Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Ann Eye Sci. Author manuscript; available in PMC 2022 October 04.

Published in final edited form as:

Ann Eye Sci. 2022 March ; 7: . doi:10.21037/aes-21-29.

RegenX: an NLP recommendation engine for neuroregeneration topics over time

Shaan Khosla[1], **Leila Abdelrahman**[2], **Joseph Johnson**[3], **Mohammad Samarah**[4], **Sanjoy K. Bhattacharya**[2]

[1]New York University, Center for Data Science, New York, NY, USA

[2]Department of Ophthalmology & Miami Integrative Metabolomics Research Center, University of Miami, Bascom Palmer Eye Institute, Miami, FL, USA

[3]Department of Marketing, University of Miami, Miami Herbert Business School, Miami, FL, USA

[4]Carroll University, Wisconsin, WI, USA

Abstract

Background: In this investigation, we explore the literature regarding neuroregeneration from the 1700s to the present. The regeneration of central nervous system neurons or the regeneration of axons from cell bodies and their reconnection with other neurons remains a major hurdle. Injuries relating to war and accidents attracted medical professionals throughout early history to regenerate and reconnect nerves. Early literature till 1990 lacked specific molecular details and is likely provide some clues to conditions that promoted neuron and/or axon regeneration. This is an avenue for the application of natural language processing (NLP) to gain actionable intelligence. Post 1990 period saw an explosion of all molecular details. With the advent of genomic, transcriptomics, proteomics, and other omics—there is an emergence of big data sets

This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the noncommercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

Correspondence to: Sanjoy K. Bhattacharya. University of Miami, 1638 NW 19th Avenue, Room 707A, Miami, Florida 33136, USA. sbhattacharya@med.miami.edu.

Contributions: (I) Conception and design: All authors; (II) Administrative support: SK Bhattacharya; (III) Provision of study materials or patients: SK Bhattacharya; (IV) Collection and assembly of data: S Khosla, L Abdelrahman; (V) Data analysis and interpretation: S Khosla, L Abdelrahman; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://aes.amegroups.com/article/view/10.21037/aes-21-29/coif). SKB serves as an unpaid editorial board member of *Annals of Eye Science* from August 2020 to July 2022. SK reports that their research is mostly online using public databases. However, infrastructure such as basic software including zoom meetings were supported by the University of Miami. Personal and software has support from NIH grant EY14801 and RPB unrestricted grants to University of Miami for research. LA reports that the research is mostly online using public databases, while meeting software (e.g., Zoom) was sponsored by the University of Miami. JJ reports that there are no conflicts of interest with regard to funding sources, meetings and travels reported in his COI form. MS reports that their research is mostly online using public databases. However, infrastructure such as basic software including zoom meetings were supported by the University of Miami. Personal and software has support from NIH grant EY14801 and RPB unrestricted grants to University of Miami for research. SKB reports that their research is mostly online using public databases. However, infrastructure such as basic software including zoom meetings were supported by the University of Miami. Personal and software has support from NIH grant EY14801 and RPB unrestricted grants to University of Miami for research.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and is another rich area for application of NLP. How the neuron and/or axon regeneration related keywords have changed over the years is a first step towards this endeavor.

Methods: Specifically, this article curates over 600 published works in the field of neuroregeneration. We then apply a dynamic topic modeling algorithm based on the Latent Dirichlet allocation (LDA) algorithm to assess how topics cluster based on topics.

Results: Based on how documents are assigned to topics, we then build a recommendation engine to assist researchers to access domain-specific literature based on how their search text matches to recommended document topics. The interface further includes interactive topic visualizations for researchers to understand how topics grow closer and further apart, and how intra-topic composition changes over time.

Conclusions: We present a recommendation engine and interactive interface that enables dynamic topic modeling for neuronal regeneration.

Keywords

Regeneration; natural language processing (NLP); machine learning; deep learning

Introduction

The scientific literature on neuro-regeneration and reinnervation began in the 18th century (1). Since then, hundreds of articles have been published on topics ranging from key regeneration genes (2–4) to the dynamics of remyelination (5). This knowledge corpus is vast and dynamic as new insights are continually emerging from research across multiple continents and topics. Given this vast accumulating body of research, it is hard to know how this literature is evolving. For example, which are the emerging research topics? Which ones have died out? Answers to such questions are important to scholars who want to pursue further research in this area or physicians who want to update their knowledge.

The purpose of this paper is to describe the design and implementation of a computational system that analyses digitally curated past research and provides researchers with an interface to query the corpus and get recommended papers. Furthermore, the web-based interactive interface provides a visualization tool to examine the evolution of topics over time. Using our proposed system, researchers can find answers to narrow queries such as the molecules that play a role in nerve regeneration or broad queries such as the gaps in the literature.

Our computational system relies on the advances made in natural language processing (NLP). Specifically, we use dynamic topic modelling (DTM) (6) to extracts topics and track changes of topics over time. It also shows users how keywords and topics evolve. Beyond applying DTM, we created a novel user interface for researchers to apply the trained model to their data. Overall, we make the following contributions to the interface of neural regeneration and computation:

I. We compiled and curated a corpus of over 600 published works in the neuroregeneration field spanning 1776 to 2020;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

II. We present an interactive time-dependent dynamic topic model specific for neuroregeneration neuroscience literature for researchers to grasp how topics change, how authors cluster based on topics, and how individual keywords rise and fall in popularity over time;

III. We created a dynamic user interface for physicians and scientists to interact with the data. The interface displays visual results and ways to see how topic topology evolves. Moreover, the interface recommends literature for users to investigate.

The remainder of this paper is organized as follows: we review the literature and discuss related contributions from prior authors. Then we specify our algorithm design and computational tools. Next, we detail our experiments and report our results. We discuss the impact of our methods and opportunities for future investigation and present conclusions.

Researchers have developed computational tools for other domains, such as lipidomics and metabolomics. These tools help them garner insights from large scale non-structured text data. For example, Schomburg first composed the BRENDA (7) database in 2002, aiming to create a relational database between enzymes, proteins, and their respective biochemical pathways. By integrating information from KEGG metabolic pathways to diseases and other biomedical concepts, the researchers compose a translational database for scientists to examine changing terms over time. In 2017, Schomburg extended BRENDA by parsing 2.6 million papers in the PubMed corpus, extracting unstructured word tokens, and adding relevant labels (e.g., gene and enzyme names). This work represents a growing trend in biomedical sciences: applying NLPnatural language processing and Big Data techniques to derive structured insights. While BRENDA is powerful, it does not focus on the Neuroregeneration domain and does not focus on how terms and topics in the literature change over time. More importantly, BRENDA is applicable to detailed structural analysis rather than the analysis of broad topics and themes.

Besides relational databases and extraction of words, computational approaches like NLP have found use in biosciences. Chen (8) developed a Bio-DTM to explore how the traditional Chinese medical herb, ginseng, is discussed in the literature over time. While the corpus covers a wide array of topics, the literature lacks domain-focused DTM algorithms related to neuroscience, specifically Neuroregeneration.

Relatedly, van Altena (9) applied Latent Dirichlet Allocation (LDA) analysis methods to the PubMed and PMC corpora to model specific topics in those fields. After filtering for stop words and ranking their terms, the authors generated word clouds and other visuals to extract specific literature themes ranging from systems and security to disease prevention. The authors optimized their methods by tuning their number of topic hyperparameters to minimize the Akaike Information Criterion (AIC) (10) to choose the best model for their task. Other authors (11–13) use topic coherence (14,15) or other performance metrics (16) for optimizing the model architecture. After optimizing their performance metric, van Altena reported results on the broad biomedical corpus, yet the results lack domain-specificity.

LDA has applications beyond summarizing topics in a corpus: Wang (17) used Bio-LDA to discover key topics in PubMed, and then utilized entropy between keywords to generate a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

semantic ontology linking gene, disease, and other biological terms in an undirected graph. Beyond ontologies, Hu (18) derived LDA embeddings for unsupervised style suggestions for Etsy users. Although the domain is not related to medicine, their work shows how LDA can be used for inference and suggesting new content for users.

As we show, most of the current literature for Big Data NLP in Biomedicine handles broad corpora like the PubMed databases and lacks domain-specificity. Our focus on curating a corpus specific to Neuroregeneration and applying LDA-topic extraction methods provides nuanced insights in an important biomedical domain. Moreover, by applying DTM methods, we show how the literature evolves, allowing researchers to identify promising avenues for future research.

Methods

Curating the literature on regeneration

We began by curating the literature. To assemble this corpus, we searched in different databases, including Google Scholar and PubMed. We gathered 700 published articles and books within the Neuroregeneration domain and sorted them into bins based on their publication year. During our search, we also included papers in other languages, including French and German. In the 18th century, the scientific community was heterogeneous, and scientists often communicated their findings in their native languages. Despite this variation in language, the majority of the corpus is in English. Figure 1 illustrates the document distribution binned by time period. We chose to bin the entire corpus into 16 sequential time periods for training the DTM model.

Document optical character recognition (OCR)

We used optical character recognition (OCR) technology to convert PDF format research papers into a machine-readable format that we can parse through. To achieve this end, we used the Adobe Document Cloud export PDF functionality (19), converting the .pdf files to .docx, which are parsed by specific Python libraries, as detailed below. When performing OCR, we discarded documents that the Document Cloud failed to process. These documents were primarily scanned images with complex figures or extremely large in disk space. Figure 2 details the overall process for handling the dataset.

Preprocessing words

After using OCR to convert the PDF documents into a usable string format, we preprocessed our text. We utilized simple regex to parse out words from non-letter characters. Additionally, we removed words lacking semantic meaning as stop words using the Python Natural Language Toolkit (NLTK)’s stop word list (20). Initial implementations yielded skewed results because the earlier periods contained some German, French, and other non-English articles. Thus, we also incorporated German, French, Dutch, and Spanish stop words to filter out words in languages beyond English.

To detect essential words in the literature, we assembled a collection of start words. We employed BioBERT (21) training tokens for this named entity recognition task. These

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

tokens were derived from the BioCreative II Gene Mention corpus (22), the NCBI disease corpus (23), the CHEMDNER corpus of chemicals and drugs (24), and the Species-800 corpus (25), among others. We further included the indices of modern textbooks, including Netter's *Concise Neuroanatomy* (26) and Bear's *Exploring the Brain* (27), which contain significant named entities found throughout the corpus. Using these reference sources for start words created a clean filter when working with the data. Using our references for start words we removed these generic terms from the corpus, which allowed us to focus on domain-specific entities. Finally, we stemmed the words to remove redundancy, often found with word plurals and possessives. We chose to use Porter stemming (28) for this final preprocessing step.

Dynamic topic modeling

We used the DTM algorithm, a time-series generative model for data collections over time. The algorithm builds off of Latent Dirichlet Allocation (LDA) which takes advantage of Bayesian probabilistic modelling of words in documents to define an underlying representation of topics in a document (29). The algorithm models each topic by an infinite mixture over an underlying set of topic probabilities to create a topic distribution for each research paper in our corpus. DTM extends this already powerful algorithm by separating the documents in our corpus by time slices, allowing us to model the evolution of topics over time. We analyzed the change in embeddings of each topic's underlying probabilities over time by sequentially ordering our corpus. To extract the DTM topics, we relied on the Python Gensim package (30). We then used these topics to analyze our corpus in a variety of ways. DTM's advantage lies in observing evolving entities and comparing documents from different periods with different word usage. In our experiments, we define the number of topics, T, and the number of time bins, B, as tunable parameters. We define these hyperparameters for selecting the optimal model. Further, we used CV Topic Coherence (31) and Domain expert validation as measures to compare our hyperparameters and evaluate the coherence of the topics produced. The systematic study of the configuration space of coherence measures is abbreviated as CV.

DTM applications for new regeneration text

As noted above, there are many applications of LDA and DTM. We present a novel application of using our trained DTM to provide recommendations related to neural regeneration. Users can input the contents of their current research, abstracts of papers they are investigating, or anything that contains the topics for a researcher's search query. We then preprocess the text following the same pipeline as the documents in our corpus. This is extremely important, as any differences in preprocessing will result in differences in interpretations from the model, yielding poor paper recommendations. The algorithm filters out words that are not start words, stems, tokenizes, and uses our dictionary to convert the tokens into a Bag of Words matrix. The DTM then calculates the log probability distribution of topics and compares this value to every research paper's topic distribution in our corpus. We do this by calculating the Hellinger Distance between each of these topic distributions. In other words, we calculate how similar the topics in the text entered are to the topics in our set of research papers. We then return the research papers with the most similar

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

topics. Below we discuss further the results of this implementation of DTM into a neural regeneration recommender system.

Results

Evaluating topic semantics

Topic coherence grid search—We ran a grid search to obtain the optimal hyperparameters for our DTM. Most notably, the tunable parameter with the most significant impact on this metric is the number of topics, *T*. Having too few topics will not give the recommendation algorithm sufficient flexibility to compare nuanced latent topics, both in the query entered by the user and the research papers in our corpus. For example, when the number of topics increases past (T>5) (T>5), we see a new topic emerge from our corpus that solely relates to the eye. Words such as ocular, retina, and vision comprise of this new topic. On the contrary, a *T* that is too high will result in incoherent topics, comprised of words with little semantic significance to Neuroregeneration researchers. Since the model's primary purpose is to serve as part of a recommendation algorithm, we erred on the side of more topics, since the benefit of outputting more specialized topics outweighs the downside of producing incoherent topics. These incoherent topics would likely return low Hellinger Distances unless the user also entered incoherent text.

Most significantly, we used the CV Topic Coherence, which calculates the similarity of top words in a topic, to examine inter-topic similarity. Topic coherence measures the robustness of the topic distributions which allowed us to test different hyperparameters and measure their impact on our model. Figure 3 displays the grid search results below.

Ultimately, using grid search and topic coherence metrics alone were insufficient to decide the most optimal combination of hyperparameters. When deciding on overall topic quality, we also shared our discovered words and topics with Neuroscience undergraduate and graduate students and faculty at a large Southeastern University. We factored in their input to determine the optimal number of topics. To complete this qualitative assessment, we retrieved the top 20 terms for each topic over each binned time period. Through discussions with these domain experts, we chose to match the ten topics to the overall concepts found in Table 1. To assign topics, we surveyed four neurologists and neurology researchers independently for their summaries on the top 20 topic words. We present here the most salient responses.

Validation from domain experts/emergent opinions of neuroscientists

The DTM model also shows how authors cluster based on topic and shows salient patterns over the years. Figure 4 shows the results of authors found in the corpus and how they cluster together based on similar topics. We provide a key for the topics shown in the supplemental file (available online: https://cdn.amegroups.cn/static/public/10.21037aes-21-29-1.xls).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keyword popularity

Another exciting application that arises from utilizing Bayesian probabilities to model Neuroregeneration research topics is the resulting log probabilities of keywords. In LDA and DTM, each word belongs to a topic, or even multiple topics. Topic assignment depends on probability of that word appearing in that topic. As a result of this, we can study how certain words rise and fall out of particular literature topics over time. In Figure 5, we see several keywords relating to neural regeneration research.

A practical user interface

As mentioned above, our algorithm's advantage lies in its robustness in recommending research papers from different periods that may have different words, syntax, and sentence structure. The model considers papers from different periods differing in everything but the latent topics embedded in that text for recommendation. Since neural regeneration is such a diverse subject, with a lengthy past, usage of this algorithm for recommendations results in finding research papers that would not ordinarily come up with similar searches on Google or PubMed. Additionally, by using Biobert entity and keyword filtering, we can fine-tune our DTM further to study the topics relating to our specific domain resulting in even stronger recommendations. Since we filtered the text in this way, we can create a flexible yet specific, set of topics closely aligned with the research domain.

The website also provides the user with the most important topic from our corpus related to their query, which presents another application. Perhaps the researcher is interested in gaining insights into what topics of regeneration their inputted text contains. In this manner, we include a powerful summarization feature on the website.

Discussion

This article is the first to propose a computational approach to dynamic topic extraction from the corpus on Neuroregeneration literature. It serves as an entry point for future experiments using NLP to facilitate the broader key topic extraction from this research domain. We describe a text-mining pipeline for processing documents involving filtering techniques using BioBert. We also introduce a new approach for understanding topics, keywords, and subjects authors tend to write about over time. Finally, we implement all of the techniques above into a practical tool for the Neuroregeneration community to use involving our DTM that matches latent topics in the domain.

This paper's limitations point to opportunities for further investigation. One such limitation is the link between the topics we derive from DTM. Specifically, to build an accurate knowledge system, we need to go beyond the bag-of-words approach and extract the text's sequential information.

Finally, the DTM results herald further analysis and investigation. Our results are limited to analyzing temporal dynamics, yet geography is another crucial factor to ascertain when and where essential discoveries in the neuroscience field occurred. How do scientific discoveries' time and location reflect geopolitical trends in history? These questions are open for future investigation, and forthcoming publications could focus on addressing these

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

questions. We also hope to implement these future threads of research into additional practical functionality for researchers in the field to leverage.

Conclusions

This paper describes a computational system to investigate the dynamics of the Neuroregeneration literature. In particular, we curated a corpus of over 600 research papers and created a time-dependent dynamic topic model based on this corpus. This model is the backbone of a web application for researchers to visualize topics changing over time, to view how authors' works cluster together, and to gain insights through an article recommendation system. By interpreting historical works, our applied tools advance neuroregeneration's future development.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are extremely thankful to the many volunteers from the University of Miami RRLR2K project, who contributed towards collection of original papers. More information with links to resources, visualizations, and the Neuroregeneration research recommendation system can be found at: http://regenx.herokuapp.com/.

Funding:

This project was partly supported by NIH grant EY14801 and an unrestricted funds from Research to Prevent Blindness to University of Miami.

References

1. Ochs S The early history of nerve regeneration beginning with Cruikshank's observations in 1776. Med Hist 1977;21:261–74. [PubMed: 333203]
2. Goldberg JL, Barres BA. Nogo in nerve regeneration. Nature 2000;403:369–70. [PubMed: 10667770]
3. Zhao LX, Zhang J, Cao F, et al. Modification of the brain-derived neurotrophic factor gene: a portal to transform mesenchymal stem cells into advantageous engineering cells for neuroregeneration and neuroprotection. Exp Neurol 2004;190:396–406. [PubMed: 15530878]
4. Thompson JA, Ziman M. Pax genes during neural development and their potential role in neuroregeneration. Prog Neurobiol 2011;95:334–51. [PubMed: 21930183]
5. Arnett HA, Wang Y, Matsushima GK, et al. Functional genomic analysis of remyelination reveals importance of inflammation in oligodendrocyte regeneration. J Neurosci 2003;23:9824–32. [PubMed: 14586011]
6. Blei DM, Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning: 2006:113–20.
7. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res 2002;30:47–9. [PubMed: 11752250]
8. Chen Q, Ai N, Liao J, et al. Revealing topics and their evolution in biomedical literature using Bio-DTM: a case study of ginseng. Chin Med 2017;12:27. [PubMed: 28919923]
9. van Altena AJ, Moerland PD, Zwinderman AH, et al. Understanding big data themes from scientific biomedical literature through topic modeling. Journal of Big Data 2016;1:1–21.
10. Sakamoto Y, Ishiguro M, Kitagawa G. Akaike information criterion statistics. Dordrecht, Netherlands: Springer Netherlands, 1986.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

11. Stevens K, Kegelmeyer P, Andrzejewski D, et al. Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning: 2012:952–61.
12. O'Callaghan D, Greene D, Carthy J, et al. An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications 2015;42:5645–57.
13. Jelodar H, Wang Y, Yuan C, et al. Applications: Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools 2019;78:15169–211.
14. David N, Grieser JHLK, Timothy B. Automatic evaluation of topic coherence In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics 2010:100–8.
15. Röder M, Both A, Hinneburg A: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining: 2015:399–408.
16. Omar M, On B-W, Lee I, et al. LDA topics: Representation and evaluation. Journal of Information Science 2015;41:662–75.
17. Wang H, Ding Y, Tang J, et al. Finding complex biological relationships in recent PubMed articles using Bio-LDA. PLoS One 2011;6:e17243. [PubMed: 21448266]
18. Hu DJ, Hall R, Attenberg J: Style in the long tail: Discovering unique interests with latent variable models in large scale social e-commerce. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining: 2014:1640–9.
19. Acrobat A Adobe Acrobat Export PDF In: Document Cloud https://documentcloudadobecom/
20. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. Sebastopol, California: O'Reilly Media, Inc., 2009.
21. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–40. [PubMed: 31501885]
22. Smith L, Tanabe LK, Ando RJ, et al. Overview of BioCreative II gene mention recognition. Genome Biol 2008;9 Suppl 2:S2.
23. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 2014;47:1–10. [PubMed: 24393765]
24. Krallinger M, Rabal O, Leitner F, et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 2015;7:S2. [PubMed: 25810773]
25. Pafilis E, Frankild SP, Fanini L, et al. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. PLoS One 2013;8:e65390. [PubMed: 23823062]
26. Rubin M, Safdieh JE. Netter's Concise Neuroanatomy Updated Edition E-Book. Netherlands: Elsevier Health Sciences, 2016.
27. Bear M, Connors B, Paradiso MA. Neuroscience: Exploring the Brain, Enhanced Edition: Exploring the Brain. Bolingbrook, Illinois: Jones & Bartlett Learning, 2020.
28. Willett P The Porter stemming algorithm: then and now. 2006 Program.
29. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. 2003;3:993–1022.
30. Khosrovian K, Pfahl D, Garousi V. Gensim 2.0: a customizable process simulation model for software process evaluation. In: International conference on software process. Springer, 2008:294–306.
31. Syed S, Spruit M. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International conference on data science and advanced analytics (DSAA). IEEE: 2017:165–74.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

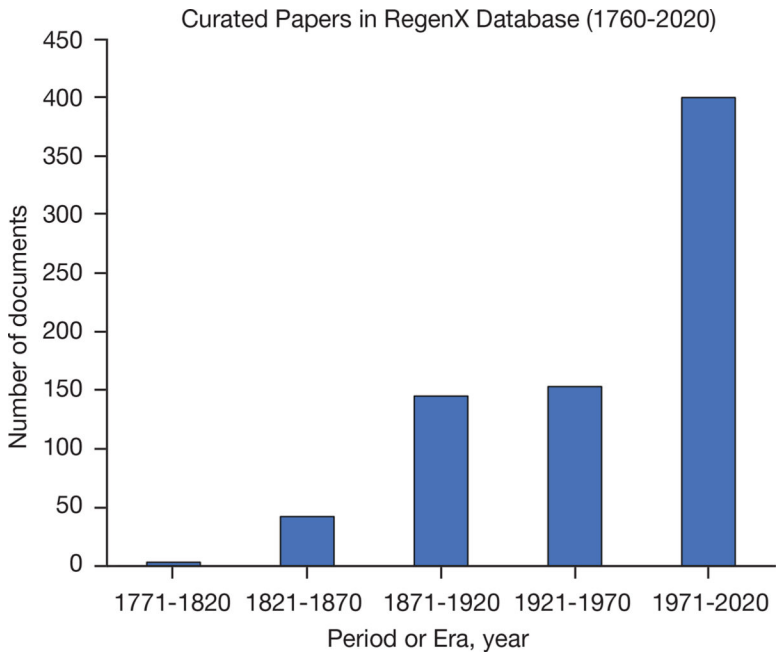

Figure 1.

Literature distribution over time: 1771 to 2020. The curated papers from 1700–2020, x- and y-axis are years of citation and number of documents obtained respectively. These documents were used in preparation of RegenX database (http://regenx.herokuapp.com).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

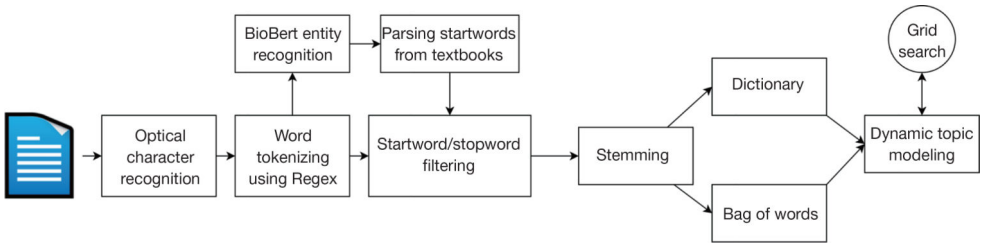

Figure 2.

Overall process pipeline. The different steps of dynamic topic modeling for RegenX database is depicted in this flow chart.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

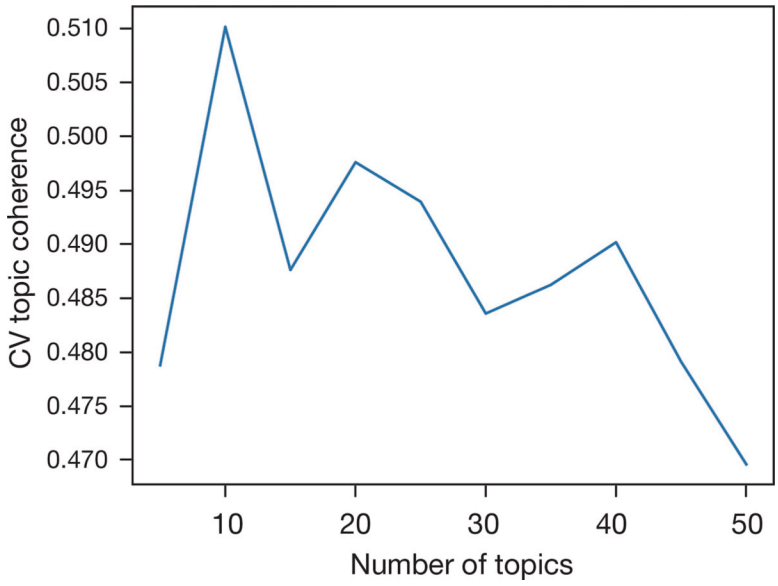

Figure 3.

The plot of coherence measures (CV) topic coherence as a function of the number of topics, using 16 time periods. We found that 10 topics have the highest topic coherence. The x- and y-axis denotes number of topics and systematic study of the configuration space of CV respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

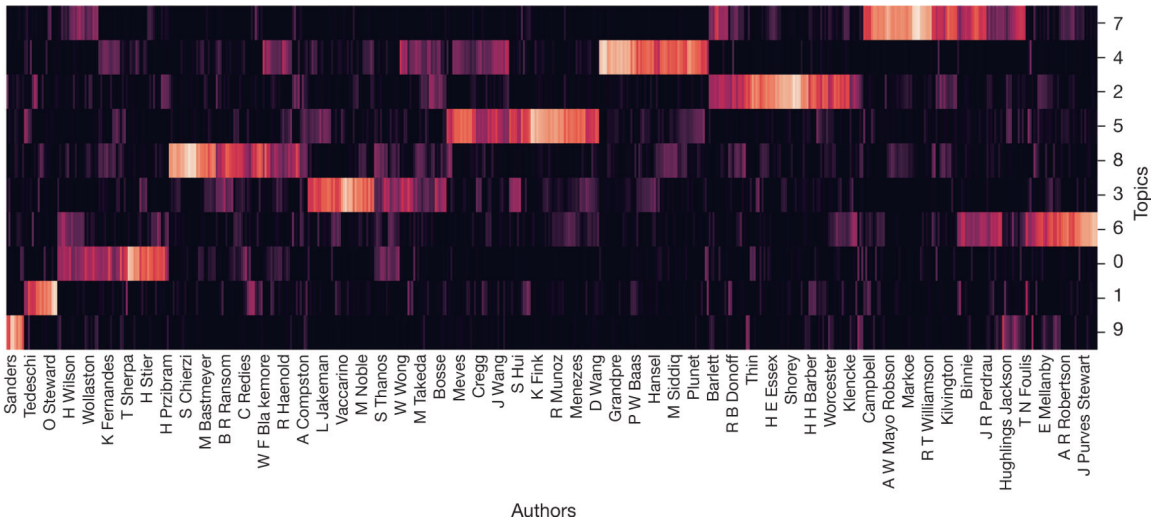

Figure 4.

A cluster map illustrating how certain authors cluster around the same topic. Brighter intensity indicates higher topic log probability, which means the respective author's work is more likely to align with a topic. The x- and y-axis denotes author name and topics respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

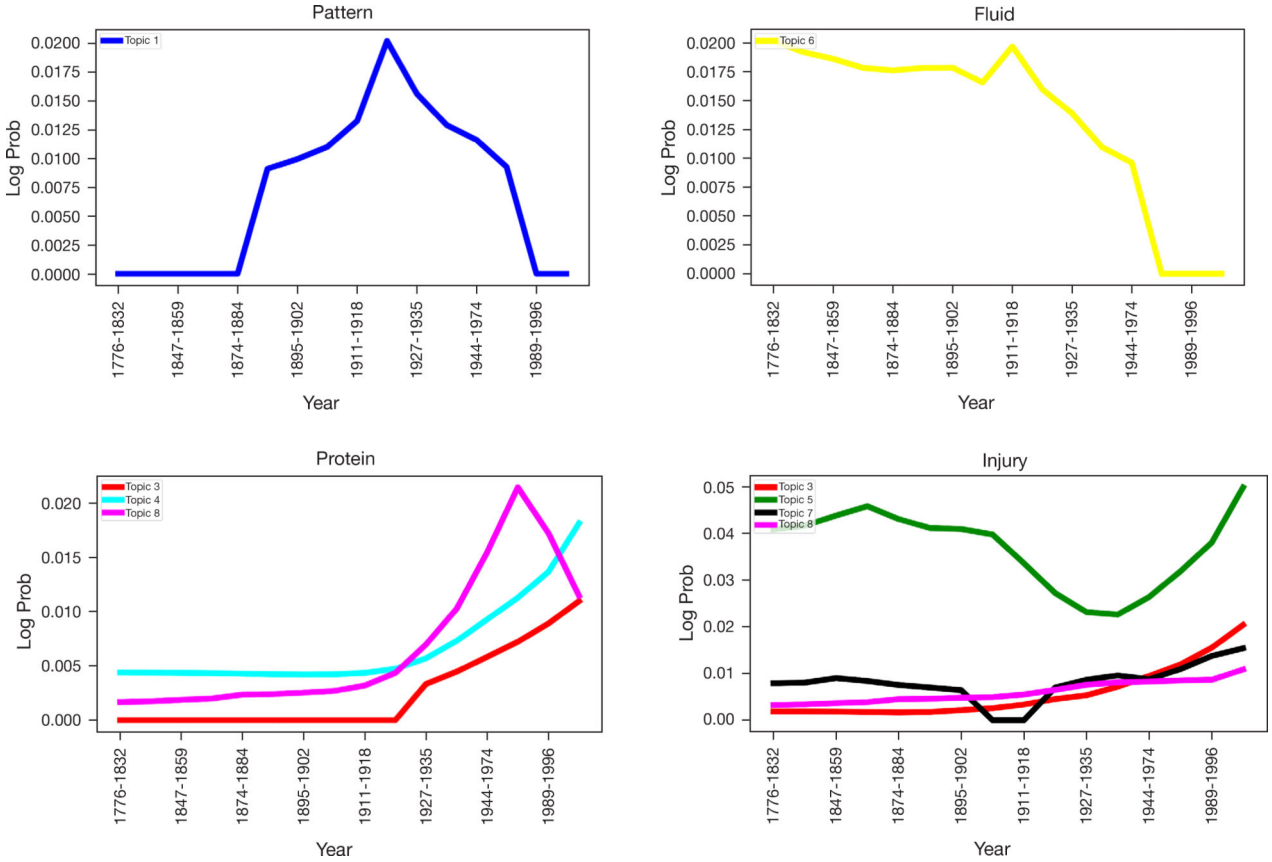

Figure 5.

Top: selected words from topics that belong to the same topic over time. Bottom: words belonging to multiple topics with their log-probability plotted as a function of time. Log Prob is logarithmic probability of the word logging to a topic over the years. Pattern, Fluid, Protein and Injury were four selected words for this analysis.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Matching the topic numbers to their concepts

Topic number	Concept
1	Vision
2	Processes and activity
3	Anatomy
4	Cells
5	Growth and regeneration
6	Spinal cord
7	Disease
8	Movement
9	Optic nerve regeneration
10	CNS neuroanatomy

CNS, central nervous system.