An Online Service for Topics and Trends Analysis in Medical Literature

Spyridon Kavvadias, George Drosatos, and Eleni Kaldoudi

Abstract

Topic modeling refers to a suite of probabilistic algorithms for extracting word patterns from a collection of documents aiming for data clustering and detection of research trends. We developed an online service that implements different variations of Latent Dirichlet Allocation (LDA) algorithm. Scientific literature origin from targeted search queries in PubMed, works as input while output files are available for every step of the process. Researchers can compare the results of different corpora, preprocessing texts and topic modeling parameters in a quick and organized way. Information regarding topics help users assign labels and group them to categories. Visualization of data is a contribution of our service with graphs generated on the fly providing information about the corpora, the topics, groups of topics and categories as well. We rely in modern technologies and follow the principles of agile software development to achieve scalability and discreet design.

Keywords

Topic modeling • Content analysis • Trend analysis
Visualization

1 Introduction

Our era is characterized by continuous advances in biomedical sciences and a corresponding large amount of scientific publications each year. Literature topics and trends analysis is increasingly employed to give insights on past and future research directions. Several statistical algorithms have been applied to model topics in scientific literature [1–5]. As such methods require considerable mathematical and programming background, recent research proposes user friendly integrated tools to enable researchers of various backgrounds to explore topics analysis [6–8]. However, currently available tools do not cover the entire topics and trends analysis workflow and require custom set up. In this paper, we propose an open source and platform independent service to support topic modeling and trends analysis for the biomedical expert. The service allows creation and description of biomedical literature corpora, supports the entire workflow of topic modeling and trends analysis and provides visual navigation of the results.

2 Topic Modeling

Topic modeling [9] refers to a suite of algorithms that aim to analyze the hidden structure of a collection of documents. Each document is characterized by a variation of topics, each topic consists of a collection of words and each word has its own statistical weight. Several topic modeling approaches are available [3–5, 10], Latent Dirichlet Allocation (LDA) being one of the most popular. The algorithm starts by randomly assigning each word of a document in one of K topics. Then, it calculates conditional probabilities for each topic in each document ($(t|d)$ where t denotes the topic and d denotes the document) and for each word in every topic ($(w|t)$ where w denotes word). Through an iterative process, it reassigns words and topics until they reach a steady state. The algorithm requires setting the initial number K of assumed topics and the parameters that define the Dirichlet prior for the per document topic distribution (parameter α) and for the per topic word distribution (parameter β).

Topic modeling has been successfully applied in many other research areas, for example to analyze and classify genomic sequences [11], classify images based on visual

S. Kavvadias (✉) · G. Drosatos · E. Kaldoudi
School of Medicine, Democritus University of Thrace,
Alexandroupoli, Greece
e-mail: skavvadi@med.duth.gr

G. Drosatos
e-mail: gdrosato@ee.duth.gr

E. Kaldoudi
e-mail: kaldoudi@med.duth.gr

L. Lhotska et al. (eds.), *World Congress on Medical Physics and Biomedical Engineering 2018*,
IFMBE Proceedings 68/1, https://doi.org/10.1007/978-981-10-9035-6_89

words topic modeling [12], detect discussion themes in social networks [13] and analyze source code [14]. Additionally, there are several implementations of topic modeling (and especially of LDA) in different programming languages [15–18]. In this paper, we integrate some of the existing implementations in an online service which provides added value functionalities, including user-friendly interface to visualize and label topics and tools to support trend analysis. The service also allows for generation of rich metadata for each step of the workflow, to fully document the topic modeling experiments.

3 Topic Modeling and Trends Analysis Service

An overview of literature topic modeling and trends analysis workflow is presented in Fig. 1. The process starts with the generation of the initial literature corpus, as a collection of relevant published papers; most often the collection is limited to paper titles and abstracts due to access restrictions. Following a rudimentary text preprocessing, the topic modeling algorithm is parametrized and applied to identify topics that are essentially word collections. Human intervention is required to label topics so that they are meaningful for human interpretation. Finally, popularity of topics over time is assessed for trends analysis.

We have developed a user-friendly web-based environment to encapsulate this entire workflow of topic modeling and trends analysis and provide this as a service for the non-expert biomedical scientist. The input of the service is a corpus of research abstracts retrieved from PubMed as a result from a specific query. The system allows the user to describe each corpus via relevant metadata, including corpus generation date, initial database query, study aim and user details. Basic corpus statistics are also calculated, e.g. number of publications per year (given as a graph), total number of articles, number of articles with an abstract and minimum–maximum year of articles.

Text preprocessing is routinely used to clean the corpus via: (1) removal of all the punctuation and escape codes; (2) exclusion of stop-words; (3) conversion of all words to their lemmas by applying the stemming procedure; and (4) exclusion of articles with no words in their abstracts or less than 3 letters in their titles. Current service implementation uses the most commonly used Krovetz stemmer [19] as a default option for the stemming process. However, the service allows importing of additional stemming algorithms [10]. The service provides basic preprocessing statistics and allows the user to generate metadata to richly describe preprocessed corpora for future reference.

The processed corpus can be archived and used as input to the topic modeling procedure along with the necessary execution parameters. Currently, we support two different

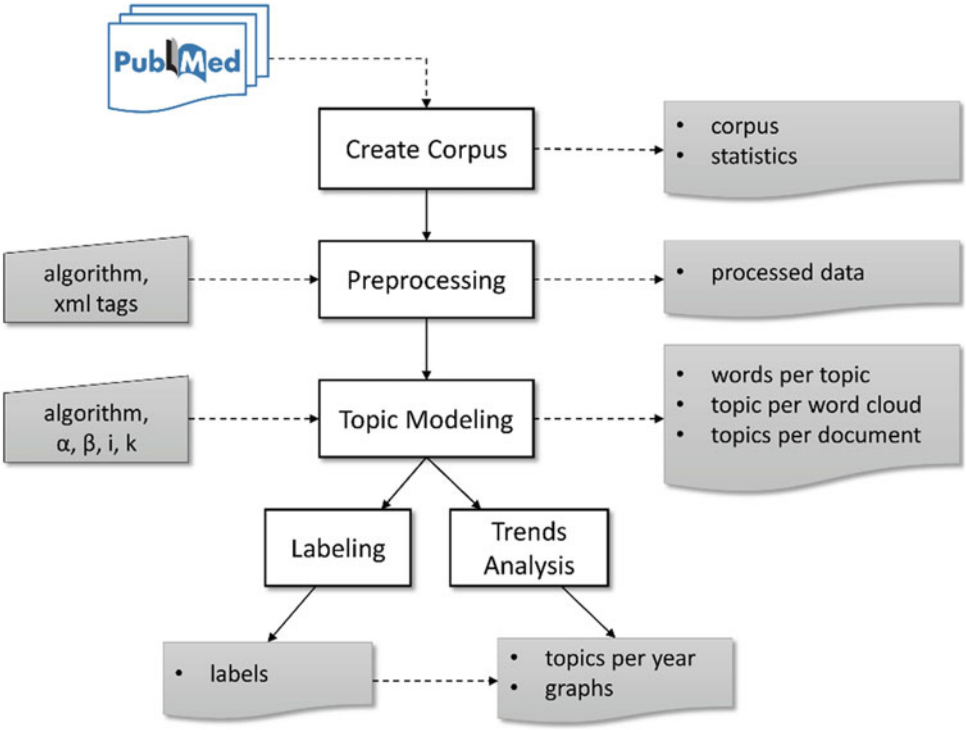
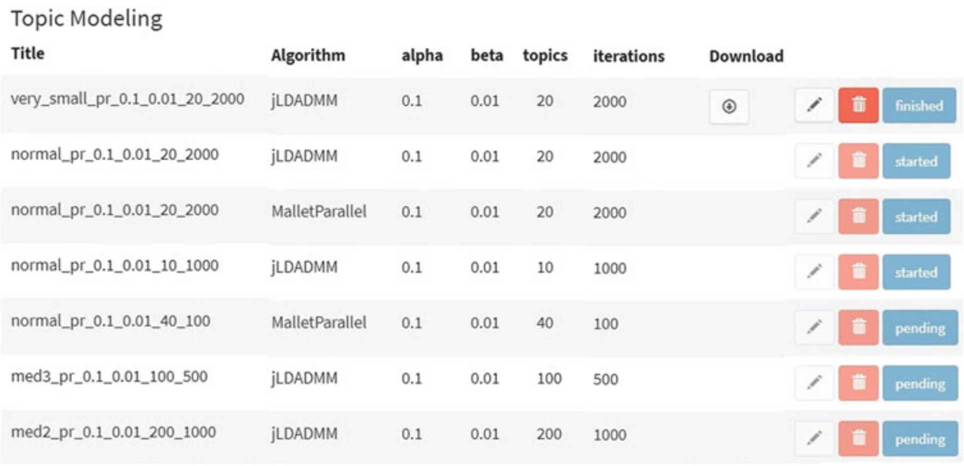


Fig. 1 The basic workflow, inputs and outputs of the platform

Fig. 2 Metadata table for topic modeling



Topic Modeling

Title	Algorithm	alpha	beta	topics	iterations	Download			
very_small_pr_0.1_0.01_20_2000	jLDADMM	0.1	0.01	20	2000	⊙	✎	🗑	finished
normal_pr_0.1_0.01_20_2000	jLDADMM	0.1	0.01	20	2000		✎	🗑	started
normal_pr_0.1_0.01_20_2000	MalletParallel	0.1	0.01	20	2000		✎	🗑	started
normal_pr_0.1_0.01_10_1000	jLDADMM	0.1	0.01	10	1000		✎	🗑	started
normal_pr_0.1_0.01_40_100	MalletParallel	0.1	0.01	40	100		✎	🗑	pending
med3_pr_0.1_0.01_100_500	jLDADMM	0.1	0.01	100	500		✎	🗑	pending
med2_pr_0.1_0.01_200_1000	jLDADMM	0.1	0.01	200	1000		✎	🗑	pending

implementations of LDA based on the Java libraries Mallet ParallelTopicModel [17] and jLDADMM [18] with input/output performance enhancements. Service architecture supports easy integration of other LDA implementations based on predefined public interface descriptions.

Topic modeling experiments are resource and time consuming while they often be repeated with different initialization parameters. The proposed service displays a current status of scheduled topic modeling experiments and supports a powerful experimental lab-bookkeeping. As shown in Fig. 2, the user is guided to insert relevant metadata that describe in detail each topic modeling experiment. Metadata can be edited and updated, while they automatically inform saved experimental results and trends analysis and visualizations produced in the following steps of the workflow.

Another important service feature of added value is the ability for the user to label each topic. The procedure of assigning labels to topics is shown in Fig. 3. For every topic that has been generated by the execution of the algorithm, the top most words (number indicated by the user) that describe the topic are ranked by statistical weight and displayed for the user in tabular form or as word clouds. The user can then assign a title to each topic and create nested categories to organize various topics.

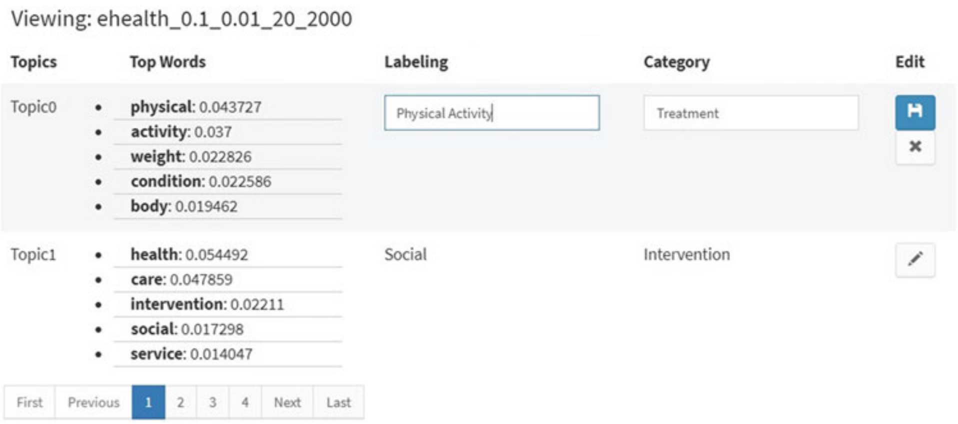
The final step in the workflow involves trends analysis on the identified topics. The popularity (t, y) of the topic (t) for each year (y) is calculated as the mean of the weight of this topic for all documents published this year (D_y):

$$P(t,y) = \frac{1}{|D_y|}\sum\nolimits_{d\in D_y}\frac{|\{\mathrm{w} \in \mathrm{d} : \mathrm{topic}(\mathrm{w}) = \mathrm{t}\}|}{|d|} \qquad (1)$$

where t represents a topic and w is a word in document d of the documents' collection D_y for year y [20]. Calculated trends are then displayed as graphs. The service supports for rich visualizations which allows comparative displays of different topics and categories and corpora, while preserving metadata information describing the different experiments whose results are compared. An example is shown in Fig. 4. The user can generate graphs on the fly for any group of selected topics or categories and compare trends for a chosen time range.

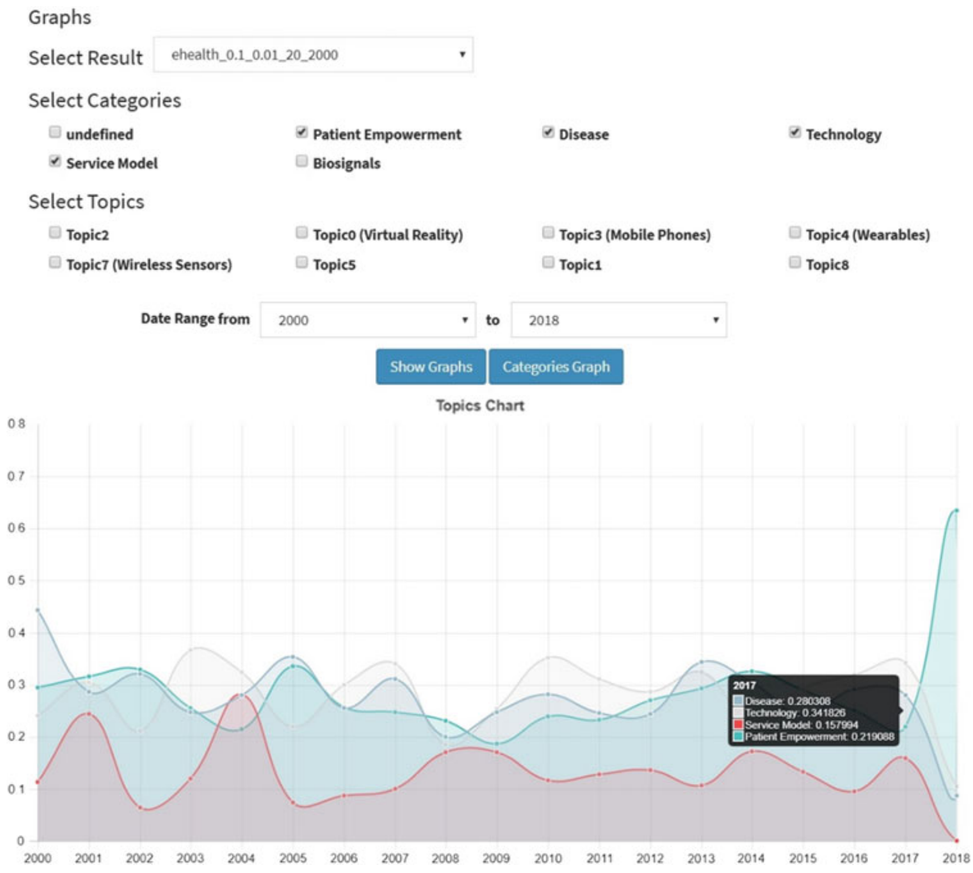
The system is implemented in NodeJS with LoopBack framework (http://loopback.io) and is accessible at https://

Fig. 3 Assigning labels and categories to topics



Viewing: ehealth_0.1_0.01_20_2000

Topics	Top Words	Labeling	Category	Edit
Topic0	**physical**: 0.043727 **activity**: 0.037 **weight**: 0.022826 **condition**: 0.022586 **body**: 0.019462	Physical Activity	Treatment	💾 ✖
Topic1	**health**: 0.054492 **care**: 0.047859 **intervention**: 0.02211 **social**: 0.017298 **service**: 0.014047	Social	Intervention	✎



Graphs

Select Result: ehealth_0.1_0.01_20_2000

Select Categories

☐ undefined ☑ Patient Empowerment ☑ Disease ☑ Technology
☑ Service Model ☐ Biosignals

Select Topics

☐ Topic2 ☐ Topic0 (Virtual Reality) ☐ Topic3 (Mobile Phones) ☐ Topic4 (Wearables)
☐ Topic7 (Wireless Sensors) ☐ Topic5 ☐ Topic1 ☐ Topic8

Date Range from 2000 to 2018

2017
🟦 Disease: 0.280308
⬜ Technology: 0.341826
🟥 Service Model: 0.157994
🟩 Patient Empowerment: 0.219088

Fig. 4 Categories Graph. The user selects the result and the categories that wants to compare. The topics of each category appear after the user check them

trends.duth.carre-project.eu/. Data storage is based on the MongoDB (https://www.mongodb.org). The frontend is powered by AngularJS framework (https://angularjs.org) and the graph visualizations are implemented using Chart.js (http://www.chartjs.org/) and Vis.js library (http://visjs.org). In the backend, we developed a mechanism for the management of parallel processes that are possible to be requested by the same user or not. For this purpose, we used a FIFO philosophy (first-in, first-out) for the execution of processes and limitations on the number of processes (e.g. three) that are possible to be executed simultaneously. This is required because our system has limited computing resources and the topic modeling algorithms require high computational cost.

4 Discussion

This paper proposes a web-based service that allows biomedical researchers with no experience in data modelling and programming to execute topic modeling and trends analysis experiments of biomedical literature corpora, keep experimental details and visualize the results. Work in progress includes to make our web service free of bugs, support more topic modeling algorithms with an easy mechanism to add new implementations of them, and to develop a mechanism that would add a batch of processes with different parameters with goal to select the appropriate ones (e.g. the number of topics). Additionally, we plan to perform an evaluation of our system regarding the system's performance and the users' satisfaction.

Acknowledgements This work was supported by the FP7-ICT project CARRE (Grant No. 611140), funded in part by the European Commission and Greek National Matching funds (DUTH KE81442).

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Paul, M., Girju, R.: Topic modeling of research fields: An interdisciplinary perspective. In: International Conference Recent Advances in Natural Language Processing (RANLP 2009), pp. 337–342 (2009).
2. Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W: An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1608 (2016).
3. Blei, M., D., Andrew, Y., J., Jordan, I., M.: Latent dirichlet allocation. Journal of Machine Learning Research, Vol. 3, pp. 993–1022 (2003).

4. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391 (1990).
5. Hofmann, T.: Probabilistic latent semantic analysis. In: 15th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc. pp. 289–296 (1999).
6. Scrivner, O., Davis, J.: Topic modeling of scholarly articles: Interactive text mining suite. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" (2016).
7. Kim, D., Swanson, B. F., Hughes, M. C., Sudderth, E. B.: Refinery: An open source topic modeling web platform. Journal of Machine Learning Research, 18(12), 1–5 (2017).
8. Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., Seppi, K.: The topic browser: An interactive tool for browsing topic models. In: NIPS Workshop on Challenges of Data Visualization (Vol. 2) (2010).
9. Blei, M.: Probabilistic topic models. Communications of the ACM, 55(4):77–84, (2012).
10. Jurafsky, D., Martin, J. H: Speech and language processing. 3rd edn. Pearson, London (2017).
11. La Rosa, M., Fiannaca, A., Rizzo, R., Urso, A.: Probabilistic topic modeling for the analysis and classification of genomic sequences. BMC Bioinformatics, 16(6), S2 (2015).
12. Rasiwasia, N., Vasconcelos, N.: Latent dirichlet allocation models for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(11), 2665–2679 (2013).
13. Lau, J. H., Collier, N., Baldwin, T.: On-line trend analysis with topic models: #twitter trends detection topic model online. In: 24th International Conference on Computational Linguistics, pp. 1519–1534 (2012).
14. Binkley, D., Heinz, D., Lawrie, D., Overfelt, J.: Understanding LDA in source code analysis. In: 22nd International Conference on Program Comprehension, pp. 26–36, ACM, New York, NY, USA (2014).
15. Topic Modeling Software, http://www.cs.columbia.edu/~blei/topicmodeling_software.html, last accessed 2018/02/05.
16. Grün, B., Hornik, K.: topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40(13), 1–30 (2011).
17. MALLET: A machine learning for language toolkit, http://mallet.cs.umass.edu, last accessed 2018/02/05.
18. jLDADMM: A Java package for the LDA and DMM topic models, http://jldadmm.sourceforge.net, last accessed 2018/02/05.
19. Krovetz, R.: Viewing morphology as an inference process. In: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 191–202, ACM, New York, NY, USA (1993).
20. Priva, U. C., Austerweil, J. L.: Analyzing the history of Cognition using topic models. Cognition, 135, 4–9 (2015).