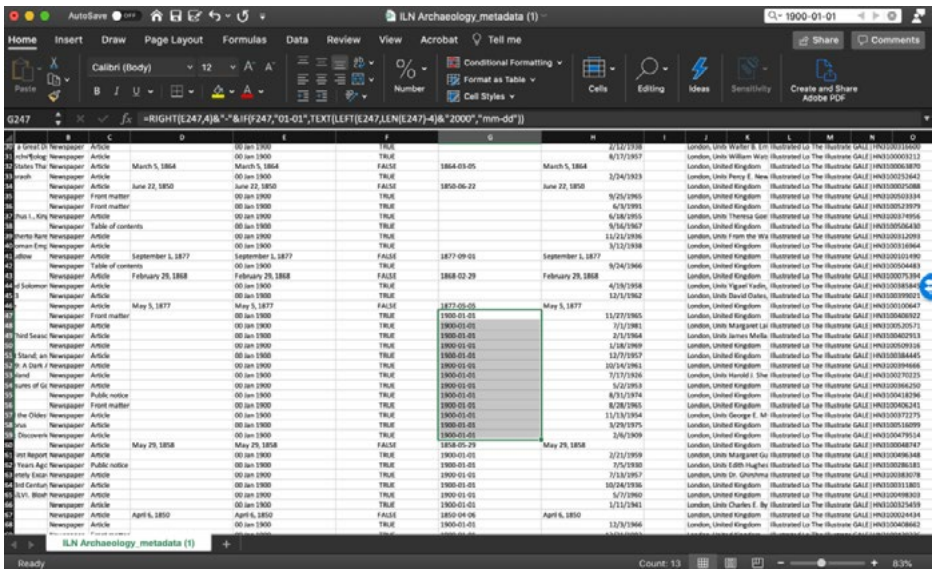
I was able to find a couple useful resources online that helped me solve the problem, but I was surprised by how complex it turned out to be.

This article provided good background context, while this post ultimately solved my issue.

I began by manually splitting out the text dates into a separate column.

D	E	F	G	
Title		Publication Date	Publisher	Place
ted London News.	April 8, 1882	April 8, 1882		Londc
ted London News.		7/22/1961		Londc
ted London News.		8/17/1935		Londc
ted London News.		2/6/1937		Londc
ted London News.		1/29/1938		Londc
ted London News.	February 3, 1877	February 3, 1877		Londc
ted London News.		10/27/1973		Londc
ted London News.		11/27/1982		Londc
ted London News.		2/13/1960		Londc
ted London News.		4/26/1986		Londc
ted London News.		3/27/1937		Londc
ted London News.		9/12/1964		Londc
ted London News.		10/3/1970		Londc
ted London News.		8/2/1958		Londc
ted London News.	August 7, 1847	August 7, 1847		Londc
ted London News.		12/3/1960		Londc
ted London News.	March 10, 1888	March 10, 1888		Londc
ted London News.	June 24, 1876	June 24, 1876		Londc
ted London News.		7/28/1984		Londc
ted London News.		10/29/1983		Londc
ted London News.		9/1/1962		Londc
ted London News.		3/17/1923		Londc
ted London News.		4/19/1930		Londc
ted London News.		5/25/1968		Londc
ted London News.	November 5, 1870	November 5, 1870		Londc
ted London News.		10/8/1966		Londc
ted London News.		8/15/1925		Londc
ted London News.		2/3/1951		Londc
ted London News.		5/16/1970		Londc
ted London News.		7/12/1969		Londc
ted London News.	November 4, 1871	November 4, 1871		Londc
ted London News.		10/27/1928		Londc
ted London News.		2/28/1925		Londc
ted London News.		7/2/1955		Londc
ted London News.		5/21/1938		Londc
ted London News.	August 28, 1852	August 28, 1852		Londc
ted London News.		10/27/1984		Londc
ted London News.		1/22/1949		Londc

I then created an additional three columns in order to perform the text-to-number conversion and to get Excel to appropriately render dates before January 1, 1900.

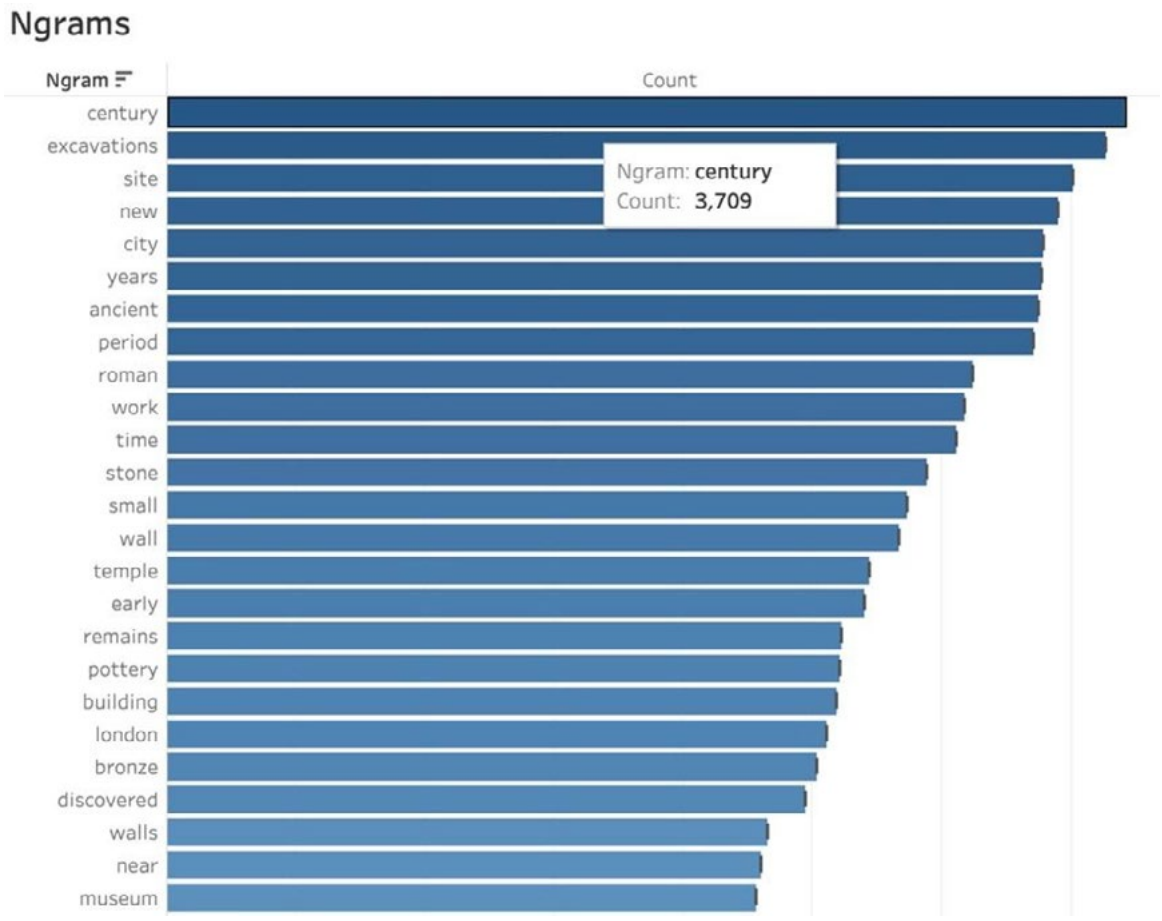


Columns E, F, and G are the added columns, and the final formula is shown in the function bar. One issue that arose was that blank fields were replaced with “1900-01-01,” and trying to clean these up with find and replace didn’t work. Again, I had to do this manually, but the outcome was a table with all dates standardized in the format YYYY-MM-DD.

Using Exported CSV Data in Tableau

Ngrams

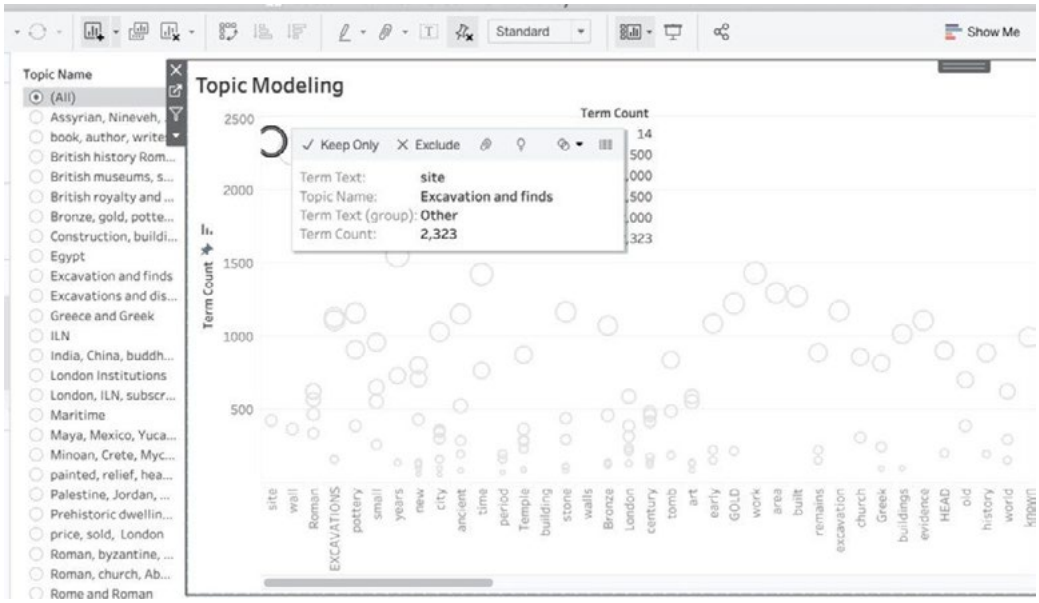
To answer the question “Which words were most prevalent in articles reporting about archaeological digs?” I visualized the nGrams CSV output in Tableau. A bar chart proved to be the most effective visualization.



Following data cleanup, the most common terms included “century,” “excavations,” and “site,” which one would probably expect in documents related to archaeology. The first civilization mentioned is “Roman,” and the city is “London,” which, given that the newspaper was published in the UK and Roman archaeological artifacts are regularly found, wasn’t surprising. Further refinements in the terms may yield interesting results about the civilizations most commonly reported on and whether the focus was on the material culture or the process of archaeology itself.

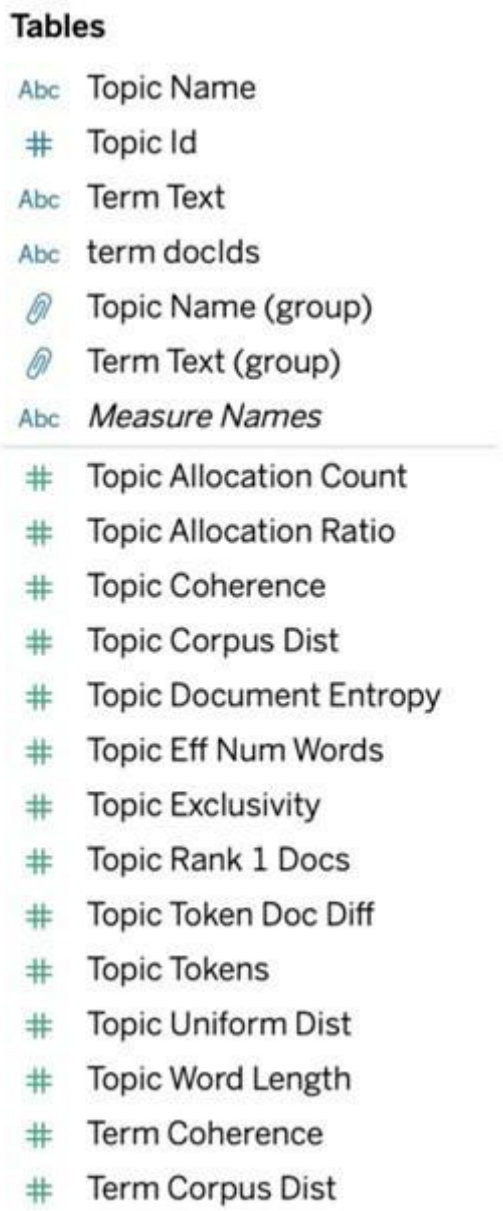
Topic Modeling

I used a circle view in Tableau to visualize my topic modeling analysis to answer the question “What themes or topics are most prevalent in the dataset?” I found the option to display all or to zoom in on a single theme to be most helpful.



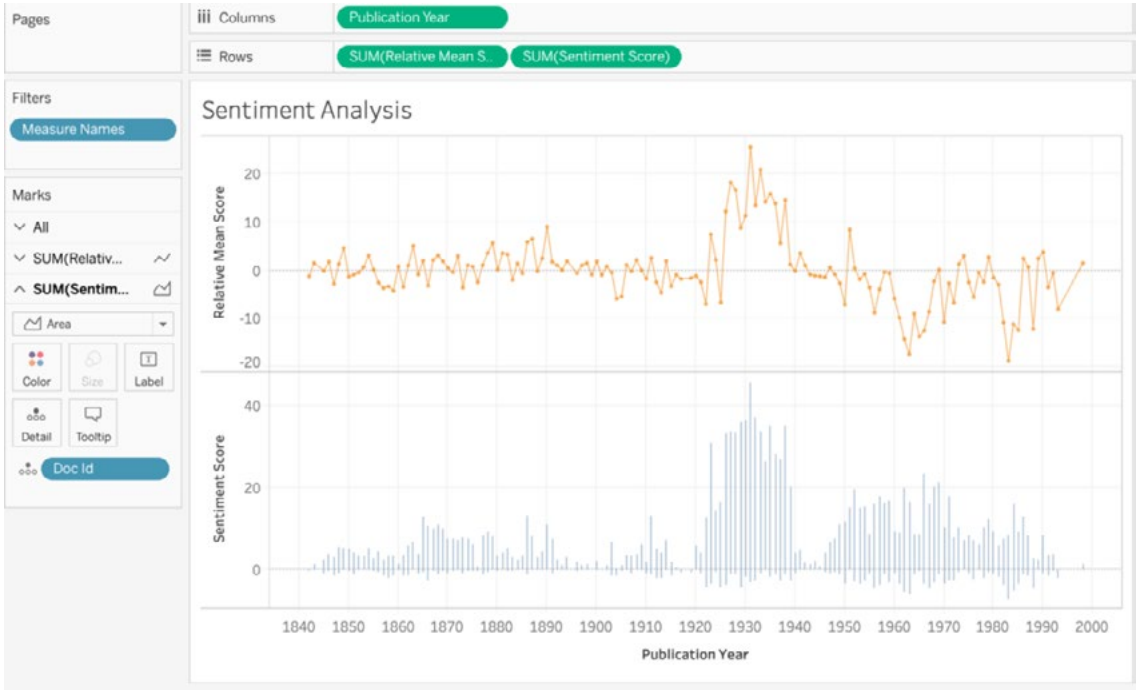
Topic modeling is a qualitative analysis, so it's incumbent on the researcher to decide what the connections are between the terms the algorithm comes up with and to then name the topic appropriately. In this case, the most common words are grouped around the theme that I named "excavations and finds." The term displayed above is "site," which occurs 2,323 times in the dataset.

I plan to do more work on this visualization, returning to the *Digital Scholar Lab* in conjunction with Tableau to experiment with various topic measures available in the platform. The granularity of this data could allow for more detailed analysis results as I continue to refine the OCR content.

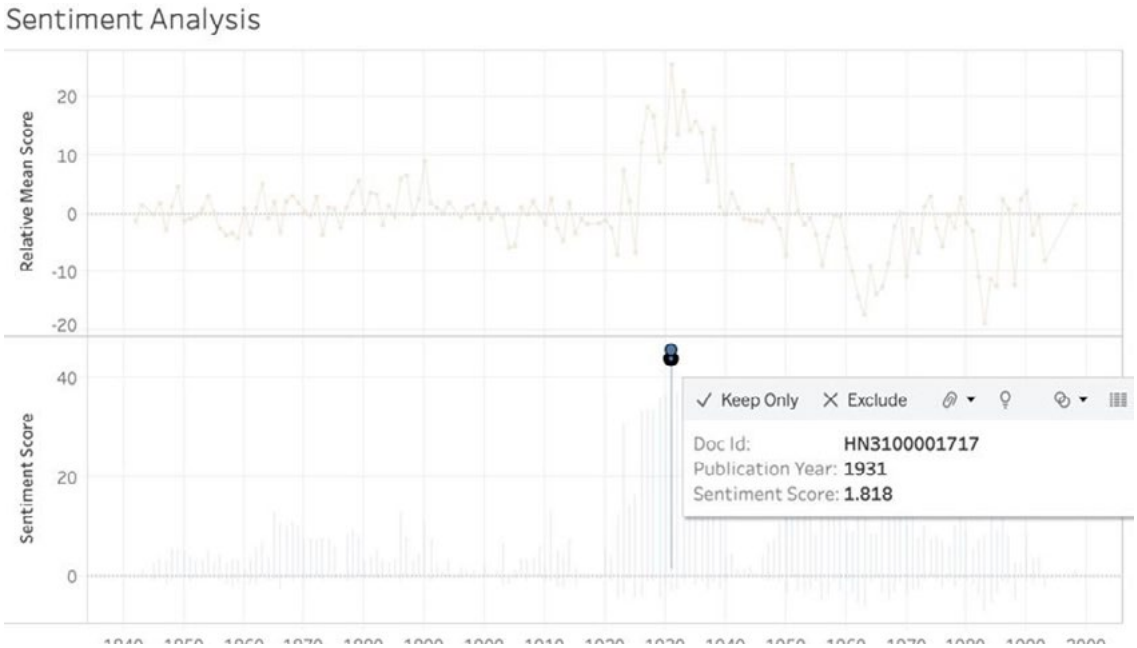


Sentiment Analysis

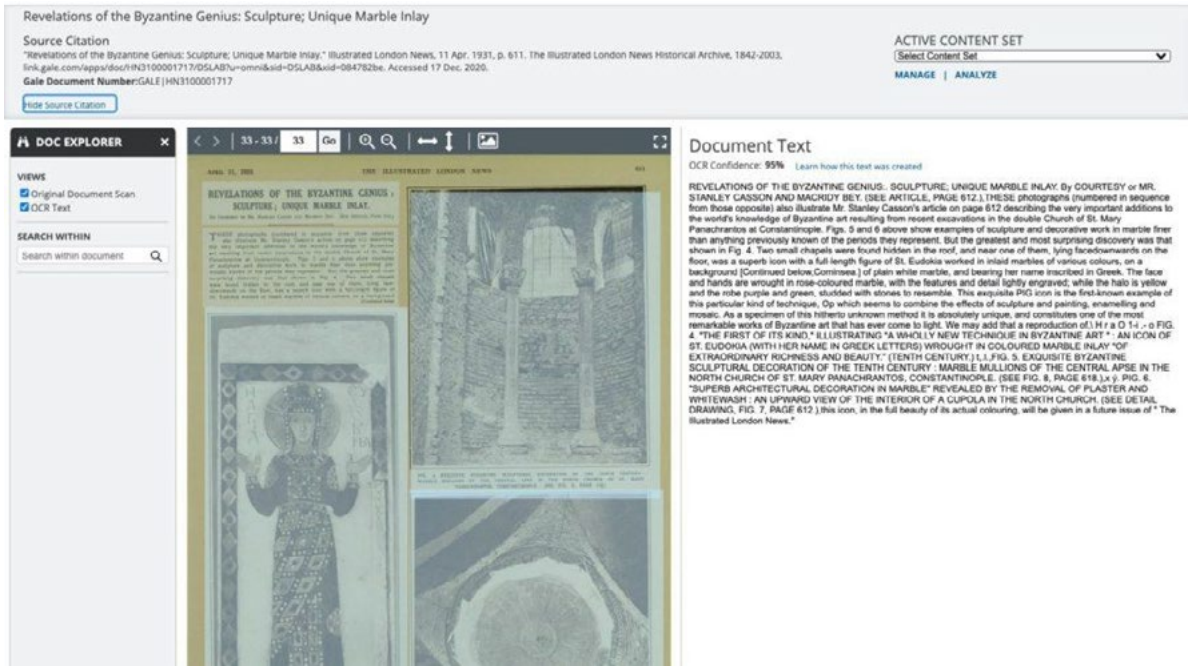
I used the sentiment analysis output to answer the questions "What was the overall feeling about this type of reporting?" and "Was it reported favorably?" I imported the CSV and created a sentiment analysis visualization in Tableau using two line charts to show the relative mean sentiment score and the sentiment score. The data was visualized over time, between 1842 and 2003, the full run of the newspaper. On the whole, the sentiment is overwhelmingly positive.



Selecting the most positive report brings up the detail of the point, along with the relevant Gale Doc ID.



I was then able to go into the Lab to find the document and figure out what made it so positive.



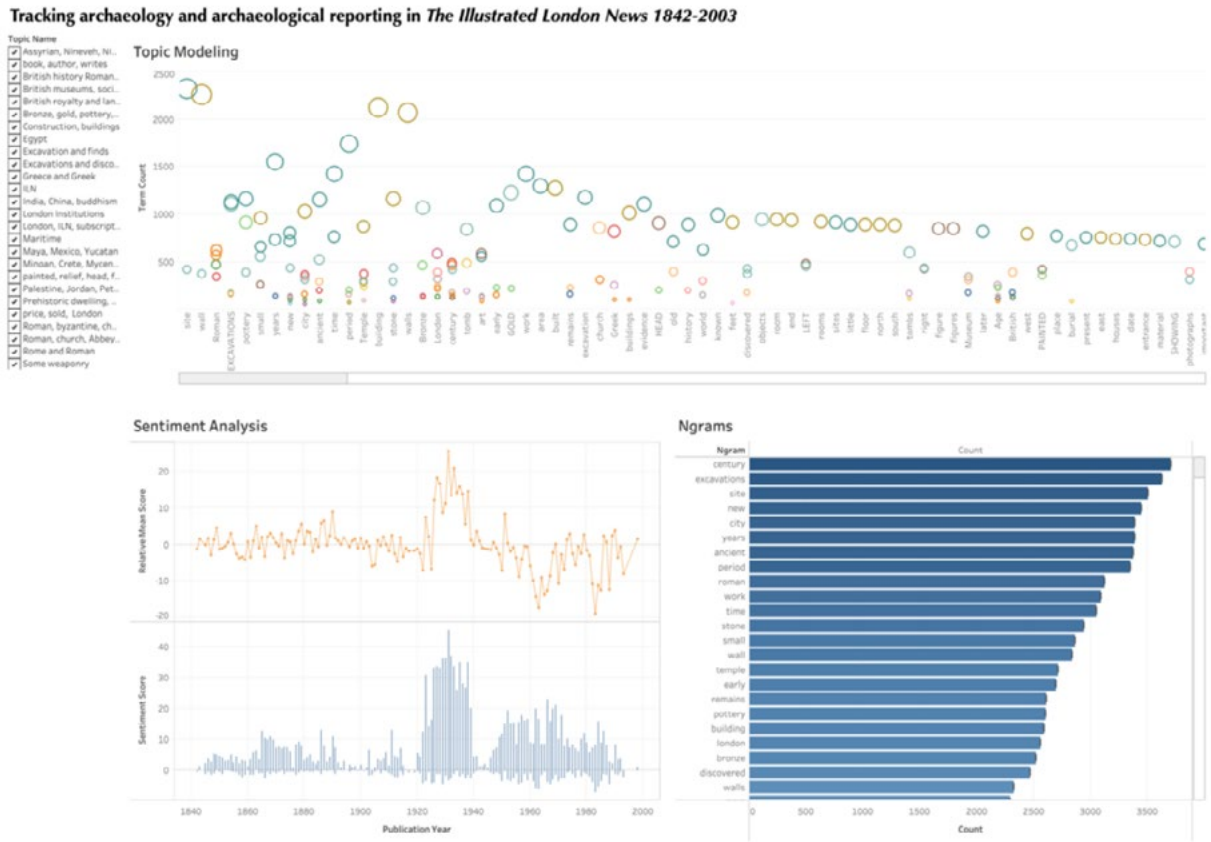
Words like “genius,” “unique,” “important,” and “finer” contributed to the positivity of the document. Looking at the period between the world wars, there was a lot of positive reporting on archaeology, and it was a time of great activity and discovery. This visualization did a nice job of capturing the spirit of discovery, while the two line charts gave a good sense of the pervading sentiment in the newspaper.

5. My final question pertained to the archaeologists themselves: “Is it possible to identify which archaeologists were directly contributing to the publication and how many contributions they made?”

The answer is yes, but interestingly, fewer of the articles prior to 1900 have an author’s name associated with them. I didn’t realize that the articles started having bylines after this date until I looked at this visualization. There are a number of well-known archaeologists authoring articles, including Max Mallowan, who excavated in Iraq and was the second husband of Agatha Christie; Howard Carter/Percy Newberry/Harry Burton/A. Mace (Tutankhamun); Henry Frankfort (early Egypt); J. D. S. Pendlebury (Egypt and Crete); John

Garstang (Egypt); and many more. It was fascinating to see how many archaeologists wrote articles for the publication. I plan to compare authorship of similar articles from other contemporary newspapers—for instance, *The Times* of London—to see if the ILN is an anomaly or the norm.

Tableau Dashboard



My final Tableau dashboard included the Ngrams, sentiment analysis, and topic modeling analyses generated by the raw exported data. Ngrams showed that while the general terms "century," "excavation," and "site" were the most common in my dataset, the most-reported excavations were those with Roman finds. This pattern was reinforced when interacting with the topic model: again, the biggest topics included those describing general archaeological terms, while the algorithm also did a good job of grouping by civilization or type of find, including Egyptian, Assyrian, Roman, Greek, Etruscan, and pottery, bronze/metal, tomb, etc., and giving a snapshot of some of the more significant archaeological activities of the day. Finally, sentiment analysis indicated that the period

between the two World Wars was most positive in nature: there were are a large number of excavations underway during this period with well-reported finds. While I didn't include my analysis of authors in the final dashboard, I was able to identify the main contributors to the newspaper, who also happened to be some of the more famous practicing archaeologists of the day. This opened a new research angle to me, as I thought about comparing published excavation reports by these individuals and their more popular writing.