A computational literature review of football performance analysis through probabilistic topic modelling

Vitor Ayres Principe[1,2,3], Rodrigo Gomes de Souza Vale[1,2], Juliana Brandão Pinto de Castro[1,2], Luiz Marcelo Carvano[3], Roberto André Pereira Henriques[4], Victor José de Almeida e Sousa Lobo[3,4,5] & Rodolfo de Alkmim Moreira Nunes[1,2]

[1]Postgraduate Program in Exercise and Sport Sciences, Instituto de Educação Física e Desportos (IEFD), Rio de Janeiro State University, Rua São Francisco Xavier, 524, Pavilhão João Lira Filho, Bloco F, 9º andar, Maracanã, Rio de Janeiro 20550-900, Brazil

[2] Laboratory of Exercise and Sports (LABEES), Rio de Janeiro State University, Rio de Janeiro, Brazil

[3] Nova Information Management School (NOVA IMS), Lisbon, Portugal

[4] Portuguese Naval School, Almada, Portugal

[5] CINAV (Naval Research Center), Almada, Portugal

This is the Author Peer Reviewed version of the following article published by Springer:

Principe, V.A., de Souza Vale, R.G., de Castro, J.B.P. *et al.* A computational literature review of football performance analysis through probabilistic topic modeling. *Artif Intell Rev* **55**, 1351–1371 (2022). https://doi.org/10.1007/s10462-021-09998-8

A computational literature review of football performance analysis through probabilistic topic modeling

Abstract

This research aims to illustrate the potential use of concepts, techniques, and mining process tools to improve the systematic review process. Therefore, we performed a review on two online databases (Scopus and ISI Web of Science) from 2012 to 2019. We identified 9,649 studies that were analyzed by probabilistic topic modeling procedures in a machine learning approach. The Latent Dirichlet Allocation (LDA) method, chosen for modeling required the stages: 1) data cleansing, 2) data modeling into topics for coherence and perplexity analysis. All research was conducted according to the standards of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) in a fully computerized way. The computational literature review (CLR) is an integral part of a broader literature review process. The results presented met three criteria: (1) literature review for a research area, (2) analysis and classification of journals, and (3) analysis and classification of academic and individual research teams. A contribution of the article is to demonstrate how the publication's network formed in this particular field of research, and the content of the abstracts can be automatically analyzed to provide a set of research topics for quick understanding and application in future projects.

Keywords Football; Performance Analysis; Literature review; Computational literature review; Topic models; LDA

Introduction

Over time, methods for conducting systematic reviews have become more rigorous, further prolonging the completion of reviews (Pham et al. 2018), due to finite resources concerning time and effort (Jennex 2015). Among this, a researcher, a doctoral student, or both, to better understanding a research area, needs to quickly get an overview of the literature associated with which journals have the most significant impact and what are the most recent and frequent topics (Mortenson and Vidgen 2016). Thus, researchers contribute to knowledge generation based on searches and promote education. For this, the use of text analysis is beneficial, given the significant increase in the number of electronic research materials in this new era (Lee et al. 2014).

Brings scientists new challenges and opportunities due to the characteristics related to the volume, variety, speed of data creation (Chen, Zhong, and Yuan 2016). The systematic literature review (SLR) provides reliable means and established methods for carrying out a comprehensive and robust literature review (Felizardo et al. 2011). However, conducting this researches becomes quite costly due to the studies' growth of 8 to 9% each year, as reported by Bornmann and Mutz (2015). Besides, to being more significant than they used to be, bibliometric datasets are becoming more complex (McLevey and McIlroy-Young 2017).

This abundant data requires computational skills to access these vast bibliometric data. Several programming languages used to make access more accessible to the academic database. The *pybliometrics*, Python package (Rose and Kitchin 2019), *rscopus*, R package (Muschelli 2018)) to access the RESTful APIs that Scopus provides, and other projects can found to access different databases like Web of Science, PubMed, Google Academic and more.

Within this context, after bibliometric data acquisition, Text Mining is a well-established practice. It is commonly used to extract non-trivial patterns and knowledge from unstructured documents or textual documents written in natural language (Felizardo et al. 2011). Among the various methods of text mining and grouping, we highlight probabilistic topic modeling (Blei et al., 2003). This method captures two essential aspects: (1) words can have multiple meanings, and (2) interpretations and documents may contain one or more topics (van Altena et al. 2016).

In this way, natural language processing (NLP) is producing visible practical results due to the advancement of machine learning techniques. One of its main applications is the classification of documents, which received significant attention. In general, document classification problems investigated by (1) coding each word or document for a numerical vector, and (2) classifying documents (Shimada, Kotani, and Iyatomi 2016).

In coding, the Latent Dirichlet Allocation (LDA) method is the most popular topic-modeling algorithm. The LDA assigns a document probability distribution to the word of each topic (Blei et al. 2003). For the document classification, we highlight the logistic regression, the artificial neural networks, the Bayesian structures, and the support vector machine, which are widely used. In recent years, sentence vector representations and recurrent neural networks have shown promising results in several problems of document classification in English (Shimada et al. 2016).

Therefore, this study aims to demonstrate new essential concepts, mainly for the *Stricto Sensu* programs of Physical Education universities, and to illustrate the potential use of concepts, techniques, and tools of process mining to improve the systematic review process known as computational literature review (CLR). The CLR can identify the main terms and interpretations found in the articles on soccer performance analysis conducted during the last seven years of scientific production.

Method

The purpose of a literature review is often to allow the researcher to map and evaluate the existing intellectual territory to specify a research question and develop additional knowledge (Tranfield, Denyer, and Smart, 2003). However, with the increase in the number of journals, the time and effort required to conduct a literature review are increasing, prompting researchers to choose where to allocate the resources to do empirical research instead of extensive literature reviews. Consequently, the quality outcome of literature reviews is declining (Jennex, 2015). One possibility to solve the problems of literature reviews is to conduct an SLR, which follows a set of transparent and reproducible steps (an algorithm). In this way, Jahangirian et al. (2011) propose the use of automation to assist in the stages of search and screening.

Research Framework and Development of the Computational Literature Review

The CLR of the present study was conducted under the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Liberati et al. 2009). It provides an overview of the literature and the most relevant topics that were published in the studies. The CLR Framework process (Figure 1) begins with the identification of the type of case (literature review, periodical analysis, research management) that will be investigated. For the CLR, the search terms are the same as those used in an SLR.

The structure, shown in Table 1, summarizes the processes and steps of how to extract the latent topics from the data of the articles. The data source for the computational review of the articles were the online databases Scopus and ISI Web of Science, searched on December 1, 2019, for relevant articles published between January 1, 2012, and December 1, 2019, using the keywords "Football," "Soccer," each associated with the terms "Performance" and "Analysis."

=== Insert Table 1 here ===

The present review limited the information sources to scientific journals to guarantee the articles' quality. This delimitation is justified, because academics and professionals, to acquire and disseminate knowledge, generally consult scientific journals (Ngai, Xiu, and Chau 2009).

In the selection of the article, an advanced search procedure used where Boolean expressions ("AND" and "OR") allow combinations of keywords (Rowley and Slack 2004). Then, the articles pre-selected in the online journals were exported in two different formats:

- Research Information Systems Incorporated (.ris) is a standardized format used by many digital libraries, such as IEEE Xplore, Scopus, ACM Portal, Scopemed, ScienceDirect, SpringerLink, Rayyan QCRI as well as leading reference/citation management applications, such as Zotero, Citavi, Mendeley, and EndNote, which can export and import citations in this format.
- Web of Science Bibliographic Reference (.ciw), currently managed by Clarivate Analytics, used by digital library Web of Science and readable in the bibliographic reference management software of the Clarivate Analytics, which is called EndNote.

Thus, the organization, identification, and exclusion of duplicated articles were performed in the Mendeley Desktop, a free and accessible software that can be used by researchers. Thus, after deleting duplicated articles, a text file was generated and exported to be used for data analysis.

Impact Analysis

When a research article refers to another article, the original article gets a quote. The number of quotes the article receives can evaluate the impact of an article. Hence, it can be created abstracts by an author (showing how many quotations an author received for the published articles), and by the place of publication (how many quotes a journal received) from counting gross citations to articles included in a CLR. However, citation counts are not problem-free. The h-index is used to evaluate the impact of a researcher and is generally accepted as a useful measure of impact (Hirsch, 2015).

Structure Analysis

Social networks are sets of connected objects represented by a graphic. They have an excellent benefit for the dissemination of information through communication among its members. The network consists of nodes and edges, where each node is a network point, and an edge is a line connecting two nodes (Simsek and Kara, 2018; Wasserman and Faust, 1994).

A social network reflects a social structure that can be represented by individuals or organizations and their relationships. Through this social structure, data, and information exchanged between individuals or organizations can be studied and analyzed at different levels of detail (Horta et al. 2018).

Scientific Social Networks are specific types of social networks that represent the social interactions of researchers that occur in the scientific environment (Horta et al., 2018). They are very popular in the academic community as a way of understanding the structure of the research community and identifying the top researchers in that community. A component of the scientific authorship and co-authorship network is one in which all authors of this component are reachable (Mortenson and Vidgen, 2016).

Content Analysis

In any literature review work, the researcher involved has the concern of identifying the "topics" contained in the documents. In many cases, the evaluation of the work is carried out based only on the review of abstracts. In pragmatic terms, this evaluation becomes reasonable because of the amount of work. The abstract purpose: "facilitate quick and accurate identification of the topic of published papers" (Luhn, 1958). The CLR uses probabilistic topic modeling to automate this analysis.

Probabilistic topic models are a collection of algorithmic approaches to machine learning adopted in the field of text mining. These models seek to find structural patterns within a collection of text documents to extract semantic information from a set of documents, called corpus. The topic templates produce groupings of words that represent the central themes present in a particular corpus. In this way, these techniques provide an automated way of identifying common subjects within the documents presented (Lee et al., 2014; Blei, 2012; Griffiths and Steyvers, 2004).

Given a corpus of documents, probabilistic topic models can find a set of recurring themes called topics. The topics are, in fact, probability distributions on the words of documents. The purpose of topic modeling is to automatically discover topics from a collection of documents (La Rosa et al., 2015). LDA is a probabilistic statistical model used to discover the underlying abstract topics in a series of documents or text data. (Blei et al. 2003). If it assumed that a document is a sequence w of words, where $d = (w_1, w_2, \dots, w_n)$, the generative model for documents can be expressed through the following probability distribution:

$$P(w_i) = \sum_{j=1}^{T} P\left(w_i \middle| z = z_j\right) P(z = z_j) \quad (1)$$

Where $P(w_i)$ is the probability of the word w_i in a given document; $P\left(z = z_j\right)$ is the probability of choosing a topic word z_j for the current document; $P\left(w_i \middle| z = z_j\right)$ is the probability of showing the word w_i on a certain topic z_j and T is the number of topics.

The LDA model is represented as a probabilistic graphical model in Figure 1. This model has been applied in different fields, such as the detection of topics in collections of articles press (Figuerola et al., 2017). The LDA presents three levels for the representation, where the set of documents is called by the letter D, while $\theta^{(d)}$ is the multinominal distribution on the topics of the document D. The set $N_{(d)}$ denominates the set of words w for a specific document D, while z is the topic to which the word w is assigned. Finally, the set T represents the number of topics, where $\varphi^{(z)}$ is the multinominal distribution on the words for the topic z. For the model called LDA, the latent variables θ, φ, and z must be estimated together with the distributed Dirichlet hyperparameters α and β (Blei et al., 2003; Griffiths et al., 2005). The hyperparameters α and β should be interpreted as smoothing factors for assignments respectively from topic to document (θ) and from word to topic (φ).

=== Insert Figure 1 here ===

Topic Modelling Implementation

The free software Python 3.6 was used to implement the steps of pre-processing, topical modeling adjustment, model selection, and post-processing.

Pre-processing of text in this study includes the tokenization of words, conversion of words to upper-case letters, removal of characters and punctuation numbers, and removal of words considered as words of semantic connection (stopwords). Additionally, extra stopwords were added, which were garbage words resulting from processing steps.

The assembly of the model and natural language processing (NLP) consisted of estimating the latent variables θ, φ, and z, which was done using the Gesim 2.2.0 library (Rehurek and Sojka, 2010).

Results

The systematic search with time and publication type filters was performed using the electronic databases Scopus and ISI Web of Science, with last updated in December 2019. The search phrase was developed with the Boolean operators [OR] (between synonyms) and [AND] (between descriptors). Initially, 11,413 articles were identified. After removal of duplicates, 9,649 studies were used (Figure 2).

=== Insert Figure 2 here ===

The inclusion criteria for these articles were: (1) be related to the temporal issue, and thus, the criterion is that the study should be published in the last seven years (2012 to 2019) for analysis, and (2) inclusion of documents solely and exclusively by the type called an article by the two databases. Subsequently, all studies available in the database when researched were select for this study. Studies only excluded when presented as duplicates.

Impact Analysis

At first, the impact was assessed using the count of publications over the years of articles published in online databases. This is simply the number of articles that were published each year according to Figure 3, and by the journal, as presented in Figure 4.

Table 2 is a pure species of the journals extracted from the databases using the search term(s) sorted by the number of published articles present on that database. This table shows the top 10 journals, although the counts of all journals are written on a spreadsheet so the researcher can conduct further inspections and analyses.

Then, the author summarizes the articles to identify which researchers have the most significant impact. Table 3 shows the top 10 researchers (out of 9,649 articles) in the data set, according to the number of published works, total citations, and h-index. Although it is possible to sort data in author order and search for duplicate authors, the volume of data makes this awkward, and we accept that some "noise" is inevitable. The impact is typically low for author data and has little or no effect on the analysis of location and article citation or in topic modeling of abstracts. It can be seen, in Table 3, that the research in the field of performance analysis in soccer is growing, which shows the interest of several authors on the subject.

=== Insert Figure 3 and 4 here ===

Figure 4 shows the authors' preference for two databases (Journal of Strength and Conditioning Research and Journal of Sports Sciences) that present more than 470 articles published in this period of analysis and which may indicate a tendency of the themes related to this study area.

In Table 2, The ten journals that obtained the most significant number of publications during the study period were selected. Therefore, it presented some other impact metrics collected on May 12, 2018, from the respective agencies (Incites Journal Citation Reports, Scimago Journal & Country Rank, and CAPES) in *.cvs* format and included in the database.

=== Insert Table 2 here ===

The h-index was created in 2005 by Jorge E. Hirsch as an attempt to measure the impact of academic research. Hirsch (2015) presented an easily computable index, which provides an estimate of the importance, significance, and broad impact of a scientist's contributions, comparing, in an unbiased manner, different individuals competing for the same resource when a critical evaluation criterion is a scientific achievement.

In this way, Plos One is the journal that has the most significant impact in the community with an h-index factor of 268 in 2019, many citations, and consequently Eigenfactor Score much higher than the others. Plos One is a free-access scientific journal available only online, published by the Public Library of Science, which mainly covers primary research from any discipline in the field of science and medicine. In this way, Plos One is a journal that needs to be more considered by the authors of this field of study.

Three journals need special attention are Journal of Science and Medicine in Sport, International Journal of Sports Physiology and Performance, and Journal of Strength and Conditioning Research, which have the high impact factors JCR, SJR, and CAPES.

Journal Citation Reports (JCR) is a popular way to evaluate indexed journals on the Web of Science and is a crucial tool to help researchers determine where to publish their work and which journals to use in their research. A little different from the JCR, the SCImago Journal Rank (SJR) indicator is very similar to the Eigenfactor score, the first worked on the Scopus database and the second on the Web of Science database (Jacsó, 2010).

The h-index locating of "someone," several databases can be used. Thus, for the composition of the data of table 3, we used the software Harzing's Publish or Perish macOS GU Edition. This software was designed to empower academics and present research impact (Harzing, 2007). The software can be purchased free of charge from the website https://harzing.com/resources/publish-or-perish. The publications years established were 1990 to 2019 and used search to inspect by Scopus. The Scopus need a free registration required by API Key.

=== Insert Table 3 here ===

Structure Analysis

In addition to the worksheets used to produce Table 3, the CLR generates a full network view and author views (Figure 5). Figure 5a is a network-wide view of authors and co-authors with 9,650 articles. Figure 5b is a view of the author Clemente, F.M, who presents 700 or more published articles with their respective co-authors in database. Figures 5c and 5d present in more detail the network of the Clemente, F.M.

=== Insert Figure 5 here ===

It can observed that the authors that research on performance analysis in soccer do not present a homogeneous community, but several segments or niches that are probably determined by their lines of research. It can determine that some authors only develop their work with the same coauthors (the same form of collaboration). However, Clemente, F.M., in his network, presents a higher range of publications in partnership, which includes 168 different authors.

Hence, highlighted Clemente's author, from the Polytech Institute of Viana do Castelo, have secure connections and sharing of works on performance analysis with Martins, F.M.L. (with 88 works together) and Mendes, R. (with 60 works published together). Both are Portuguese researchers of the School of Higher Education, of the Polytechnic Institute of Coimbra, Portugal, in the field of Physical Education and Mathematics, respectively, which demonstrates the interest on the application of mathematical models in the analysis of performance in soccer.

Six different nationalities are among the ten most published authors from 2012 to

2019, shown in Table 3. Brazil appears among them with 3 researchers. Highlight Loturco, I. with 1,030 citations based on information collected by Harzing's Publish or Perish software and 435 articles in this database.

Content Analysis

The next task was to build a topic template for abstracts. At first, we performed an extensive data cleansing, which requires relatively little work in this regard, in addition to the standard case, blank, parting, and so on. However, there are still some essential concerns.

The predictive likelihood measures proposed to evaluate the quality of the generated topics. Nevertheless, its correlation is negative with human interpretation (Chang et al., 2009). In this way, data less consistent with a personal point of view created. This correlation is especially essential when generated topics are used in document collections to understand trends and development within a specific research area (Syed and Spruit, 2017). Röder et al. (2015) systematically and empirically explored the topic coherence measures and their correlation with the human topic classification data. Thus, their approach revealed a new measure of unexplored coherence denominate *CV*.

Similarly, Mimno et al. (2011) present a coherence new metric *UMass* where results can classify over the ROC curve area. *UMass* coherence is an asymmetric confirmation measure between major word pairs. Thus, a smoothed conditional probability and perplexity measurement is a predictive measure of the probabilistic model where a low perplexity indicates how good the probability distribution is in the sample (Brown et al., 1992).

First, the number of topics (K) to be used was determined. After the analysis of *CV* coherence (0.508), *UMass* (-4.842) and perplexity (-10.566) some experimentation, and considering the size of the data set, we selected a value of 20 (Figure 5).

=== Insert Figure 6 here ===

Second, the researcher chooses to remove some words from the data set because they have limited discriminatory value. For example, terms like "performance" or "analysis" occur in almost all extracted abstracts. Additionally, again based on visually inspecting the outputs, we also removed "noise" words, such as "p-value," "American," "et," "role," "however," among others. Although this does not necessarily limit the effectiveness of the model, it makes the results more challenging to interpret, as these terms appear in almost every topic as recommended by Mortenson and Vidgen (2016).

With these preliminary steps performed and defining some of the other required parameters, the model executed. The majority of the 20 topics extracted represent distinct research areas. As an example, it can be seen, in Table 4, the first five topics that show the three main most frequent terms for each topic, being the size of the word determined by the probability of the word and human interpretation of the distribution presented by the model.

=== Insert Table 4 here ===

Therefore, as in exploratory factor analysis, topic-modeling software does not include label topics – this is something the user must do based on the content of the topics. When working with a large number of documents, we can observe the size of the documents by topic. Thus, Figure 6 shows the number of documents against word distribution. Therefore, below (Figure 7) follows the information of 4 topics determined by the research and its relationship with the documents.

=== Insert Figure 7 here ===

A fascinating visualization technique is provided by the pyLDAvis package, which is a Python library for viewing interactive topic templates based on the package written in R. This package provides an overview of all topics, shows the differences between topics, and allows the researcher to read the most highly associated terms for each topic individually. It is a powerful tool that allows the user to examine specific topics, keeping the entire topic scenario on display, and therefore useful to the user when interpreting and labeling topics (Sievert and Shirley, 2014).

Figure 7 shows the pyLDAvis interface. Selecting a topic on the left side (in this case, the topic that seems to address subjects related to physical training types) highlights the most useful terms for interpreting the selected topic on the right side. On the other hand, selecting a term on the right side exposes the conditional distribution on the topics to the left of the selected term.

It is possible to see the research topics related to several areas of performance analysis, such as injuries, strength, and distance, not showing any type of stratification by sex, which demonstrates that the issues through performance analysis in women's football are still incipient. Thus, we highlight topic one related to athlete development and coach support, the topic related to the risk of injury present in all sports, and topic three related to the physical demands on training and games. Other topics can be viewed interactively, for view this use on an IPython notebook, but can also be saved in a standalone HTML file for easy sharing and distribution, as it can check at http://bit.ly/LDA_football.

=== Insert Figure 8 here ===

After running the topic template, there will be a set of probabilities for each of the articles against the chosen topic numbers. The later probabilities, inferred from the model, demonstrate the "topical distribution" of each document. A document has a probability of 0.5 for a given topic, and this suggests that around 50% of the content of the document is related to that topic.

Study limitations

The study limitations come from a more robust data cleansing model since when working with articles abstracts, the number of 'noise' words is quite high and also the insertion of other sports within the same search context as rugby, Australian football (AFL), handball and others. The missing data and errors in the data insertion in the Scopus and ISI Web of Science database also appear as a limitation. Another limitation was the use of only two sources of information (2 online databases). Although Scopus and ISI Web of Science index a large number of scientific journals worldwide, it does not represent all publications about football performance analysis from 2012 to 2019. So, like not using a more optimized process, for example, using the API of these scientific research platforms.

Conclusion

The CLR offers a fully functional and automated tool that allows the researcher to evaluate large volumes of data on the existing literature regarding impact, structure, and content. Consequently, the CLR offers an approach that may provide greater validity within the academic context on literature reviews. When performed in similar datasets, the CLR results are replicable, and their approach is transparent, providing a more objective way to determine the relevance and importance of the sources.

Another approach is related to the speed and productivity of research stemming from how academic research can encompass the opportunities of more sophisticated data analysis and the use of large volumes of data in a consistent manner. Although growing in organizations of all kinds, data analysis, conducted using artificial intelligence, has its application in academia, particularly in Physical Education research as an area that is still little explored.

Unusual in the Physical Education setting, programming languages can assist both systematic review and applicability within the sports context. From this context, the review study identified yet unexplored gaps when considering performance analysis within football. A tiny amount of studies has addressed soccer players, not present for the word woman or significant female probability. Also, we found no studies related to goalkeeper function. Several studies are addressing physical issues, mainly with the use of technologies such as global positioning system (GPS) for both professional and young athletes.

As Big Data applications continue to grow in influence in the community as well as the opportunities it offers to conduct new methods of analysis, these professionals' skills may also be more valued. The use of knowledge in the fields of mathematics, machine learning, and artificial intelligence can develop the ability and confidence to use algorithms, through software such as Python, which includes the CLR to support literature reviews with more agility and efficiency.

Acknowledgment

This work was supported by the Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro [grant number E-26/202.638/2018].

References

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1109/MSP.2010.938079

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Bornmann, L., & Mutz, R. (2015). Growth Rates of Modern Science : A Bibliometric

Analysis Based on the Number of Publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. https://doi.org/10.1002/asi

Brown, P. E., Pietra, V. J. Della, Mercer, R. L., Pietra, S. a Della, & Lai, J. C. (1992). An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics*, *10598*(1), 31–40. http://acl.ldc.upenn.edu/J/J92/J92-1002.pdf

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 288–296.

Chen, Z., Zhong, F., & Yuan, X. (2016). Framework of Integrated Big Data: A Review. *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, 1–5. https://doi.org/10.1109/ICBDA.2016.7509815

Felizardo, K. R., Salleh, N., Martins, R. M., Mendes, E., MacDonell, S. G., & Maldonado, J. C. (2011). Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews. *2011 International Symposium on Empirical Software Engineering and Measurement*, 77–86. https://doi.org/10.1109/ESEM.2011.16

Figuerola, C. G., García Marco, F. J., & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics*, *112*(3), 1507–1535. https://doi.org/10.1007/s11192-017-2432-9

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. *Advances in neural information processing systems*, *17*, 537–544. https://doi.org/10.1.1.73.9813

Harzing, A. W. (2007). Publish or Perish. https://harzing.com/resources/publish-or-perish

Hirsch, J. E. (2015). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.2336195100

Horta, V., Ströele, V., Braga, R., David, J. M. N., & Campos, F. (2018). Analyzing scientific context of researchers and communities by using complex network and semantic technologies. *Future Generation Computer Systems*, *89*, 584–605. https://doi.org/10.1016/j.future.2018.07.012

Jacsó, P. (2010). Comparison of journal impact rankings in the SCImago Journal & Country Rank and the Journal Citation Reports databases. *Online Information Review*, *34*(4), 642–657. https://doi.org/10.1108/14684521011073034

Jahangirian, M., Eldabi, T., Garg, L., Jun, G. T., Naseer, A., Patel, B., et al. (2011). A rapid review method for extremely large corpora of literature: Applications to the domains of modelling, simulation, and management. *International Journal of Information Management*, *31*(3), 234–243. https://doi.org/10.1016/j.ijinfomgt.2010.07.004

Jennex, M. E. (2015). Literature Reviews and the Review Process : An Editor-in-Chief ' s Perspective. *Communications of the Association for Information Systems*, *36*, 139–146.

La Rosa, M., Fiannaca, A., Rizzo, R., & Urso, A. (2015). Probabilistic topic modeling for the analysis and classification of genomic sequences. *Bmc Bioinformatics*, *16*(Suppl 6), 9. https://doi.org/10.1186/1471-2105-16-s6-s2

Lee, H., Kwak, J., Song, M., & Kim, C. O. (2014). Coherence analysis of research and education using topic modeling. *Scientometrics*, *102*(2), 1119–1137. https://doi.org/10.1021/acsnano.7b00569

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine*, *6*(7). https://doi.org/10.1371/journal.pmed.1000100

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. https://doi.org/10.1147/rd.22.0159

McLevey, J., & McIlroy-Young, R. (2017). Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *Journal of Informetrics*, *11*(1), 176–197. https://doi.org/10.1016/j.joi.2016.12.005

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (2), 262–272.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, *339*, b2535. https://doi.org/10.1136/bmj.b2535

Mortenson, M. J., & Vidgen, R. (2016). A computational literature review of the technology acceptance model. *International Journal of Information Management*, *36*(6), 1248–1259. https://doi.org/10.1016/j.ijinfomgt.2016.07.007

Muschelli, J. (2018). Gathering bibliometric information from the Scopus API using rscopus. *R Journal*.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2 PART 2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Pham, B., Bagheri, E., Rios, P., Pourmasoumi, A., Robson, R. C., Hwee, J., et al. (2018). Improving the conduct of systematic reviews: a process mining perspective. *Journal of Clinical Epidemiology*, *103*, 101–111. https://doi.org/10.1016/j.jclinepi.2018.06.011

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. https://doi.org/10.13140/2.1.2393.1847

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 399–408. https://doi.org/10.1145/2684822.2685324

Rose, M. E., & Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX*, *10*, 100263. https://doi.org/10.1016/j.softx.2019.100263

Rowley, J., & Slack, F. (2004). Conducting a literature review. *Management Research News*, *27*(6), 31–39. https://doi.org/10.1108/01409170410784185

Shimada, D., Kotani, R., & Iyatomi, H. (2016). Document classification through image-based character embedding and wildcard training. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 3922–3927. https://doi.org/10.1109/BigData.2016.7841067

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.1.1.100.1089

Simsek, A., & Kara, R. (2018). Using swarm intelligence algorithms to detect influential individuals for influence maximization in social networks. *Expert Systems with Applications*, *114*, 224–236. https://doi.org/10.1016/j.eswa.2018.07.038

Syed, S., & Spruit, M. (2017). Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation. *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017*, *2018-January*, 165–174. https://doi.org/10.1109/DSAA.2017.61

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review, *14*, 207–222. https://doi.org/10.1111/1467-8551.00375

van Altena, A. J., Moerland, P. D., Zwinderman, A. H., & Olabarriaga, S. D. (2016). Understanding big data themes from scientific biomedical literature through topic modeling. *Journal of Big Data*, *3*(1). https://doi.org/10.1186/s40537-016-0057-0

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786. https://doi.org/10.1007/s11192-014-1321-8