2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #32907

Desperately Seeking Standards: Using Text Processing to Save Your Time

Ms. Halle Burns, University of Nevada, Las Vegas

Halle Burns is the Data Librarian at the University of Nevada, Las Vegas University Libraries. In addition, she is certified as an instructor with The Carpentries. Her current research interests include data literacy, digital humanities, and improving the accessibility of data science and technology education.

Ms. Susan B. Wainscott, University of Nevada, Las Vegas

Susan Wainscott is the Engineering Librarian for the University of Nevada, Las Vegas University Libraries. She holds a Master of Library and Information Science from San Jose State University and a Master of Science in Biological Sciences from Illinois State University. As liaison librarian to several departments at UNLV, she teaches information literacy for many students, provides reference assistance to the campus and community, and maintains the collection in assigned subject areas. Her research interests include information literacy instruction and assessment, the notion of threshold concepts, the effect a student's emotional state has on their learning, and improving access to technical literature.

Desperately seeking standards: using text processing to save your time

Abstract

Purpose/Hypothesis

We aim to analyze our standards-use, interlibrary loan, and document-delivery-request data on a more regular basis to inform collections management decisions. However, manually searching for standards titles within interlibrary loan and document-delivery-request data is time consuming and unlikely to occur on a regular basis. We were also interested in a method that could be applied to large blocks of text, such as theses and dissertations.

Design/Method

To detect the presence of engineering standards and other standards documents in tabular datasets as well as in large blocks of text, the first step was to develop a regular expression, using Python in Jupyter Notebooks. Regular expressions (or regex), used for text processing and querying, identifies patterns within written text. This pattern was tested to match a series of standards, within sample text that included known standards such as ANS 10.5-2006. In addition, it was checked against words and phrases it should not match against, including web addresses and mathematical equations. As a proof of concept, the text processing code was evaluated against a collection of sample pdf dissertations, one of which included standards documents in the text and references list.

As there are many iterations of what a standard can be called, we were unable to restrict the regex matching criteria any further. This means that false-positives appeared, such as the “state name and zip code” combination, report numbers, and chemical formulas. To help identify results from false-positives, we

expanded the regular expression to also pull words surrounding the match, giving context to the results. This does not prevent the false-positives but allows us to quickly distinguish a false-positive from an actual match.

Once the pattern was identified, it was then applied (using Python and the pandas package) to compiled spreadsheets to identify standards in tabular collections data. We compared these results to an earlier manual search performed on the same data set. We also tested the text processing method on a set of dissertations.

Results

The new method required 25% less time to complete, and the outcomes were similar. While we predicted that more standards would be located using the text processing method compared to a manual search, the text processing method missed three standards that were previously detected, and located one standard that had not been previously detected. The regular expression also successfully detected standards documents mentioned in large blocks of text.

Conclusion

We developed and assessed an open source text processing method to flag potential standards mentioned in text and tabular datasets. This method is a substantial improvement over manual searching, providing similar results in a quarter of the time. The new method requires less than half a standard workday to analyze 10,000 interlibrary loan or document delivery requests. Our pilot test of the method on large blocks of text shows that it will also detect standards used in materials that are not regularly indexed for citations such as theses and dissertations, as well as technical reports and other gray literature.

Introduction

Engineering and other standards documents are potentially useful to many library patrons and may be integral to some research or design projects [1]. Understanding patrons' needs for these items would be informative for collection management. However, collection management decisions are often informed by analyses that are better suited to more commonly collected and used items, such as monographs and journal articles. These analyses may compare the use of existing collections to patron requests for additional similar materials [2], and analyses of local researcher citation practices [3] can be performed.

Interlibrary loan (ILL) data have been used by librarians to provide useful insights into what materials patrons are seeking and in which subject areas a library's collections may need more depth [4]. ILL request data has also been used in combination with circulation, electronic resource use counts, and turnaway data (counts of user attempts to access electronic resources that result in a failure to access full-text) to inform collection development decisions, primarily for monographs and journal articles [5], [6] but also for other formats such as standard documents [7]. ILL data has additionally been used to gauge the value of an improved discovery system (federated search tool used by libraries to provide patrons a single search box that operates simultaneously across multiple collections and databases) and link resolvers (software tool that seeks full-text access to items across multiple collections) to direct patrons to library-provided items [8], [9].

There are several impediments to applying such analyses to standards and other forms of gray literature, such as grant reports or other technical reports and manuals which are often produced by organizations where publishing is not their primary mission [10]. Unlike a publishing-focused organization, these gray literature publishers may lack the infrastructure and/or motivation to issue common identifiers such as ISSN, ISBN, or DOI to their materials. The gray literature publishers are also unlikely to participate in indexing of their materials by academic and library databases or discovery services or to provide data compatible with link resolver systems used in many modern library catalogs or discovery systems [1]. While unique identifiers for each publication can be applied by the gray literature producers, those identifiers may be unfamiliar to library workers or patrons. In addition to this, academic libraries are inconsistent in how or even if they catalog the standards within their collections [11], [12].

Thus, standards are difficult for patrons to locate in collections because they are seldom indexed or cataloged, and patrons may submit an ILL request for items that may or may not be in the local collection. However, completing an ILL request form for standards may also be difficult due to unfamiliar identifiers assigned by publishers, and the limitations of ILL request forms designed for books and journal articles. These difficulties result in wide variability in how patrons may complete ILL requests for the same item. Thus, we can infer that standards are both infrequently and imperfectly requested. Seeking standards requests within ILL datasets can be like looking for inconsistently-formed needles in a haystack.

We sought to develop a text processing method to detect mentions of engineering standards and other standards documents in interlibrary loan datasets, as well as in large blocks of text, such as full-texts of theses and dissertations or within other publications. Using a portion of the same dataset analyzed by Wainscott and Zweircan [7], we applied a text processing method to identify probable requests for standards documents and compared our results to theirs. We also tested this method on five full-text dissertation documents.

Method

We analyzed the method described by Wainscott and Zwiercan [7] and their unpublished data to get more detail and the approximate time spent on each step. The data set included four years of ILL requests, a total of 38,230 records, or requested items. We describe their method as a “Scan then Examine” set of steps. The Scan step refers to the initial visual pattern search of the dataset to determine what could be a potential standard. The pattern was based upon the detection of an acronym with or without subsequent numbers. The Examine step references the more intensive process of reading the text of the remaining records for meaning. This determined whether each requested item was likely a standard. We sought to replace the Scan step with text processing. Below we show these steps in a block quote from Wainscott and Zwiercan's paper, and also identify data preparation steps.

> **(Preparation)** We obtained the ILL request data for 2012-2018 and searched each year for items containing the string standard. **(Scan)** Those items were then evaluated to determine if they were the standard format type. **(Preparation)** Remaining items (not containing the word standard) were then filtered to exclude titles with the string journal or proceedings, **(Scan)** and the resulting set were manually scanned for acronyms in all capital letters within fields with data entered by patrons. **(Examine)** Entries with all capital letter acronyms were flagged and further assessed for the standard format type. All items determined to be standards were then reviewed to determine if the request was fulfilled by ILL staff by any means. [7]

To replace the Scan step, Author 1 first developed a regular expression, using Python in Jupyter Notebooks (code, documentation, and de-identified dataset can be found on Author 1's GitHub: https://github.com/hburns2/desperately-seeking-standards). Regular expressions (or regex), used for text processing and querying, identify patterns within written text. A regular expression, as defined by the organization Library Carpentries, "is a method of using a sequence of characters to define a search to match strings "with 'string' being "a contiguous sequence of symbols or values" [13]. This specific expression was created to identify any string of words that matched:

1. Starting with two or more capital letters
2. Followed by any character except specific punctuation
3. Followed by zero or more capital letters

4. Followed by one or more numerical characters
5. Followed by zero or one instance of any character except specific punctuation
6. Followed by zero or more numerical characters

The regular expression underwent several iterations, as new patterns were discovered that it could be built to match. The final expression was tested to match a series of standards, including GO-102001-1287, DOE-2, and ASTM E1641. In addition, it was checked against words and phrases it should not match, including web addresses and mathematical equations.

Once the pattern was identified, it was then applied (using Python and the pandas package) to compiled spreadsheets to identify standards in ILL requested library items. In addition to using the regex pattern Author 1 used pandas to identify any use and variant of the word *standard* in the Photo Journal Title, Photo Article Title, and Loan Title columns of the ILL dataset. When compared against the original ILL dataset, consisting of over 38,000 records, the regex pattern identified 486 possible record matches while the search for forms of *standard* identified 169. Two records were located by both, for a total of 653 records identified by the text processing method (Table 1). This text processing method took about 15 minutes to set up and run.

The Preparation steps used by Wainscott and Zwiercan [7] were replicated and timed, taking a total of 15 minutes. Unpublished calendar data from Wainscott (Author 2 of the present analysis) were then used to estimate the time spent on the initial manual Scan portion of their process, shown in Table 1. We determined that 16 hours (960 minutes) were spent on this Scan step.

Table 1.

Comparison Of Standards Search Methods, Time Spent Per Search, And Total Records Evaluated Per Step

Search methods steps	**Wainscott and Zwiercan**	**Text Processing**
Preparation for processing	15 minutes (38,230 records)	15 minutes (38,230 records)
Initial manual Scan	960 minutes (29,289 records)	N/A
Examine	35 minutes (235 records)	97 minutes (653 records)
Total time	**1,010 minutes**	**112 minutes**

We next replicated the Examine step used by Wainscott and Zweircan [7] to generate a time estimate. To address potential subjectivity in determining which requested items are standards, we created a new data table containing the 653 identified records (from the new text processing method) and the 235 records (evaluated in the Wainscott and Zwiercan Examine step) and noted which method had detected each record. Fifty-three duplicate records (records found by both the text processing and the original Wainscott and Zweircan methods) were then merged and the annotation updated to indicate that both methods detected the record. This yielded a combined dataset with 835 records. Notation of which method(s) had detected the records were hidden from view, and Author 2 examined each of the 835 records to determine if the request was for a standard. This evaluation included both quick determinations that some records were not requests for standards, and occasional searching for requested items in the WorldCat database or using an internet search tool to locate the item on a publisher's website. Forty of the total Examined records were determined to be requests for standards documents. The replicated Examine step in this new combined dataset took 2 hours, for an average time of 9 seconds per record. We extrapolated that the likely amount of time spent on the Wainscott and Zwiercan Examine step was 35 minutes.

As a proof of concept, the text processing code was also evaluated against five full-text formatted dissertations from the university library’s institutional repository. Using the Python Tika package and its library, Parser, we were able to convert the file formats of these dissertations from PDFs to machine-readable text. This enabled us to then apply our regular expression to the collection of documents. On top of pulling the text matches, we added an element to the code that would also extract the entire line in which the result appears. This provides context, which will ultimately assist us in determining whether the match is true or a false positive. We have not yet analyzed our results from this supplementary effort.

Results

When the compiled dataset of 835 probable records was examined by Author 2 to determine if they were standards, 16 records previously identified as standards by Author 2 in her role during the earlier study [7] were not determined by the current analysis to be standards. They were instead discovered to be a variety of other gray literature formats. Several of these 16 were requests for the American Association of State Highway and Transportation Officials Load and Resistance Factor Design (AASHTO LRFD) Bridge Design Specifications, a few were for American Society of Civil Engineers (ASCE) Manuals of Practice, or for other materials published by professional societies such as the National Criminal Justice Reference Service (NCJRS). Of the 40 requested items determined to be standards documents in the current analysis, both methods were similar in efficacy, with 36 standards requests detected by both methods (Table 2).

Table 2.

Standards Requests Detected By Each Method, Or By Both

Standards Requests	**Both**	**Manual Scan Only**	**Text Processing Only**
40	36	3	1

Both methods yielded false positives. For instance, the regular expression pattern does not pick up on indications of standards only composed of capital letters. It will not pick up on “ASTM Engine Coolants and Related Fluids,” which can be assumed to include engineering standards, but does recognize “ASTM C94 / C94M - 17a” as a match. It will also pick up records that are clearly not standards, such as a request for a particular model Texas Instruments brand calculator, or the graphic novel "Toppu GP 1." The method used by Wainscott and Zwiercan also yielded false positives, including the 16 requests discussed above.

As there are many iterations of what a standard can be called, we were unable to restrict the regex matching criteria any further. This means that false positives appeared, such as state abbreviation, zip code combinations, and chemical formulas. For the full-text experiment, to help identify results from false positives, we expanded the regular expression to also pull words surrounding the match, giving context to the results. This does not prevent the false positives but allowed us to quickly distinguish a false positive from an actual match.

The text processing method did result in three false negatives. The three records missed by the new method provide some insight into potential improvements to the text processing method. The first missed record, a 2015 request, had a form of the word *standard* in the Loan Author field. We could expand the columns searched in the ILL dataset to include this column. The second missed record was a 2018 request that did not have any alphanumeric unique identifier, nor did the request include any form of the word *standard* in the request fields. It did have "code" in the Loan Title field. This may be an additional text string to search for with pandas.

The third missed record was a 2017 request that included "ASTM" within the request fields, but no use of the strings *standard* or *code* in the data, and no alphanumeric unique identifier for the requested standard was included in the request. ASTM is a commonly-requested engineering-standard developing organization. This was likely selected during the manual scan method for further examination based upon evaluator familiarity with the name of the organization.

There was one standard request detected only by the text processing method. This request was for an American Association of Museums (AAM) document with a Loan Title of "National Standards". It should have been selected by Wainscott and Zwiercan [7] for further investigation in their Scan step. This demonstrates the potential for human error in a time-consuming, manual scan of a large number of records.

While the text processing method described in this paper is not more accurate than the earlier study's fully manual methods, it does result in a decrease in time needed to complete a full analysis. The numbers of records scanned and examined each affected the amount of time required to complete the analysis. Wainscott and Zwiercan's Scan step for four years of ILL data took approximately 16 hours, or about four hours for each year’s data. This step was eliminated with the text processing method. The text processing method presented here yielded a set nearly three times larger (653 records) than the set that Wainscott and Zwiercan [7] identified for the Examine step (235 records). However, the increase in time for the Examine step was only slightly over one hour (Table 1).

Overall, the text processing method presented in this paper is approximately 9x faster than the Wainscott and Zwiercan method. For a library with a similar ILL request load of about 10,000 ILL requests per year, using this new text processing method would take about an hour to complete on an annual basis.

Regarding the full-text dissertations and theses, we did not fully analyze the results of using regex with the dissertation texts. However, the proof of concept was successful in locating several strings of text that appear to be standards, as well as several strings that clearly are not. An example of a false-positive includes the string "CHAPTER 1." Similar strings would be easy to exclude in further analyses.

Conclusion

Standards documents might be quite important to researchers and students, and there may be no suitable substitute for a particular needed standard document if it is a requirement of a researcher’s project or design. Standards are often expensive for a patron to obtain on their own, and this may be an unrealistic expectation for students working on a class project. Academic libraries in particular have an interest in understanding the use of standards by their patrons, to inform collection development and budgets for ILL.

The text processing method we have developed can be used with open-source software with a minimal time investment, allowing for more frequent analysis of ILL data, or for citation analysis of local authors' publications and student work. The time and energy saved by this text processing method allow for a more nuanced investigation of probable standards requests and can search for mentions of standards within large text documents. This will enable further investigation of standards documents within theses, dissertations, and other gray literature documents that are less likely to be included in bibliometric analyses.

Literature Cited

[1] M. Phillips, “Standards Collections: Considerations for the Future,” *Collect. Manag.*, vol. 44, no. 2–4, pp. 334–347, Jul. 2019, doi: 10.1080/01462679.2018.1562396.

[2] J. R. Matthews, *The Evaluation and Measurement of Library Services, 2nd Edition*. ABC-CLIO, 2017.

[3] S. Edwards, “Citation analysis as a collection development tool: A bibliometric study of polymer science theses and dissertations,” *Ser. Rev.*, vol. 25, no. 1, pp. 11–20, Jan. 1999, doi:

10.1080/00987913.1999.10764479.

[4] E. J. Schares, “ILL Communication : Analyzing five years of Iowa State University’s print Interlibrary Loan requests,” in *American Society for Engineering Education 126th Annual Conference & Exposition*, 2019, p. no page numbers.

[5] G. P. Barton, G. E. Relyea, and S. A. Knowlton, “Rethinking the subscription paradigm for journals: Using interlibrary loan in collection development for serials,” *Coll. Res. Libr.*, vol. 79, no. 2, pp. 279–290, 2018, doi: 10.5860/crl.79.2.279.

[6] M. Smith, “Living in denial: The relationship between access denied turnaways and ILL requests,” *Ser. Libr.*, vol. 75, no. 1–4, pp. 31–41, 2019, doi: 10.1080/0361526X.2019.1593914.

[7] S. B. Wainscott and R. J. Zwiercan, “Improving access to standards,” *ASEE Annu. Conf. Expo. Conf. Proc.*, vol. 2020-June, p. 13, 2020, doi: 10.18260/1-2--34790.

[8] P. P. Frank and R. L. Bothmann, “Assessing undergraduate interlibrary loan use,” *J. Interlibrary Loan, Doc. Deliv. Electron. Reserv.*, vol. 18, no. 1, pp. 33–48, 2008, doi: 10.1300/J474v18n01_05.

[9] L. R. Musser and B. M. Coopey, “Impact of a discovery system on interlibrary loan,” *Coll. Res. Libr.*, vol. 77, no. 5, pp. 643–653, 2016, doi: 10.5860/crl.77.5.643.

[10] The New York Academy of Medicine, “What is Grey Literature? | Grey Literature Database.” http://www.greylit.org/about (accessed Mar. 03, 2021).

[11] K. A. Kozak, “Standards, Standards: Where might you be?,” in *2014 ASEE North Midwest Section Conference*, 2014, pp. 1–8, doi: 10.17077/aseenmw2014.1039.

[12] J. Gelfand, I. Lawal, J. Powell, and A. Rauh, “Collecting Standards for Scholarship, Organization, Industry, and Innovation,” 2018, [Online]. Available: https://surface.syr.edu/sul/185/.

[13] S. Laster *et al.*, “LibraryCarpentry/lc-data-intro: Library Carpentry: Introduction to Data (Regular Expressions). (Version v2019.06.1),” *Zenodo*, [Online]. July 2019. doi: 10.5281/zenodo.3264946.