#### Philadelphia College of Osteopathic Medicine

From the SelectedWorks of Skye Bickett

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.



Available at: https://works.bepress.com/

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

P. F. Anderson<sup>1</sup> <pfa@umich.edu>; Skye Bickett<sup>2</sup>; Joanne Doucette<sup>3</sup>; Pamela Herring<sup>4</sup>; Andrea Kepsel<sup>5</sup>; Tierney Lyons<sup>6</sup>; Scott McLachlan<sup>7</sup>; Carol Shannon<sup>1</sup>; Lin Wu <sup>8</sup>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.

# Text Mining Images

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

🥸 Voyant Tools										
🌐 Terms 🛛 🔩 Links	s Ellocates	₿∎0?	Reader	? Irends I Document Terms						
Term	Collocate	Count (context)	9999 1 12 2016 c 2016 wiley periodicals inc, 27601782 the self paced maximal awaren untake vo2max test say which is based on the	patients emethods ersults study ousing odrug based omethod						
virtual	screening	232	borg 6 20 ratings of perceived exertion rpe scale allows participants to self regulate							
virtual	reality	204	their exercise intensity during a closed loop incremental maximal exercise test as	3000-						
virtual	based	60	the purpose to this study was to assess the effect of trial familiarisation on the	2800-						
virtual	using	58	validity and reproducibility of a field based spv test in a cross sectional study	2600-						
virtual	planning	46	based spv tests the gxt was continuous and incremental until the attainment of vo2max	2400						
virtual	3d	44	the spy which was completed on an outdoor 400m athletic track consisted of five x	2400-						
virtual	simulation	43	there were no differences in the vo2max reported between the gxt 635+/ 101 mlkg 1min	2200-						
virtual	models	43	1 and each spv test 655+/ 87 654+/ 70 and 667+/ 77 mlkg 1min 1 for spv1 spv2 and spv3	2000-						
virtual	molecular	40	tests p>05 high intraclass correlation coefficients were reported between the gxt							
virtual	model	40	and the spv and between each spv test >/=80 although participants ran faster and further							
virtual	docking	38	demonstrated that a field based spv is a valid and reliable vo2max test as trial familiarisation							
virtual	images	37	did not moderate vo2max values from the spv the application of a single spv test is	Mar 1400-						
virtual	methods	36	an appropriate stand alone protocol for gauging vo2max, field exercise maximal oxygen uptake ratings of perceived exertion reliability	1200-						
virtual	training	35	26750717, introduction three 18f labeled radiopharmaceuticals have been food and drug	1000-						
virtual	design	35	administration approved for the identification of cortical amyloidosis in clinical settings although there has been strong debate among professionals as to the ethical							
virtual	used	31	and social consequences of disclosing such information increasing numbers of participants							
virtual	surgical	31	are being recruited into secondary prevention trials for which they are likely to and/or desire to receive their positron emission tomography pet imaging results methods							
virtual	study	31	healthy older adults $n = 63$ mean age = 62 years enrolled in a preclinical alzheimer's	400 -						
virtual	environment	31	disease ad biomarkers trial and 11 requested disclosure of pet amyloid imaging results to their treating neurologist per institutional review board approved study protocol	200 -						
virtual	imaging	29		1 2 3 4 6 6 1 8 9						
tual ×	2,549 context •			Document segments						

# 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.



# 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.

Summary E Documents E Phrases	?	E Contexts	Bubblelines		
s corpus has 1 document with 7,628,648 total words and 155,180 unique word forms. Created about 5 days ago.		Document	Left	Term	Right
cabulary Density: 0.020		1) 2016	of this research was to	de	delivery system snedds for improving
erage Words Per Sentence: 141271.3	n 🗄	1) 2016	present investigation is aimed to	de	delivery system smedds of eprosartan
	E	1) 2016	to guide future chagas drug	de	therapy chemistry techniques synthetic drug
st trequent words in the corpus: patients (25973); methods (25864); results (23383); study (22437); using (20771)	Œ	1) 2016	used as targets for drug	de	design humans metabolic networks and
	E	1) 2016	provide blueprints for future drug	de	effects allosteric site drug effects
	Œ	1) 2016	amorphous drug products, dosage form	de	delivery 27260152, background moebius syndrome
	Œ	1) 2016	the present study was to	de	device combination product in particular
	E	1) 2016	will influence future drug formulation	de	delivery nasal route older people
	E	1) 2016	and future prospective on the	de	delivery systems based on mscs
	E	1) 2016	agricultural fungicides antifungal drugs drug	de	-likeness lead-likeness mycoses 26955356
	E	1) 2016	disclosed included tolerability impacts study	de	formulation alignment with the clinical
	E	1) 2016	issues related to clinical trial	de	development need to be considered
		1) 2016	halp the existing days	do	coursesing the stope strengly depend

# Figure 1: Word Cloud of Whole Corpus



# Figure 2: Top Three Technologies Identified in Analysis

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.

# Analysis: Challenges & Solutions

Solutions:

# Challenges:

- FLink exports as CSV directly from PubMed, but only permits export of 10,000 records, no abstracts.
- 2. Inadequate (under-powered) hardware.
- 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.
- 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.

# e simulat\* • 3d\* • robo\*

# Figure 3: 3D Printing in Context



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

# 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

# **Process: Software & Technology**

1<sup>or</sup> 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



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?