May, 2017

Text-Mining PubMed Search Results to Identify Emerging Technologies Relevant to Medical Librarians.pdf

Skye Bickett
P. F. Anderson
Joanne Doucette
Pamela Herring
Andrea Kepsel, et al.

Text-Mining PubMed Search Results to Identify Emerging Technologies Relevant to Medical Librarians

P. F. Anderson[1] <pfa@umich.edu>; Skye Bickett[2]; Joanne Doucette[3]; Pamela Herring[4]; Andrea Kepsel[5]; Tierney Lyons[6]; Scott McLachlan[7]; Carol Shannon[1]; Lin Wu[8]
University of Michigan-Ann Arbor; 2) Georgia Campus-Philadelphia College of Osteopathic Medicine; 3) MCPHS University, Boston; 4) University of Central Florida College ofMedicine; 5) Michigan State University; 6) Cerebros Medical Systems, Jessup, PA; 7) Ruskin College, Oxford; 8) University of Tennessee, Memphis

Objectives

The Emerging Technologies Team, part of the Medical Library Association (MLA) systematic review (SR) projects, conducted a pilot study to identify emerging technologies relevant to medical librarians. The team analyzed results from its previously reported PubMed Search filter using text mining to identify patterns, themes, and trends important to the practice of medical librarianship and the communities we support.

Methods

We began by establishing a common competency base through custom training sessions from higher education data-mining experts. Next, the team 1) reviewed and finalized the emerging technologies PubMed search strategy created for the project; 2) exported the data; 3) used automated tools to clean extraneous data from the data set; and 4) tested the data by running preliminary text-mining scans. Steps 3 and 4 were repeated to refine and focus the results. Tools such as GREP, R, FLink, pubmed.mineR were evaluated and tested for data export and cleaning, with the ultimate choices settling on a combination of EndNote, Voyant, OpenRefine for data cleaning, and Voyant, OpenRefine/GoogleRefine and AntConc for analysis.

Data Cleaning

After running the finalized search strategy in PubMed, the resulting list was exported from the database in the MEDLINE format, creating a TXT file. To support the proposed text mining analysis of this dataset using Voyant or OpenRefine, a CSV file needed to be built. FLink, an NLM product developed to create CSV files, was used initially. Unfortunately, the program was only able to handle 10,000 records and could not produce a CSV file that included abstracts. To create the appropriate CSV file, results were exported from PubMed as TXT file, imported to EndNote, deduplicated within EndNote, then exported using a custom output style created by the team. The resulting CSV file of 162,339 PubMed records was downloaded to Excel where all fields except for PMID, Title, Abstract and MeSH or Keywords were deleted. The remaining content was cleaned by removal of punctuation (using nested SUBSTITUTE functions) and changing all text to lowercase (using the LOWER function). Considerations during punctuation removal included separation of MeSH headings and subheadings by removal of the "/" character, the importance of "." character in numerical values, and the decision of whether or not to keep numeric data.

.

Analysis: Challenges & Solutions

Challenges:

1. FLink exports as CSV directly from PubMed, but only permits export of 10,000 records, no abstracts.
2. Inadequate (under-powered) hardware.
3. Large file size created challenges with opening file and file conversion.
4. Unable to install current version of text mining software (OpenRefine).
5. IT policies (blocking) and support at some institutions.

Solutions:

1. Export full records, import to Endnote, use custom filter for initial cleaning, export to CSV file.
2. Do you have access to more powerful computers elsewhere?
3. Break file into smaller chunks for cleaning; pool for final analysis.
4. Upgrade computer to have more memory, or use more powerful computer elsewhere.
5. Ensure good technical support and administrative backing for project.

Process: Software & Technology

1st stage: Voyant was used to identify words/word cloud to create custom stoplist. Stop word list was created by the team leader, and peer reviewed within the team. Collocations used to identify major tech concepts from word cloud concepts. Additional visualizations to refine understanding of top three tech concepts.

2nd stage: OpenRefine will be used to open projects & expand analysis. (*OpenRefine challenges*: with full dataset, it stalls out in the project creation stage on team desktops. Works in Chrome, but not in Opera or Safari.)

3rd stage: AntConc will be used for a deep dive into the specific terms/phrases of interest to discover their context and related concepts.

Text Mining Images

All images were created through the Voyant-Tools analysis: <http://voyant-tools.org/>

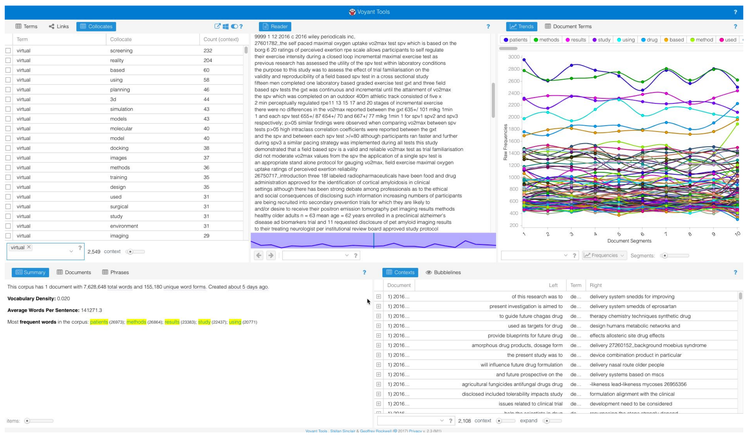


Figure 1: Word Cloud of Whole Corpus

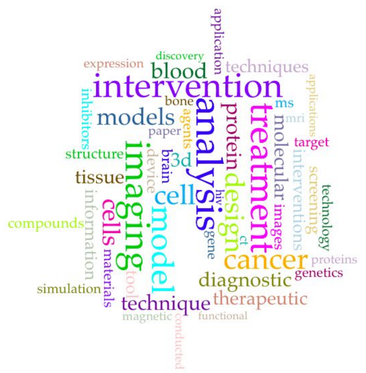


Figure 2: Top Three Technologies Identified in Analysis

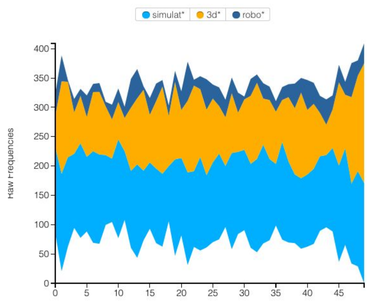
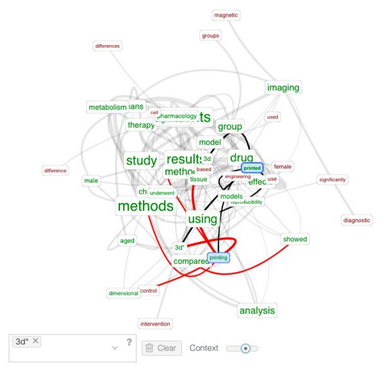


Figure 3: 3D Printing in Context



Results

The finalized search strategy results in a five- year file with 162,339 records. Deduping in EndNote resulted in 162,221 records. We tested the analysis with 5-year [162,221], 3-year [107,531], and 1-year sets for each of the five years. For this poster we tested the analysis process with the single year set from 2016 [35,535].

In our initial project planning, we had identified five main areas of interest (technology, information, public health, education, and the body). This analysis made clear additional clusters of interesting content, especially new methodologies (e.g., big data and data visualization) and emerging interdisciplinary trends (such as precision medicine). As those had not been included in the original planning, this showed the potential benefit of text mining for discovering unknown areas of relevance.

The primary areas of the body which were strongly represented in the data included blood, bone, brain, and urine. Related concepts which were strongly represented in the data set included cancer, diagnostics, treatment, and biomarkers.

The three top technologies that arose from the text mining process were robotics, simulations, and 3D technologies, especially 3D printing. All three were being used most heavily in surgery. Simulations were also prominent in education/training.

Next Steps & Recommendations

Voyant was used for basic analysis to identify big concepts from 2016, yet we can dig deeper with additional tools to generate unknown items. Moving forward, we will extend the years of analysis for trends and patterns and continue gathering more data. We will continue the text mining process, using visualization techniques such as Google Refine and OpenRefine for analysis in context and AntConc for concordance and fringe concepts. We will then publish the results.

Once one has a dataset, the dataset itself can be useful to look for trends in specific areas, such as surgery or education, potentially in response to areas of interest within the library's target audience.

Sources / Resources

BIBLIOGRAPHY

Higgins JPT, Deeks JJ, (eds.) 2011. Selecting studies and collecting data, London: The Cochrane Collaboration.

Mane KK, Borner K. 2004. Mapping topics and topic bursts in PNAS. Proceedings of the National Academy of Sciences 101(suppl. 1), 5287-5290.

Mikova N. 2016. Recent trends in technology mining approaches: Quantitative analysis of GTM Conference Proceedings. In: Daim TU, Chiavetta D, Porter AL, Saritas O (eds.) Anticipating future innovation pathways through large data analysis. Cham, Switzerland: Springer International Publishing.

Porter AL, Cunningham SW. 2005. Tech mining: Exploiting new technologies for competitive advantage, Hoboken, NJ, John Wiley & Sons, Inc.

Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou PP, Guyatt GH. 2011. Interpreting results and drawing conclusions. In: Higgins JPT, Green S (eds.) Cochrane handbook for systematic reviews of interventions Version 5.1.0 (updated March 2011). London: The Cochrane Collaboration.

Stevens A, Milne R, Lilford R, Gabbay J. 1999. Keeping pace with new technologies: Systems needed to identify and evaluate them. BMJ 319, 1291-3.

RESOURCES

Endnote: endnote.com/
Voyant: https://voyant-tools.org/
OpenRefine: http://openrefine.org/
AntConc: www.laurenceanthony.net/software/antconc/

Find us

MLASR6 Google Plus Community: goo.gl/RxtOFg

The Medical Library Association initiated a large systematic review project to assess the level of evidence available to support the profession and practice of medical librarianship in several very important questions. Team 6 has been assigned to explore this topic:

The explosion of information, expanding of technology (especially mobile technology), and complexity of healthcare environment present medical librarians and medical libraries opportunities and challenges. To live up with the opportunities and challenges, what kinds of skill sets or information structure do medical librarians or medical libraries are required to have or acquire so as to be strong partners or contributors of continuing effectiveness to the changing environment?