19

Searching for Similarity

Computational Analysis and the US Film Industry Trade Press of the Early 1920s

Eric Hoyt, Ben Pettis, Lesley Stevenson, and Sam Hansen

During the early 1920s, few niche businesses were more crowded than the trade press of the US film industry. *Moving Picture World, Motion Picture News*, and *Exhibitor's Trade Review* published in New York City and competed for dominance among a nationwide readership. Many more regional papers sprang up in the nation's distribution hubs, such as Atlanta, Chicago, Kansas City, and Minneapolis, to serve their local industry communities. And yet still more trade papers covered the movies: *Camera!* published in Los Angeles for the production community; *Harrison's Reports* issued weekly reviews that were "free from the influence of film advertising"; and the most famous entertainment trade paper of all, *Variety*, reported on the movies alongside vaudeville and "legitimate theatre."

How unique were these trade papers? This is a question relevant to today's researchers, who encounter all of the above-mentioned trade papers (and more) when running searches within the Media History Digital Library's search platform, Lantern. Is a review in *Exhibitor's Trade Review* interchangeable with a review in *Exhibitors Herald?* Should a news item that appears in *Motion Picture News* be interpreted any differently than one that appears in *Moving Picture World?*

Questions of similarity (and its inverse, distinctiveness) were also on the minds of the publications' original readers and editors more than a century ago. Exhibitors "are watching the motion picture journals more or less critically," observed W. Stephen Bush in 1917, who had recently left his position editing *Moving Picture World* to take a leadership role at *Exhibitor's Trade Review*.[1] The motion picture distributors purchased large amounts of advertising within the same trade papers that reviewed their products. Could the reviews be trusted? How much of the new content was the work of the papers' own writers, editors, and correspondents? And how much of it was barely edited reprints of press releases, the work of studio publicists? “The reading pages of the motion picture trade papers are loaded with press matter from the various manufacturers,” alleged one industry executive at the time.[2]

Within this environment of mistrust, the trade papers competed for readers and subscribers by emphasizing their independence, originality, and distinctiveness. In *Ink-Stained Hollywood: The Triumph of American Cinema's Trade Press*, one-quarter of our team (Eric Hoyt) chronicled the rivalries among trade papers and their significance to communities within the industry.[3] His research utilized a range of sources and methods, including investigating court archives, performing quantitative content analysis, and reading countless issues of the trades. But one question lingers: Just how much overlap and similarity were there among the trade papers in using the same press releases and language?

This question matters not simply for assessing each publication's claim to a singular identity but also for understanding the early cultural norms that shaped an industry that now dominates global media, communication, and culture. If these papers truly offered unique insights and stories, then we can view the landscape of contemporary Hollywood as emerging from a genuine dialogue that represented the widely varying perspectives of people with different roles in the industry. Conversely, if the papers promoted themselves as distinct but merely parroted the same information and even spoke in similar styles, should we instead understand early industry workers as cogs in a machine that has never recognized their labor as distinctive or agential?

We turn to these questions with computational analysis methods. Text similarity measurement algorithms are widely used throughout the internet, for purposes as varied as purchasing concert tickets and flagging papers for plagiarism. If we ran similar algorithms on a corpus of trade papers from the year 1922, what patterns might emerge? Many publications carefully crafted distinct identities and claims to individuality, but how unique was the content that appeared within their pages? How might the results confirm, complicate, or complement what we already know? The nuances of the language in each publication would have helped create in-groups and out-groups that not only segmented groups within the film industry but also defined the boundaries of the industry itself. Understanding the relative similarities and differences among publications allows us to assess these publications' claims to individuality. Even more significantly for scholars of film, journalism, and media industry history, these measurements also help us understand the environment in which individual laborers were producing, distributing, and exhibiting films.

In this chapter, we discuss the process and outcomes of an exploratory study on the use of computational methods to assess large volumes of motion picture trade papers. We begin by introducing our corpus—twenty-one digitized volumes of trade papers and fan magazines. We briefly discuss the publications' backgrounds and industry profiles, including the ways they presented themselves and the ways in which they were perceived by readers. Second, we explain the computational methods that we used for measuring text similarity. We try to keep our descriptions clear and succinct while pointing readers who want to dive deeper into the specific techniques and suggestions for where to turn next. Finally, we share the results of our research study and reflect on the process and its potential broader utility for film history research. We contend that computational methods like text similarity measurement are a useful complement to traditional research methods of archival research and close reading, enabling research questions that might otherwise not be feasible for human researchers to investigate alone, particularly when working with large corpora.

UNDERSTANDING THE CORPUS: FILM INDUSTRY TRADE PRESS OF 1922

We began this project, quite naively, with the idea to run similarity measurements across the nearly three million pages available online in the Media History Digital Library (MHDL).[4] However, this proved unfeasible due to the computational processing power and time that would be required.[5] Moreover, we realized there would be advantages to narrowing our focus to a single year. We selected the year 1922, with an emphasis on July 1922, for two chief reasons. First, the MHDL had already digitized a wide cross section of trade papers from that year, including—appropriately for this book—several published outside of the United States. Second, we knew from Eric's earlier research that there was a great deal of competition within the American film industry's trade press during this period.

In 1922, *Variety* and the Chicago-based *Exhibitors Herald* were pursuing strategies to grow their readership and influence within the industry, emphasizing independence, integrity, and uniqueness as distinguishing factors. During the following year, *Exhibitors Herald* created the "'Herald Only' Club," emphasizing the loyalty of subscribers who exclusively wrote into *Exhibitors Herald* and read that paper at the exclusion of its rivals.[6] Given the competitive bent of the 1920s trade press, how distinct was each publication? Would the "'Herald Only' Club" have any factual grounding once the word patterns, sentences, and page structures were analyzed at scale?

In addition to the above-mentioned trade papers, we included sixteen additional unique journals. Our corpus included fan magazines (*Photoplay*, *Shadowland*, and *The Picturegoer*), a technical journal (*American Cinematographer*), English-language trade papers published outside the US (*Canadian Moving Picture Digest* and *The Film Renter and Moving Picture News*), and studio-generated publicity (*Universal Weekly* and *Paramount Pep*). When subjected to computational analysis, this mix of film publications held the potential for both expected and surprising similarities to emerge (see table 19.1). [tabref 19.1]

TABLE 19.1 Corpus of Selected 1922 Trade Papers

Publication	Location	Dates	URL
American Cinematographer, The	Los Angeles, US	July 1922	http://archive.org/details/americancinemato00amer
Camera	Los Angeles, US	April 1922–April 1923	http://archive.org/details/camera05unse
Canadian Moving Picture Digest	Toronto, CA	May–October 1922	https://archive.org/details/canadian-moving-picture-digest-1922-05
Cine-Mundial	New York, US	1922	http://archive.org/details/cinemundial07unse
Cinéa	Paris, FR	1922	http://archive.org/details/cina22pari
Exhibitor's Trade Review	New York, US	June–August 1922	http://archive.org/details/exhibitorstra00newy
Exhibitors Herald	Chicago, US	July–September 1922	http://archive.org/details/exhibitorsherald15exhi
Exhibitors Herald	Chicago, US	October–December 1922	http://archive.org/details/exhibitorsherald15exhi_0
Film Daily, The	New York, US	1922	http://www.archive.org/details/filmdaily2122newy
Film Renter and Moving Picture News, The	London, UK	July–August 1922	https://archive.org/details/film-renter-and-moving-picture-news-1922-07
Great Selection: "First National First" Season 1922–1923, The	New York, US	1922	http://archive.org/details/greatsel00firs
Kinematograph, Der	Düsseldorf, DE	July 1922	https://archive.org/details/kinematograph-1922-07
Motion Picture News	New York, US	July–August 1922	http://archive.org/details/motionpicturenew26july
Motion Picture Studio, The	London, UK	June 1922–February 1923	http://archive.org/details/motionpicturestu02unse
Moving Picture World	New York, US	July–August 1922	http://archive.org/details/movingpicturewor57july
Paramount Pep	New York, US	July–December 1922	http://archive.org/details/paramountpepjuld07unse
Photoplay	Chicago, US	July–December 1922	http://www.archive.org/details/photoplayvolume222chic
Picturegoer	London, UK	1922	http://archive.org/details/picturegoer34odha
Shadowland	New York, US	January–May 1922	http://archive.org/details/shadowland192200brew

(*Continued*)

TABLE 19.1 (*Continued*)

Publication	Location	Dates	URL
Tess of the Storm Country (United Artists Pressbook)	Los Angeles, US	1922	http://archive.org/details/pressbook-ua-tess
Universal Weekly	New York, US	1922	http://archive.org/details/universal1516univ
Variety	New York, US	July 1922	https://archive.org/details/variety67-1922-07

NOTE: The date range varies between publications depending on how each was compiled and digitized. Each volume may contain multiple issues of a given publication.

TABLE 19.2 Top Volume Pairings Arranged by Set Distance

Pairing	Set distance	Volume A	Volume B
1	94.3168875	*The Great Selection*	*Motion Picture News* (July–August 1922)
2	92.6631758	*Exhibitors Herald* (July–September 1922)	*The Great Selection*
3	92.2992333	*Film Daily* (1922)	*The Great Selection*
4	91.8996811	*The Great Selection*	*Exhibitors Herald* (October–December 1922)
5	91.7403586	*Exhibitor's Trade Review* (June–August 1922)	*The Great Selection*

For all of the strengths of this corpus, though, we acknowledge that it is nevertheless an incomplete cross section of the 1920s trade press. Due to the limited availability of digitized scans, there were many journals that we were unable to include. The Philadelphia-based *Harrison's Reports*, for example, featured film reviews, a fiery editorial page, and no advertisements; editor P. S. Harrison's proclaimed independence from outside interests and allegiance to independent exhibitors would make this publication a valuable point of comparison to other trade papers of the time. Unfortunately, the MHDL's thirty-four-year digitized run of *Harrison's Reports* does not begin until 1928—nine years after it began publishing—due to the MHDL's inability to access print originals for scanning. We also lacked digitized copies of the once numerous, now rare-to-find, regional trade papers that sprang up in the late 1910s and early 1920s to serve distribution exchange cities and territories such as Atlanta, Kansas City, and Minneapolis. The two papers in our corpus that were sometimes classified as regionals—Chicago's *Exhibitors Herald* and Toronto's *Canadian Moving Picture Digest*—both vigorously resented and pushed back against the "regional" designation by the early 1920s.

Despite these limitations, the 1922 corpus features many significant US trade papers of that era plus many other publications that could potentially serve as litmus tests for the process as a whole. If, for example, our computational methods indicated that the German-language *Der Kinematograph* is highly similar to *The Film Daily* or *Photoplay*, then something about the process must be inaccurate. But if our various similarity tests could reliably identify high-level similarities and significant differences between publications, it would enable us to select specific texts from the larger corpus to perform traditional close readings on.

The use of computational methods alongside traditional humanities approaches can help researchers work with enormous volumes of content. If, after analysis by a computer and close reading by a researcher, the discourse among these publications registers substantial differences, then we have an indication that robust conversation gave early industrial figures opportunities to make choices about the direction of the business. If, conversely, all publications were to receive high similarity scores across these metrics, then we would have an indication that certain topics, events, and even actual content repeats across the industrial ecosystem. The ability to perform close readings across a large corpus of text is a valuable tool for assessing not just a single publication, but broader industrial trends as well.

COMPUTATIONAL METHODS FOR SIMILARITY DETECTION

Our project situates computational methods as a complement to traditional methods of reading and analysis. Automation and scripting cannot, and should not, fully replace the role of the human researcher who interprets and synthesizes meaning from a text. Computers are highly efficient when working with enormous volumes of data, but they lack the precision and ability to interpret nuance within a text. A human researcher works more slowly but can understand that nuance in context. The text similarity algorithms that we discuss below, therefore, are not a replacement for the role of a human researcher but instead function as an assistant that can help us by processing a large set of input text and directing our time and attention toward the close readings that are most likely to yield interesting similarities. Though it is not necessary for all humanities scholars to become experts in mathematics or computer science, a working familiarity with the processes and key concepts was useful to inform our analyses of text similarity across motion picture trade papers.

Large-scale computational analyses require us to reframe how we conceptualize what the text *is*. Most humanities scholars are used to thinking of a document in a holistic sense—it contains numerous words, which are placed in a particular order to convey meaning. For many computational methods, however, the order of words in a document is entirely ignored. Instead, we construct numerical representations of written text to mathematically assess similarity by measuring the distance between numbers. For our exploratory analysis, we use the "bag-of-words" model, which understands a document as a "collection of words that are used in differing proportions."[7] Instead of treating a document as words in a particular order with meanings, we simply count the frequency of words in a given document and compare these frequencies to those of other texts. Large language models (LLMs) such as OpenAI's ChatGPT software use more advanced "embeddings" to represent text mathematically, but the underlying principle of representing text in a numerical form is similar. The transformation methods used by many LLMs can be more accurate at evaluating written text but are slow and computationally expensive; for our initial exploratory analysis, we used simpler methods such as term frequencies and calculating Levenshtein distances. These computational approaches are gross oversimplifications of textual meaning and overlook important nuance, but they also make it feasible to quickly process large volumes of text.

There are many additional caveats to our computational method, as well as several possible types of preprocessing work to address them. First, many computational methods for processing text are sensitive to differences in the lengths of documents. Texts within the MHDL corpus are not of a consistent length; some volumes contain multiple issues of a single publication, while other volumes may be separated into individual issues. Furthermore, each issue of a publication contains separate articles and sections. Separating these parts into individual documents beforehand can improve the accuracy of the calculations but at the expense of requiring more manual preparation. But standardizing page counts and sections is just one kind of significant preprocessing work that can be performed on a corpus before running similarity comparisons. For example, many projects remove stop words—common words such as *the, a*, or *and*—from input texts to avoid overrepresenting them in results. In addition, all MHDL files are processed using optical character recognition (OCR), which identifies text within a scanned image and provides data in a format that is usable in a computer script. While OCR technologies have been continually improving, it is an unavoidable fact that errors will occur. Many factors influence the accuracy of OCR text: the quality of scans, varying page layouts, differing typefaces, and even something as seemingly simple as an image appearing on a page.[8]

These are the kinds of tradeoffs that must be considered when using computational approaches, and they are an important reminder that such methods will never fully replace the role of the human researcher. For our initial exploratory analysis, we did not perform any preprocessing and instead sought to assess the effectiveness of using raw data directly from the MHDL. We selected a variety of text similarity algorithms that balance these caveats with the utility of processing large volumes of text and used their resulting similarity measures as guides to shape our ongoing research process.

EUCLIDEAN AND COSINE DISTANCE

The most basic measurements of text similarity that we used were the Euclidean distance and cosine distance between each volume of the selected corpus. Both methods measure the relative frequency of words that appear within each text and provide a numeric representation of how "far" each document is from each other.[9] They are quick to calculate and offer a general approximation of similarity but are not well suited for representing context.

To demonstrate how these distances are measured, let's consider a smaller example: two short strings of text, rather than an entire volume. Here are two sentences that regularly appear in *Exhibitors Herald* in its "What the Picture Did for Me" section:

> Sentence A: "TELL US WHAT THE PICTURE DID FOR YOU and read in the HERALD every week what the picture did for the other fellow, thereby getting the only possible guide to box office values."
>
> Sentence B: "Join in This Co-operative Service Report Regularly on Pictures You Exhibit and Read in The Herald Every Week What Pictures Are Doing for Other Exhibitors"

The first step in calculating Euclidean and Cosine distances is to identify each unique word that appears in the two texts and count its frequency in each. For example, the word *exhibit* appears zero times in sentence A and one time in sentence B. Words that may appear similar to a human reader—such as *picture* and *pictures*—are considered entirely different to the computational model. We take these word frequencies and plot them, using each word as an axis and the number of occurrences represented as the point's distance from the origin. This results in Euclidean cosine distances, which are determined by measuring the distance between plotted points. The Euclidean distance is the length of the line segment directly between the plotted points. The cosine distance is determined by drawing a line from (0,0) to each plotted point, and then measuring the angle between the two lines.[10] When working by hand, it is only feasible to compare two or three words at a time; after all, what would a graph with four or more dimensions even look like? But the underlying mathematics is the same, even with additional axes. Computers can plot points across greater numbers of dimensions, effectively comparing the relative frequencies of any given number of words.

When applied to entire volumes, these distances provide a useful overview of general similarity between texts and are a useful starting point for comparisons and analysis. Euclidean distance can range anywhere from zero to much larger values in the hundreds or thousands, with smaller values representing texts that are more similar. When comparing texts of a similar length, Euclidean distance is useful for revealing minute differences. Cosine distance, however, is more effective for comparing texts of different lengths. This measurement ranges from 0 to 1, with lower numbers being more similar.

Because these distance measures ignore the *order* of words and only consider their frequencies, they are of limited utility on their own. They are useful for providing a zoomed-out view of many texts within a corpus but (as we discuss later) yielded limited insight when applied to the MHDL corpus. Other text similarity methods were better suited for assessing motion picture trade papers.

LEVENSHTEIN DISTANCE

The other methods we used to test for similarity between texts in our corpus are variations of Levenshtein distance, a measure first proposed by mathematician Vladimir Levenshtein in the 1960s.[11] In general, Levenshtein distance is a measure of the number of edits it takes to turn one string into another. An edit can be inserting a new letter, deleting a letter, or replacing one letter with another. Accordingly, the order of the words matters, unlike in the calculation of Euclidean and cosine distances.[12] For Levenshtein distance, lower numbers of edits indicate that the two texts are more similar.

For example, consider the following words:

Word A: *color*

Word B: *colour*

To go from word A to word B, only a *u* needs to be inserted, and to go from word B to word A, only a *u* must be deleted, so they have a Levenshtein distance of 1 (out of a maximum of 6) and a normalized distance of 1/6, or about 16.67 percent. If measured using cosine distance, they would have a cosine difference of 1, or 100 percent different texts. Though technically accurate for tallying the instances of the two words, the cosine difference does not reflect the actual similarity of these terms.

When analyzing millions of words, this may mislead us into thinking two texts are more different than they actually are. Levenshtein distances help mitigate this concern by showing the similarities *within* the words themselves. Since its introduction, mathematicians have developed a number of variants of Levenshtein distance:

- *InDel distance*: Only insertions and deletions are allowed as edits.
- *Normalized distance*: The calculated Levenshtein distance is divided by the maximum possible value. By representing all distance measures between zero and one, it becomes more feasible to compare different text pairings.
- *Sorted distance*: The words from each text are alphabetized before calculating the distance. The order of the words no longer matters.
- *Set distance*: Each unique word in a string is only listed once before the sorted distance is measured.

Levenshtein distance and its variants are very useful for finding similarity where there may be regional spelling variations (e.g., *theater* and *theatre*) or where text is slightly changed during its reuse. For example, consider the following sentences:

Sentence A: The park sees hundreds of visitors a day.

Sentence B: Hundreds of visitors a day see the park.

The original Levenshtein distance is 28. Using sorted distance instead, sentences A and B have a distance of 1; only the *s* at the end of *sees* would need to be deleted since all the other words are identical. Sorting the words alphabetically first and negating the impact of their order in the sentence results in very few changes between the sets of words, suggesting that it is very likely the two sentences are highly similar.

Each of these variants can suggest distinct—and even conflicting!—interpretations of the relative similarity among texts. Using multiple measurements in combination can offer greater insight into the results. In our analyses, we used the InDel distance and sorted distance variants.

WORKFLOW

One important consideration when selecting an algorithm to use for calculating text similarity is the computational complexity and time requirements. We had to wait more than twenty-four hours for the algorithm to process each pair of texts and deliver results with multiple variants of Levenshtein distance. Current understandings of computer science suggest that it is not possible to significantly reduce this computational complexity or decrease processing time.[13] While calculating only Euclidean and cosine distances was significantly quicker, it is still not a "plug and play" process. We provide an overview of our workflow not as a step-by-step tutorial but rather to give a sense of the work that is still required even when using "automated" computational methods.

First, we downloaded raw text files of each document from the MHDL. For tracking purposes, we ensured that each file maintained the volume's unique ID.[14] Recent upgrades to the MHDL and Lantern websites have made large-scale querying and downloading possible.[15] We did not perform further preprocessing steps for our initial analysis. Assessing OCR accuracy, removing stop words, and conducting consistent stemming and tokenization may improve our process.

After preparing the text files, we used a series of Python scripts to run each comparison algorithm. We used the "pandas" and "rapidfuzz" libraries to assist with processing our text files.[16] Three different distance metrics were generated with rapidfuzz and then normalized: InDel distance (called a ratio score in the rapidfuzz library), sort score, and set score. In all three cases, higher values indicated higher similarity.

Our Python scripts created two kinds of output files: CSV files with specific similarity values for each "candidate text" and summary text files listing the top similarity values for each measurement. The text similarity algorithms helped us focus our time and attention on where we were most likely to find interesting similarities, particularly among texts that registered as similar across multiple measurements of distance.

ANALYZING THE RESULTS

Though we highlight only a few findings here due to limited space, we encourage interested readers to download the compiled data sets, available on the Media History Digital Library, and explore the calculated distances and rankings for themselves.[17]

The volumes that were most similar to one another were *Exhibitors Herald* (July–September 2022) and *Exhibitors Herald* (October–December 2022). This pair had an InDel distance of 47.96/100 and a sorted distance of 91.69/100. This finding is unsurprising; two consecutive volumes of *Exhibitors Herald* were viewed as being most similar to each other. Many structural components of a publication, such as mastheads and section headings, are likely to appear in *all* volumes, regardless of the actual content and topics included within a given issue.

The next highest sorted distances were between pairs of New York weekly trade papers:

- *Motion Picture News* (July–August 1922) and *Moving Picture World* (July–August 1922) with a sorted distance of 86.239513; and
- *Motion Picture News* (July–August 1922) and *Exhibitor's Trade Review* (June–August 1922) with a sorted distance of 85.8518859.

At face value, this would seem to support the perceptions of the aforementioned "'Herald Only' Club" members who viewed the Chicago-based *Exhibitors Herald* as distinctively different from its rivals based in New York.[18] However, further highly ranked pairings suggest that *Exhibitors Herald* had similar text to *Exhibitor's Trade Review* and *Motion Picture News* (with sort distances ranging from 85.3 to 83.8). Whether published in New York or Chicago, the sorted distances suggest that weekly US film trade papers are more similar to each other than to monthly fan magazines, non-US weekly trade papers, or even a daily trade paper from the US like *The Film Daily*.

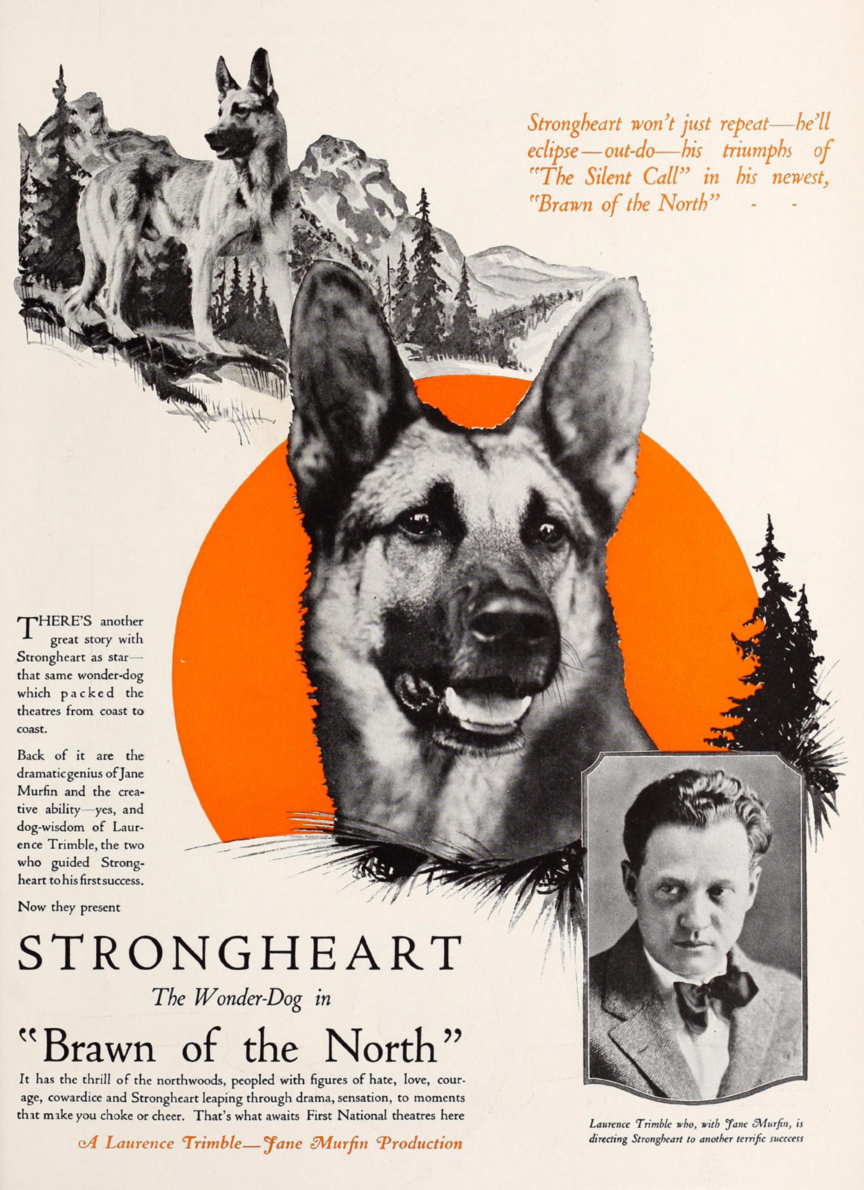
If we use set distance scores, the results change significantly. As previously mentioned, set distance is the calculation taken when duplicate words in a string are eliminated before the sorted distance is measured—in other words, the frequency of the words does not matter. Using this calculation can be helpful for comparing volumes of different lengths. We noticed that the volume *The Great Selection:*

"First National First" Season 1922–1923 appeared in the five highest-scoring pairs of volumes according to the sorted distance (see table 19.2). This was a promotional booklet generated by First National to market its upcoming productions to exhibitors, and it contained titles, names, advertising copy, and publicity text that appeared throughout US film industry trade papers.[19] When we looked at the volume that ranked as most similar to *The Great Selection, Motion Picture News* (July–August 1922), we saw that it contained twenty consecutive pages of the same promotional material contained in *The Great Selection*, including a full-page ad for a canine star, "Strongheart the Wonderdog in 'Brawn of the North'" (see figure 19.1).[20]

The next four highest-ranked set distances paired *The Great Selection* with other US trade papers: *Exhibitors Herald* [92.66], *The Film Daily* [92.3], *Exhibitors Herald* [91.9], and *Exhibitor's Trade Review* [91.74]. Blanketing the field to promote its films to exhibitors, First National produced its own promotional booklet in house and paid leading trade papers to carry the same promotions as advertisements. The set distance measurement helped us identify this same promotional text reappearing across multiple publications. [tabref 19.2]

Which publications scored the lowest? That is, which publications were the *least* similar to anything else in the corpus? Among the least similar pairings, we found that one text dominated the list across each of our measures. Our only Spanish-language magazine in the corpus, *Cine-Mundial*, appeared in all ten of the highest Euclidean distance pairings—signaling a lack of similarity. Similarly, the German-language trade paper *Der Kinematograph* was in eight of the lowest Levenshtein pairs and six of the highest cosine distances. All of this makes sense: the algorithms, just like a human reader, recognize that the patterns of language are different in those non-English language magazines.

For this reason, the most intriguing low scoring result was *Camera!*, an English-language US trade paper, which appeared in all the lowest-scoring pairs among the Levenshtein variants we computed. Founded in 1918, *Camera!* was the film industry's first weekly trade paper to consistently publish from Los Angeles.[21] *Camera!* cultivated creative workers on the West Coast as both its primary readers and advertisers; the paper provided industry news alongside ads taken out by aspiring writers and actors seeking employment on such productions.[22] While *Camera!* covered industry news related to First National in 1922, First National did not purchase advertising for its movies in *Camera!* since the customers it sought to reach (exhibitors) did not subscribe to the paper. The editors of *Camera!* addressed its readers as in-group members of a creative community and industry who were different from the exhibitor community and non-showbusiness people living in Los Angeles (see figure 19.2).[23] Over the next few years, several other film industry trade papers emerged in Los Angeles, competing against *Camera!* and eventually succeeding it. In 1922, though, *Camera!* occupied a unique position within the



Strongheart won't just repeat—he'll eclipse—out-do—his triumphs of "The Silent Call" in his newest, "Brawn of the North" - -

THERE'S another great story with Strongheart as star—that same wonder-dog which packed the theatres from coast to coast.

Back of it are the dramatic genius of Jane Murfin and the creative ability—yes, and dog-wisdom of Laurence Trimble, the two who guided Strongheart to his first success.

Now they present

STRONGHEART

The Wonder-Dog in

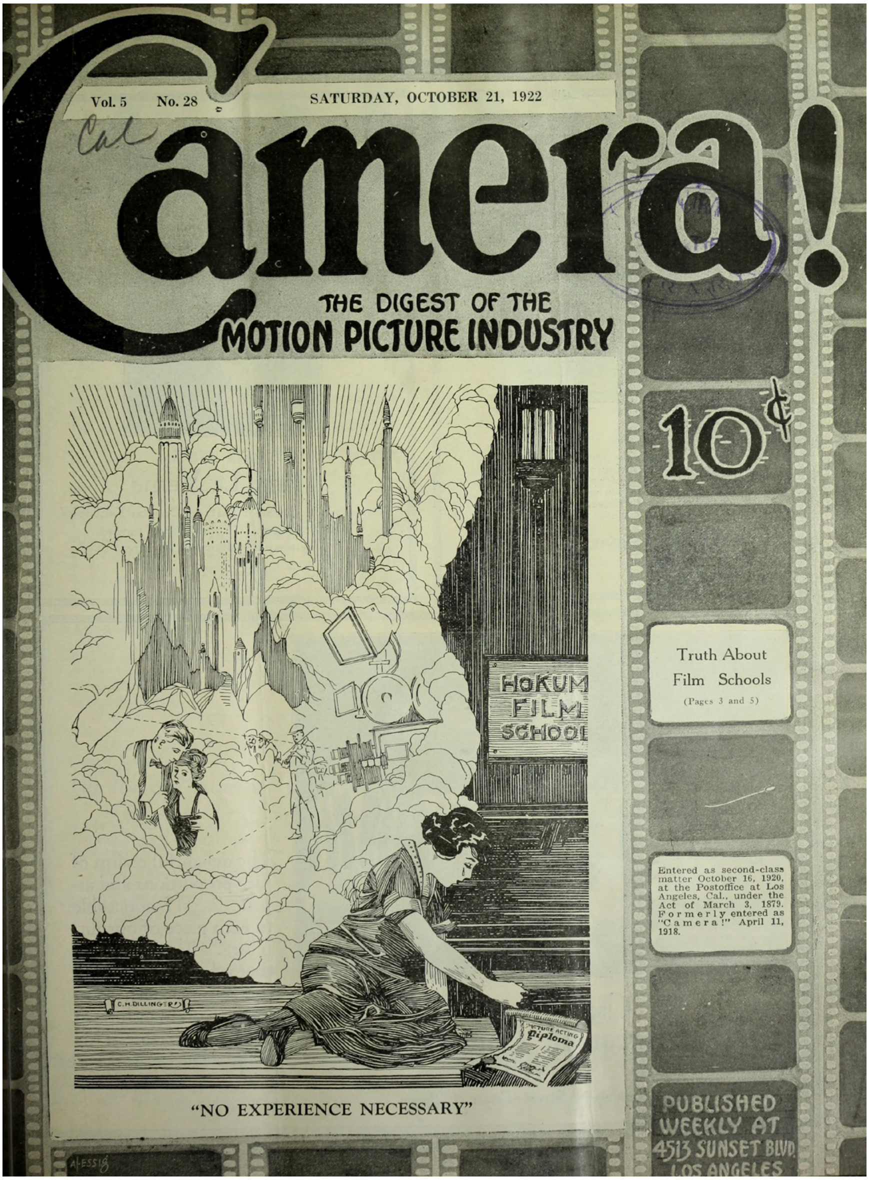
"Brawn of the North"

It has the thrill of the northwoods, peopled with figures of hate, love, courage, cowardice and Strongheart leaping through drama, sensation, to moments that make you choke or cheer. That's what awaits First National theatres here

A Laurence Trimble—Jane Murfin Production

Laurence Trimble who, with Jane Murfin, is directing Strongheart to another terrific succcess

FIGURE 19.1. Advertisement for "Strongheart the Wonderdog in 'Brawn of the North,'" *Motion Picture News,* 1922, https://lantern.mediahist.org/catalog/motionpicturenew26july_1003. The same ad was shared in *The Great Selection: "First National First" Season 1922–1923.*



Vol. 5 No. 28 SATURDAY, OCTOBER 21, 1922

Cal

Camera!

THE DIGEST OF THE MOTION PICTURE INDUSTRY

10¢

Truth About Film Schools

(Pages 3 and 5)

Entered as second-class matter October 16, 1920, at the Postoffice at Los Angeles, Cal., under the Act of March 3, 1879. Formerly entered as "Camera!" April 11, 1918.

"NO EXPERIENCE NECESSARY"

PUBLISHED WEEKLY AT 4513 SUNSET BLVD. LOS ANGELES

FIGURE 19.2. Cover of a 1922 issue of *Camera!*, criticizing profit-seeking film schools that misled the students who enrolled. https://lantern.mediahist.org/catalog/camera05unse_0571.

industry—one that, a full century later, our computer algorithms identified as distinct among other magazines that year.

CONCLUSIONS

Our algorithmic analyses suggest that there was a great deal of similarity among the four top weekly US trade papers oriented toward exhibitor readers (*Moving Picture World, Motion Picture News, Exhibitor's Trade Review*, and *Exhibitors Herald*). Despite *Exhibitors Herald*'s emphasis on its uniqueness and midwestern location, the paper's structure, use of language, and overlaps in advertising had much more in common with the New York weeklies than with the fan magazines, LA trade paper, and non-US publications in our corpus. Ultimately, *Exhibitors Herald* publisher Martin Quigley would acquire all three of those competing trade papers, rebranding the consolidated publication at the end of 1930 as *Motion Picture Herald*. Our computational analyses indicate a great deal of similarity across the papers in 1922. While these findings do not come as great surprises, they have enriched our understanding of both the historic magazines and the use of computational methods for large-scale text analysis.

What the computational results cannot tell us is what the trade papers meant to the people who originally created them, read them, and used them. The editors of *Motion Picture News* and *Exhibitors Herald*, William A. Johnston and Martin Quigley, respectively, each cultivated distinctive personas within the industry. They competed with one another for influence, power, and reader loyalty. Quigley emphasized his independence and integrity at every turn. Johnston successfully sued the editors of *Exhibitor's Trade Review* for libel when they attacked him in print. These aspects of the trade papers' histories require close reading of the magazines, as well as locating and analyzing sources outside of the papers themselves (e.g., court documents, private correspondences, audit bureau circulation reports). Algorithms are no substitute.

Yet computational methods do let us find and see things differently. Without the search indexing algorithms within Lantern, we would never have found many of the relevant pages that we read and analyzed closely with our eyes. The text similarity testing algorithms described in this chapter are, in part, attempts to achieve an even wider form of search—querying advertisements and strings of publicity text that reoccur across multiple publications, even when the specific words, phrases, and occurrences are not yet known. The promising results from the set distance rankings, with *The Great Selection: "First National First" Season 1922 –1923* scoring highest, have informed the work we are now undertaking in researching the reuse of text and graphics from Hollywood pressbooks across trade papers, fan magazines, and US newspapers. As we move forward, we are approaching the work with curiosity, humility, and the knowledge that no algorithmic results or score ranking will ever tell the whole story. We invite others to do the same, with the hope of locating many more stories to tell.

NOTES

[1] W. Stephen Bush, “Looking Forward,” *Exhibitor’s Trade Review*, June 16, 1917, 91.

[2] Leander Richardson to William A. Johnston (open letter), “One Trade Paper Enough,” *Variety*, December 28, 1917, 239, https://lantern.mediahist.org/catalog/variety49-1917-12_0400.

[3] Eric Hoyt, *Ink-Stained Hollywood: The Triumph of American Cinema’s Trade Press* (Oakland: University of California Press, 2022). Open access publication: https://doi.org/10.1525/luminos.122.

[4] For more on using distant reading to explore the MHDL, see Charles R. Acland and Eric Hoyt, eds., *The Arclight Guidebook to Media History and the Digital Humanities* (Falmer, UK: REFRAME/Project Arclight, 2016). Open access publication: http://projectarclight.org/book.

[5] We estimated that even the most basic algorithms could require several months of computing time to calculate. Although it may have been possible to decrease this time through code optimization and distributed computing strategies, we chose to select a much smaller corpus for our initial exploratory analyses.

[6] C. M. Hartman, quoted in “‘Herald Only’ Club Gains Six; Veteran and Newcomer Give Reasons for Joining,” *Exhibitors Herald*, March 29, 1924, 63, http://lantern.mediahist.org/catalog/exhibitorsherald18exhi_0_0073; George Rea letter to Exhibitors Herald, *Exhibitors Herald*, May 26, 1923, 69, http://lantern.mediahist.org/catalog/exhibitorsherald16exhi_0_0869.

[7] Shawn Graham, Ian Milligan, and Scott B. Weingart, *Exploring Big Historical Data* (London: Imperial College Press, 2015), 114.

[8] Ryan Cordell, “‘Q i-Jtb the Raven’: Taking Dirty OCR Seriously,” January 7, 2016. https://ryancordell.org/research/qijtb-the-raven-mla/.

[9] John R. Ladd, “Understanding and Using Common Similarity Measures for Text Analysis,” *Programming Historian* 9 (2020), https://doi.org/10.46430/phen0089.

[10] Michel Marie Deza, *Encyclopedia of Distances* (New York: Springer, 2014), 335–36.

[11] Vladimir I. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals,” *Soviet Physics Doklady* 10, no. 8 (February 1966): 707–10.

[12] Gonzalo Navarro, “A Guided Tour to Approximate String Matching,” *ACM Computing Surveys* 33, no. 1 (March 1, 2001): 31–88, https://doi.org/10.1145/375360.375365.

[13] Arturs Backurs and Piotr Indyk, “Edit Distance Cannot Be Computed in Strongly Subquadratic Time (Unless SETH Is False),” *arXiv*, August 15, 2017, https://doi.org/10.48550/arXiv.1412.0348.

[14] See table 19.1 for publication names and IDs.

[15] For more information on using application programming interfaces (APIs) to programmatically search and retrieve data from the Media History Digital Library, see https://lantern.mediahist.org/api.

[16] For pandas, see https://pandas.pydata.org/; for rapidfuzz, see https://maxbachmann.github.io/RapidFuzz/index.html.

[17] The compiled data sets are available for download at https://doi.org/10.5061/dryad.gtht76htc.

[18] C. M. Hartman, quoted in “‘Herald Only’ Club Gains Six; Veteran and Newcomer Give Reasons for Joining,” *Exhibitors Herald*, March 29, 1924, 63, http://lantern.mediahist.org/catalog/exhibitorsherald18exhi_0_0073; George Rea letter to Exhibitors Herald, *Exhibitors Herald*, May 26, 1923, 69, http://lantern.mediahist.org/catalog/exhibitorsherald16exhi_0_0869.

[19] Associated First National Pictures, Inc., *The Great Selection: “First National First” Season 1922–1923*, ca. 1922, https://lantern.mediahist.org/catalog/greatsel00firs_0005.

[20] “Associated First National Pictures, Inc. presents the first group of its Great Selection of Incomparable Fall Attractions” [advertisement], *Motion Picture News*, August 26, 1922, 961, https://lantern.mediahist.org/catalog/motionpicturenew26july_0987.

[21] Hoyt, *Ink-Stained Hollywood*, 111–13.

[22] “The Pulse of the Studio,” *Camera!*, April 13, 1919, 8, http://lantern.mediahist.org/catalog/camera1919losa_0034; “Where to Sell Your Scenarios,” *Camera!*, April 13, 1919, 13, http://lantern.mediahist.org/catalog/camera1919losa_0039.

[23] Hoyt, *Ink-Stained Hollywood*, 113.

PUBLICATIONS REFERENCED

The American Cinematographer
Camera!
Canadian Moving Picture Digest
Cine-Mundial
Cinéa
Der Kinematograph
Exhibitors Herald
Exhibitor's Trade Review
The Film Daily
The Film Renter and Moving Picture News
The Great Selection: "First National First" Season 1922–1923
Harrison's Reports
Motion Picture News
The Motion Picture Studio
Moving Picture World
Paramount Pep
Photoplay
The Picturegoer
Shadowland
Tess of the Storm Country (United Artists Pressbook)
Universal Weekly
Variety

BIBLIOGRAPHY

Acland, Charles R., and Eric Hoyt, eds. *The Arclight Guidebook to Media History and the Digital Humanities*. Falmer, UK: REFRAME/Project Arclight, 2016. Open access publication: http://projectarclight.org/book.

Backurs, Arturs, and Piotr Indyk. "Edit Distance Cannot Be Computed in Strongly Subquadratic Time (Unless SETH Is False)." *arXiv*, August 15, 2017. https://doi.org/10.48550/arXiv.1412.0348.

Cordell, Ryan. "'Q i-Jtb the Raven': Taking Dirty OCR Seriously." January 7, 2016. https://ryancordell.org/research/qijtb-the-raven-mla/.

Deza, Michel Marie. *Encyclopedia of Distances*. New York: Springer, 2014.

Graham, Shawn, Ian Milligan, and Scott B. Weingart. *Exploring Big Historical Data*. London: Imperial College Press, 2015.

Hoyt, Eric. *Ink-Stained Hollywood: The Triumph of American Cinema's Trade Press*. Oakland: University of California Press, 2022. Open access publication: https://doi.org/10.1525/luminos.122.

Ladd, John R. "Understanding and Using Common Similarity Measures for Text Analysis." *Programming Historian* 9 (2020). https://doi.org/10.46430/phen0089.

Levenshtein, Vladimir I. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." *Soviet Physics Doklady* 10, no. 8 (February 1966): 707–10.

Navarro, Gonzalo. "A Guided Tour to Approximate String Matching." *ACM Computing Surveys* 33, no. 1 (March 1, 2001): 31–88. https://doi.org/10.1145/375360.375365.