

Topic Modeling: how and why to use in management research

Abstract

Topic modeling can be a valuable research approach to generate management theory from textual data. They provide, without pre-defined labels, an automated procedure for coding the content of texts into a set of substantively meaningful coding categories called "topics". Topic Modeling is able to match the results of human-coded text analysis, making it an important toolkit for the 21st century Big Data-embedded scenario. This article has two sections. In the first, I map critical published studies in social sciences that employed properly topic modeling. Secondly, I illustrate how to do topic modeling by applying topic modeling in an analysis of the last five years of published research in the Iberoamerican Journal of Strategic Management. Topic Modeling can empower researchers in their theory-building process by allowing textual data to be used in a quantitative approach in follow-up analyses. This can be crucial in shifting the paradigm that textual data can only be framed as qualitative research.



Key-words: topic modeling, latent Dirichlet allocation, computer-aided text analysis, machine learning, big data.


1. Introduction

Big Data is defined as "large-scale data streams taken from the Internet, social media sites, or archives" (Morh & Bogdanov, 2013, p. 561). It can be opportunistically used in research because it proportionates access to almost unlimited data. Most of Big Data is comprised of textual uncategorized data. Analyzing such unstructured data can be a challenge since researchers cannot apply traditional textual analysis (coding for instance) given the large dimension of textual data (Bendle & Wang, 2016). Topic modeling (Hannigan et al., 2019), and specifically **Latent Dirichlet Allocation (LDA)** (Blei et al., 2003) can analyze huge amounts of text and describe the content as focusing on unseen attributes in a specific weighting. The use of computer-aided text analysis (CATA) in Management and Social Sciences literature is growing (Nelson, 2017; Nelson, Burk, Knudsen, & McCall, 2018) and can complement, if not fully **replace** (Baumer et


al., 2017), traditional text analysis approaches.  Management researchers could avail to keep Topic Modeling approaches in their toolkit.

My aim is to introduce topic modeling as a valuable research approach to generate management theory from textual data. A researcher can, through topic modeling, label and categorize textual data in order to generate a quantity or measure that can be later used for statistical analysis and hypothesis testing. Textual data, which was mostly used in management research in an exploratory and qualitative approach, can be employed in a descriptive and quantitative approach. Researchers can use topic modeling to shift the paradigm of textual data from qualitative propositions to quantitative hypothesis.

To exemplify how topic modeling can be used in management research, my objectives  are two-fold. First, I introduce topic modeling as a social sciences research tool and map critical published studies in management and other social sciences that employed topic modeling in a  proper manner. Second, I illustrate how to do topic modeling by applying topic modeling in an analysis of the last five years of published research in this journal: the Iberoamerican Journal of Strategic Management (IJSM). I treated every abstract from each article published as a document and derived 6 topics. The main results show that competitive advantage along with entrepreneurship once predominant are declining, and that international and finance are increasing in importance. The *R* and *Python* code used for all data collection, data analysis, images, and tables generation are available in an online *Open Science Framework* public repository (Storopoli, 2019).

I contribute  apping the core concepts around Topic Modeling to guide researchers in their future endeavors. Also, I show a curated sample of good examples of topic modeling research articles for further inquiries. I continue my approach towards Topic Modeling by explaining what are the main procedures and precautions of conducting a Topic Modeling analysis. Moreover, I illustrate which further quantitative analyses the researcher can do with the labeling and quantification of textual data generated through Topic Modeling. Finally, I apply Topic Modeling to a sample of five years of published articles in IJSM to demonstrate how the technique can be useful for management research with unlabelled or uncategorized textual data.

2. Topic Modeling

To address my first objective, this section introduces topic modeling defining it and giving a brief overview. Next, I explain the main technique behind topic modeling: Latent Dirichlet Allocation (LDA)  Furthermore, I provide the main benefits of topic modeling and continue the presentation by addressing prescriptive issues such as: how to do topic modeling and how to choose the best model. Then, I move to more theoretical issues like how to build theory with topic models; and, finally, I conclude with a curated selection of examples of published research in social sciences that have employed topic modeling.

2.1. Origins and Definition

Topic modeling is a class of text analysis that has arisen with the advent of both machine learning and big data. Today personal computers have a substantially sizeable computational power than when they first appeared in the early 1950s. For example, the computer responsible for guiding the Apollo mission to the Moon in 1969 had only 2 kilobytes (KB) of memory size (RAM). This may sound infimal since many of today's smartphones have at least 4 gigabytes (GB) of RAM and mostly commercially-available notebooks can have up to 16GB of RAM. This massive computational power makes it possible for many scholars and researchers to do heavy bouts of data analysis. In the beginnings of quantitative data analysis, it was impossible to run it on a personal computer, and the researcher had to have access to a mainframe computer--mostly located in universities, research institutes or resourceful firms. So the rise of the computational power of personal computers gave freedom to run massive data analysis on your lap.

Big Data, the second factor behind the advent of Topic Modeling, is defined as "large-scale data streams taken from the Internet, social media sites, or archives" (Morh & Bogdanov, 2013, p. 561). It can be opportunistically used in research because it proportionates access to almost unlimited data, with unfathomed proportions. By allying accessible high computation power with an unlimited supply of data, researchers could run complex algorithms and quantitative techniques to large text data in order to generate a body of broad categories and comprehensive analysis to understand what underlying phenomena may be behind the data.

Topic modeling is an instance of probabilistic modeling (Mohr & Bogdanov, 2013) and uses statistical associations of words in a text to generate latent topics—clusters of co-occurring words that jointly represent higher-order concepts—but without the aid of pre-defined, explicit

dictionaries or interpretive rules (Hannigan et al., 2019). It does so not as providing an automatic text analysis application but rather as providing a lens that allows researchers working on a problem to view a relevant textual data in a different light and at a different scale (Morh & Bogdanov, 2013). They provide, without pre-defined codes or categories of meaning, an automated procedure for coding the content of texts (including abundant textual data obtained by Big Data) into a set of substantively meaningful coding categories called "topics". The most used topic modeling technique is the **Latent Dirichlet Allocation** (LDA): a generative probabilistic model for collections of discrete data such as textual data (Blei et al., 2003). In LDA, each topic can be viewed as a theme because they are a set of distribution over all observed words in the texts; in other words, a bag-of-words that frequently appear together across documents. Moreover, every document (text) analyzed can have topic probabilities that stipulate which main topics they are mostly associated with. For technical statistical specifications, LDA assumes that each length of documents is Poisson distributed and the proportion of the document in each topic is Dirichlet distributed. Dirichlet distributions are commonly used as prior distributions in Bayesian statistics.

The benefits of Topic Modeling are plenty. First, they do not impose the researcher dictionaries or interpretative rules regarding the data, enabling identification of important themes that human readers are unable to discern. Also, it allows for polysemy because topics are not mutually exclusive; individual words appear across topics with differing probabilities, and topics themselves may overlap or cluster (DiMaggio, Nag & Blei, 2013). Second, when dealing with a large extent of textual data (such as in a Big Data scenario), it provides a way for researchers to obtain reasonable automated content coding, enabling to take the measure of large-scale social phenomena that we could not have previously been able to do (Morh & Bogdanov, 2013). Third, it removes the burden on the research from manually coding text data to interpret and validate the results of topic models, epitomizing a shift from interpretive methods borrowed from the humanities to disciplining the results through statistical validation (DiMaggio, 2015).

So how does **Topic Modelling** CATA technique, fare against a traditional human-powered text analysis? Surprisingly well, some might argue. In a study that compared Topic Modeling (Blei et al., 2003) versus Grounded Theory (Glaser & Strauss, 1967) the authors identified several correspondences between the grounded theory themes and the algorithmically

generated topics (Baumer et al., 2017). In another study, Topic Modeling was able to match the results of human-coded text analysis. The consequence of CATA applied as the primary approach to qualitative research can result in "an efficient, rigorous, and fully reproducible computational grounded theory" (Nelson, 2017, p. 32). The CATA framework can also be applied to any qualitative text as data, including transcribed speeches, interviews, open-ended survey data, or ethnographic field notes, and can address many potential research questions (Nelson, 2017). This change to Topic Modeling requires many of us, social scientists, to move outside our comfort zone in accepting interpretive uncertainty and to develop robust ways to interpret and validate the results of our models (DiMaggio, 2015).

2.2. How to do Topic Modeling

The first procedure is to collect data, which in topic modeling means textual data. This step can be done in several ways: by transcribing interviews, rendering reports to text, web scraping; or any other source of collecting and generating textual data. It is important to note that the number of texts in a sample can be decisive in a Topic Modeling analysis. The most common approach is to have each document as an individual text in a sample. However, sometimes the sample is comprised of one large mass of text, such as an entire book or a lengthy interview. The researcher can either choose to keep the sample to 1 text or to break up the document in chunks that are interdependent---an interview can be broken by topics and a book in its chapters.

Once the researcher has the data he or she needs, it is time to pre-process the data. Pre-processing the data means performing upon it some fundamental transformations, in order to have data that will be much more useful for performing some further, more meaningful analysis (Denny & Spirling, 2018). The first pre-processing procedure is text normalization, which is a set of transformations with the purpose to render textual data that can be quantified and compared within itself. This transformation includes: (1) converting all letters to lower or upper case; (2) converting numbers into words or removing numbers; (3) removing punctuations, accent marks, and other diacritics; (4) removing white spaces---leading and ending spaces in a text; and (5) expanding abbreviations. The second pre-processing procedure is the removal of "stop words", defined as the "most common words in a language like 'the', 'a', 'on', 'is', 'all'" (Debortoli et al., 2016, p. 111). These words do not carry significant meaning and are usually removed from texts.

The third and final pre-processing procedure is called stemming or lemmatization. Stemming is reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form; and one of the most applied stemming algorithms is Porter's stemming algorithm (Porter, 1980).

After the text is pre-processed, the researcher can carry on with the Topic Modeling by applying an algorithm to render a probabilistic model. The simplest and most widely used model is **Latent Dirichlet Allocation** (LDA) introduced by Blei et al. (2003). LDA inputs are: (1) a set of documents that can be represented as a document-word matrix (DTM); (2) the number of topics to be estimated by the algorithm. The DTM is a matrix in which the rows are each document in the sample, the columns are each unique word in the sample, and the cells are the number of times each word occurs in each document. The second input is the number of topics to be estimated by the algorithm. This manual input makes the topic quantity selection to be an obstacle to the researcher. Most researchers deal with it by generating a model for each number of topics and analyzes which ones have the best or optimal congruence to the data. The LDA outputs are: (1) a topic-word matrix; and (2) topic-document matrix. Both matrices are vectors of weights, with the topic-word being the weights of words in each topic and the topic-document weights of topics in each document. The main operation behind the LDA algorithm is vector space calculations based on similarity comparisons while using the inputs in order to generate the outputs. LDA assumes that the similarity comparisons are probabilistic in nature and that each word in a document is modeled as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of 'topics'.

2.3. How to choose the right model

With no fixed number of topics to be generated, researchers have to generate one model for each number of topics and then assess which one is the right model. The assessment can be either quantitative or qualitative. The quantitative metrics are a measure of the fit of the model, and the qualitative assessments are interpretative and discretionary to the researcher. One must not solely base his analysis on quantitative metrics since they can be unreliable (Maier et al., 2018). Topic models that perform better on quantitative metrics tend to infer topics that humans judge to be semantically less meaningful (DiMaggio, 2015). There is no statistical test for the optimal

number of topics or the quality of a solution. The point is not to estimate population parameters correctly but to identify the lens through which one can see the data most clearly.

Regarding quantitative metrics about the model fit, there are three: (1) perplexity; (2) log-likelihood; and (3) coherence. Perplexity is a global indicator of the model, and it represents the model's "surprise" at the data (Blei et al., 2003). A lower perplexity score indicates better generalization performance. Log-likelihood is also a global indicator and represents how plausible model parameters are given the data (Bordag, 2008). Coherence is a metric for assessing topic quality; it is a local indicator for each topic that examines the words in topics, decide if they make sense (Mimno et al., 2011). Coherence is noted as the optimal quantitative metric for **topic modeling** assessment (DiMaggio et al., 2013). Indeed, a statistical test for an overall solution (as opposed to for the quality of particular topics) would be misleading, because models often shunt noisy data into uninterpretable topics in ways that strengthen the coherence of topics that remain. Thus, the test of the model as a whole is its ability to identify a number of substantively meaningful and analytically useful topics, not its success in optimizing across all topics.

Qualitative assessment is based on two types of validity: internal and external (DiMaggio, 2015). Semantic or internal validity confirms that the model meaningfully discriminates between different senses of the same or similar terms. This validity answers the following question: "how can we be sure that our interpretation of the meaning of a topic is better than an alternative interpretation?". Predictive or external validity determines whether particular topics correspond to information external to the topic model (e.g., by confirming that certain topics became more salient when an external event relevant to those topics occurred). It recognizes that the same text will speak in different ways and be interpreted differently by different audiences. The researcher must locate the optimal balance between the two logics of validity.

2.4. Building Theory from Topic Models

In order to generate theory from topic models, one must understand that the main contribution of **topic modeling** is the development of a system of inductive classification. In a classic scenario of content classification and coding (mostly textual content), researchers usually are looking for shared structures of meaning that are not formally materialized. Topic modeling

can emulate the same purpose of finding these shared structures but without introducing researcher bias -- note that the only human intervention would be to choose the number of topics to be generated for the model.

The iteration between theory and the topics that emerge from the chosen model create new theoretical artifacts or build theory with them. A researcher must always ask whether such topics represent meaningful structures. That is, for every topic, one must ponder if it resonates theoretically with the content from which it was derived and also if it provides substantial theoretical contributions and discussions.

Presenting topics without particular concern for theoretical artifacts risks presenting disembodied arguments about the artifacts' importance and role regarding the theory and the data. If one naively apply topic modeling crudely, one may omit essential distinctions on how to capture an essential meaning and meaning structures in the data to generate significant theoretical discussions and contributions.

Topic modeling may not be the final destination of analysis and theory building in a study. Researchers may use topic modeling as a means to generate unbiased classifications and metrics of textual (qualitative) data. Textual data can be then measured and used in quantitative analysis, especially in hypothesis testing. It shifts the paradigm and assumption that textual data belongs only in the realm of qualitative analysis and exploratory research settings. Researchers may find new and innovative ways of measuring variables in order to test hypotheses in contexts that were not delved before.

2.5. Good Examples of Topic Modeling

Researchers have been applying topic modeling and other CATA since its inception in 2003 (Blei et al., 2003). In table 1, I present a curated sample of 23 good examples of topic modeling research articles. Most of the articles are from management (65%), but there are some from other social sciences. Also, the articles' sample sizes are notoriously large: the median is 8,000, and the mean is 18,063.

Table 1. Critical studies using topic modeling

Area of Knowledge	Article	Type of Probabilistic Model	Object	Sample Size	Number of Topics	Further Analysis	Programming Language	Package	Main Findings
Management	Dong, C., & Zhang, Y. (2019). NPOs' Voice in CSR Partnership: An Exploratory Study Using Topic Modeling. <i>International Journal of Business Communication</i> , 00(0), 1–21. https://doi.org/10.1177/2329488418819136	LDA	65 Nonprofit Organizations (NPOs)	5,661 tweets	8	ANOVA	R	stm	Identified three corporate social responsibility communication strategies, which were characterized by a distinctive emphasis on stakeholder engagement.

Management	Haans, R. F. J. (2019). What's the value of being different when everyone is? The effects of distinctiveness on performance in homogeneous versus heterogeneous categories. <i>Strategic Management Journal</i> , 40(1), 3–27. https://doi.org/10.1002/smj.2978	LDA	2,279 firms	69,188 organizational websites	100	logistic regression	-	-	U-shaped effect in benefits of differentiation from competitors.
Information Technology	Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. <i>Decision</i>	LDA	1 automobile insurance company	37,082 insurance claims	5	t-test	-	-	Extraction the text features hiding in the text descriptions of the accidents appearing in the claims for detecting fraudulent claims.

	<p><i>Support Systems</i>, 105, 87–95.</p> <p>https://doi.org/10.1016/j.dss.2017.11.001</p>								
Management	<p>Guo, L., Sharma, R., Yin, L., Lu, R., & Rong, K. (2017). Automated competitor analysis using big data analytics. <i>Business Process Management Journal</i>, 23(3), 735–762.</p> <p>https://doi.org/10.1108/BPMJ-05-2015-0065</p>	LDA	535 fitness apps	535 fitness apps	535	cluster analysis	-	-	Algorithm to reveal a fitness mobile app's position in relation to its peers.

Management	<p>Cho, Y.-J., Fu, P.-W., & Wu, C.-C. (2017). Popular Research Topics in Marketing Journals, 1995–2014. <i>Journal of Interactive Marketing</i>, 40, 52–72.</p> <p>https://doi.org/10.1016/j.intmar.2017.06.003</p>	LDA	25 journals	17,249 research papers	100	social network analysis	R	lda	The most impactful journals are the most diverse, whereas each runner-up has a unique focus.
Management	<p>Zou, H., Chen, H. M., & Dey, S. (2015). Exploring user engagement strategies and their impacts with social media mining: the case of public libraries. <i>Journal of Management Analytics</i>, 2(4), 295–</p>	LDA	10 public libraries	10,000 tweets	4	random forests machine learning algorithm	JAVA	MALLET	Identify user engagement strategies are used by libraries on Twitter, and suggest the best practices for libraries interested in pursuing social media initiatives to use to engage their users effectively.

	313. https://doi.org/10.1080/23270012.2015.1100969								
Management	<p>Wilson, A. J., & Joseph, J. (2015). Organizational Attention and Technological Search in the Multibusiness Firm: Motorola from 1974 to 1997. <i>Advances in Strategic Management</i>, 32, 407–435.</p> <p>https://doi.org/10.1108/S0742-332220150000032013</p>	LDA	Motorola patents	12,787 patent backgrounds	100	logistic regression	R	topic models	Motorola subunits with specialized attention are not myopic but instead explore broadly and tight attentional coupling across units increases the breadth of search.

Management	<p>Haans, R. F. J., & van Witteloostuijn, A. (2019). Regional stickiness of novel ideas in the scholarly international business community. <i>Cross Cultural & Strategic Management</i>, 26(2), 145–165.</p> <p>https://doi.org/10.1108/CCSM-07-2018-0102</p>	LDA	Journal of International Business Studies	1,525 articles	100	logistic regression	-	-	Strong path dependency between the geographic origin of topics and their spread across the world. This suggests the existence of geographically narrow mental maps in the field, which the authors find have remained constant in North America, widened yet are still present in East Asia, and disappeared in Europe and other regions of the world over time.
------------	---	-----	---	----------------	-----	---------------------	---	---	--

Management	Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. <i>European Research on Management and Business Economics</i> , 24(1), 1–7. https://doi.org/10.1016/J.IEDEEN.2017.06.002	LDA	"Big Data" search in Scopus (business only)	1,560 articles	27	-	R	topic models	Research is bipartite between technological and research domains, with Big Data publications not clearly aligning cutting edge techniques toward Marketing benefits. Big Data applications to Marketing is still in an embryonic stage.
Management	Kaur, J., Dara, R. A., Obimbo, C., Song, F., & Menard, K. (2018). A comprehensive keyword analysis of online privacy policies. <i>Information</i>	LDA	online privacy policies	2,000 policies	9	Jaccard similarity	-	-	Regulations have an impact on the selection of terminologies used in the privacy policies. The results also suggested that

	<i>Security Journal: A Global Perspective</i> , 27(5–6), 260–275. https://doi.org/10.1080/19393555.2019.1606368								European policies use fewer ambiguous words but use more words such as cookie and compliance with the regional regulations.
Management	Hannigan, T., Haans, R. F. J., Vakili, K., Tchaljian, H., Glaser, V., Wang, M., ... Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. <i>Academy of Management Annals</i> . https://doi.org/10.5465/annals.2017.0099	LDA	66 articles from a "topic model*" search in Web of Science and Scopus (business only)	5,362 paragraphs from the 66 articles	35	-	Python	gensim (using MALLET wrapper)	Identify and discuss how topic modeling has advanced management theory in five areas: detecting novelty and emergence, developing inductive classification systems, understanding online audiences and products, analyzing frames and social movements, and

									understanding cultural dynamics.
Management	Lee, H., & Kang, P. (2018). Identifying core topics in technology and innovation management studies: a topic model approach. <i>The Journal of Technology Transfer</i> , 43(5), 1291–1317. https://doi.org/10.1007/s10961-017-9561-4	LDA	11 technology and innovation management journals	11,693 articles	50	-	R	topic models	Shift of technology and innovation management research focus from topics of general management to those of technology and innovation management. It was also revealed that the changes of editor-in-chief caused considerable changes in topic portfolios in the technology and innovation

									management journals.
Management	<p>Szekely, N., & Vom Brocke, J. (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. <i>PLoS ONE</i>, 12(4),</p>	LDA	<p>3,906 organizations' sustainability reports</p>	<p>9,514 sustainability reports</p>	70	-	<i>Python</i>	gensim (using MALLET wrapper)	<p>Ten propositions for future research and practice that are of immediate value for organizations and researchers.</p>

	e0174807. https://doi.org/10.1371/journal.pone.0174807								
Political Science	Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. <i>Political Analysis</i> , 23(2), 254–277. https://doi.org/10.1093/pan/mpu019	STM	101 prominent Jihadist and non-Jihadist Muslim clerics	27,248 islamic texts	15	-	R	stm	Evidence of a trade-off for many Jihadists between focusing on fighting the West and focusing on excommunicating fellow Muslims they feel are inadequately supporting the Jihadist cause.

Consumer Behavior	Geva, H., Oestreicher-Singer, G., & Saar-Tsechansky, M. (2019). Using Retweets When Shaping Our Online Persona: Topic Modeling Approach. <i>MIS Quarterly</i> , 43(2), 501–524. https://doi.org/10.25300/MISQ/2019/14346	LDA	Twitter users	2,850 tweets	50	-	JAVA	MALLET	In the course of constructing their full personas, users tend to use the retweet option to enrich their self-produced personas in terms of depth, retweeting more about topics that they themselves already discuss.
Political Science	Ryoo, J. (Jun H., & Bendle, N. (2017). Understanding the Social Media Strategies of U.S. Primary Candidates. <i>Journal of Political Marketing</i> , 16(3–4), 244–266.	LDA	candidates' Facebook messages	14,386 Facebook posts	3-5 depending on each candidate	-	R	topic models	Clinton's focus on Trump increases toward the end of the primary campaign.

	https://doi.org/10.1080/15377857.2017.1338207								
Management	<p>Qiao, D., Zhang, J., Wei, Q., & Chen, G. (2017). Finding competitive keywords from query logs to enhance search engine advertising. <i>Information & Management</i>, 54(4), 531–543.</p> <p>https://doi.org/10.1016/j.im.2016.11.003</p>	LDA	Search engines' keyword association recommendations	8,500 query logs	100	-	-	-	Topic based competitive keywords suggestion method called TCK to enhance search engine advertising.

Management	<p>Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. <i>Technological Forecasting and Social Change</i>, 115, 131–142. https://doi.org/10.1016/j.techfore.2016.09.028</p>	LDA	mobile telecommunication industry's patents	157,718 patents	75	cluster analysis and social network analysis	<i>Python</i>	numpy	Company-specific differences in their knowledge profiles, as well as show the evolution of the knowledge profiles of industry leaders from hardware to software focussed technology strategies.
Economics	<p>Jelveh, Z., Kogut, B., & Naidu, S. (2014). <i>Political Language in Economics</i>. SSRN. https://doi.org/10.2139/ssrn.2535453</p>	CTM	academic writings by economists	62,888 articles	30, 50 and 100	logistic regression	-	-	Show considerable sorting of economists into fields of research by predicted partisanship, and yet can detect differences in partisanship among

									economists even within a field, even across those estimating the same theoretical parameter.
Poetics	Mohr, J. W., Wagner-Pacifici, R., & Breiger, R. L. (2013). Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics. <i>Poetics</i> , 41(6), 670–700. https://doi.org/10.1016/J.POETIC.2013.08.003	LDA	U.S. “National Security Strategy” documents	11 documents	15	-	-	-	15 topics: Terrorism, Treaties, Human Rights, Soviet/Europe, Economic Development, Energy, WMD, Conflict, Military Force, Trade, Peace, Intelligence, Global Security Strategy, Cooperation, Military Operations.

Poetics	Marshall, E. A. (2013). Defining population problems: Using topic models for cross-national comparison of disciplinary development. <i>Poetics</i> , 41(6), 701–724. https://doi.org/10.1016/J.POETIC.2013.08.001	CTM	articles from leading demographic journals	3,458	75	-	Python	numpy	Demographic research agendas reflected both cultural and institutional differences that shaped different understandings of fertility decline. While British demography focused on high-fertility contexts, French demography focused on lower-fertility contexts.
Poetics	DiMaggio, P., Nag, M., & Blei, D. M. (2013). Exploiting affinities between topic modeling and the sociological perspective on	LDA	press accounts of public support for the arts in the U.S.	8,000 newspapers articles	12	-	-		Differences in topic prevalence were driven by both the stories the papers covered and the ways they covered them, due to varying

	<p>culture: Application to newspaper coverage of U.S. government arts funding. <i>Poetics</i>, 41(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004</p>								missions, different news beats, and, possibly, differences in political orientation.
Management	<p>Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. <i>Strategic Management Journal</i>, 36(10), 1435–1457. https://doi.org/10.1002/smj.2294</p>	LDA	nanotechnology patents	2,826 patents abstract	100	structural equation modeling	R	topic models	Patents that originate new topics are more likely to be associated with local search, while economic value is the product of broader recombinations as well as novelty.

Regarding the type of probabilistic model employed by the articles is LDA by the majority (87%) with a few exceptions. Two studies employed **correlated topic modeling** (CTM). CTM topic proportions allow for topic correlation (Blei & Lafferty, 2007), admitting that some topics are closer to each other and share words with each other (Nikolenko, Koltcov & Koltsova, 2017). CTMs use logistic normal distribution instead of the Dirichlet allocation from LDA to model correlations between topics. One study employed structural topic modeling (STM). The STM provides a flexible way to incorporate metadata associated with the text into the analysis, such as: when the text was written; where (e.g., which country) it was written; who wrote it; and characteristics of the author (Robert et al., 2014). In turn, it allows the understanding of relationships between metadata and topics in their texts (Lucas et al., 2015).

The studies in **table 1** usually employ either a large number of topics (75, 100 or more) or a small number of topics (less than 20). This polarization of choice of the number of topics to be generated in topic modeling is caused by a split in the influences that the studies based their analyses. If a researcher comes from a traditional social sciences background, he or she will usually choose a small number of topics in order to focus on a comprehensive and meticulous description and discussion of each topic. On the other hand, if the background of the researcher is a more **quantitative "hard science"**, then the number of topics will be higher because of the focus might be getting the best model fit to the data (this is generally achieved by an immense number of topics in a model). There is no consensus in the prescription of the number of topics that a researcher must anchor their decision. Some argue that it depends on the level of 'resolution' a social scientist desires to obtain (Nikolenko, Koltcov & Koltsova, 2017). While others argue that it depends on the performance metrics of the model: such as perplexity (Blei et al., 2003) or coherence (Mimno et al., 2011). Perhaps a more sensible approach would be to let the performance metrics guide you but leave the final decision pending a thorough inspection by the researchers (DiMaggio, 2015).

Some studies depicted in **table 1** also employ further quantitative analysis of the data generated by topic models. As already covered in previous sections, researchers can apply topic modeling to textual data to generate non-biased categories and labels. These measures may be later included in quantitative analysis to test hypotheses. Most of the studies in table 1 employ a

supervised statistical analysis, which means that a dependent variable is guiding the analysis (ANOVA, *t*-tests, regressions, structured equation modeling, etc.). There is also a minority of studies that are concerned about unsupervised statistical analyses that do not have an important variable to guide the analysis, but the focus is to find similarities amongst their sample (clustering, social networks, measures of similarities, etc.).

Finally, topic modeling can be freely applied using either *R*, *Python* or *Java*. For *R* there are three main packages: 'topicmodels' (Hornik & Grün, 2011); 'lda' (Chang, 2011); and 'stm' (Roberts, Stewart & Tingley, 2014). The most updated of these is the 'topicmodels', which can apply either LDA or CTM as types of probabilistic model, but not STM. For STM, researchers working in *R* ecosystem must use the 'stm' package. The least updated of the *R* packages is 'lda' that can only perform the LDA type of probabilistic models. Researchers looking for topic modeling in *Python* setting must turn their efforts towards the 'gensim' library (Rehurek & Sojka, 2010). The 'gensim' is more frequently updated than any of the *R* packages while being used by firms such as Amazon, Cisco and Capital One. 'gensim' can only perform LDA, it cannot perform STM or CTM. Despite those drawbacks, 'gensim' has some advantages over the *R* packages: scalability (can be deployed in large environments and process huge chunks of data); and text parsing (can parse and work with different types and sources of textual data, such as wiki and XML). Also due to the python ecosystem, 'gensim' can interact natively with popular machine learning *python* libraries, for example, 'TensorFlow' and 'scikit-learn'. Thus, making 'gensim' very attractive towards computer scientists and social scientists looking to derive the fittest model to explain the data (mind the overfitting issue). Finally, the first tool that was available for topic modeling is the *Java*-based 'MALLET' (McCallum, 2002). It is still currently maintained and updated, but like 'gensim' can only perform LDA, not CTM or STM. It has an interface with both *R* and *Python* ecosystems by employing wrappers: 'gensim' library for *Python* and 'mallet' package for *R* (Mimno, 2013). All of the packages and libraries described here can be downloaded and used for free.

3. Case Study - RIAE last five years

This section will address my second objective: to illustrate how to do topic modeling by applying it in an analysis of the last five years of published research in this journal: the Iberoamerican Journal of Strategic Management.


All the data and procedures described in this section can be found in an online Open Science Framework public repository (Storopoli, 2019). I encourage the curious reader to access it and browse the code and files in order to capture any procedure that might sufficiently raise interest. Also, the online repository addresses replicability issues and other ethical concerns. This section is comprised of an extensive description of the procedures of data collection and data analysis. Ultimately, I present the results in a thorough manner.

3.1. Data Collection

The first procedure was to extract the last five years of published articles in the IJSM. This encompasses a timespan from 2014 to 2018. I chose full years in order to make the data collection and, subsequently, the analysis and the results replicable for future studies.

To generate the data, I have downloaded a BibTeX file exported from Scientific Periodicals Electronic Library - SPELL (www.spell.org.br) with all the published articles in IJSM. IJSM has published a total of 197 documents in SPELL from 2014 to 2018. To parse the content of the BibTeX file, I used an R package called 'bib2df' (Ottolinger, 2019). The variables that it was able to gather from SPELL for each of the 197 documents published are: title, authors, year, volume, number, pages, full abstract, URL of submission and DOI. Since my focus is on the published articles, I have removed 25 editorials and 8 book reviews from the sample; arriving in a final sample of 164 articles published in the last five years.

The 164 articles are the inputs documents to the topic modeling. However, first, it was necessary to pre-process the abstracts. The first pre-processing procedure was text normalization: (1) converting all letters to lower or upper case; (2) removing numbers; (3) removing punctuations, accent marks, and other diacritics; (4) and removing white spaces. Since IJSM adopted a structured abstract template during the later years of the analysis timeframe, I also employed a procedure to convert structured abstracts to regular abstracts. The second pre-processing procedure was to remove the stop words by using **spaCy** (Honnibal & Montani 2017), which is a free, open-source library for advanced Natural Language Processing (NLP) in *Python*. The classes of stop words that were removed with spaCy in this procedure are adverbs (very, tomorrow, down, where, there), pronouns (I, you, he, she, myself, themselves, somebody), conjunctions (and, or, but), determiners (a, an, the) and adpositions (in, to, during). The third and

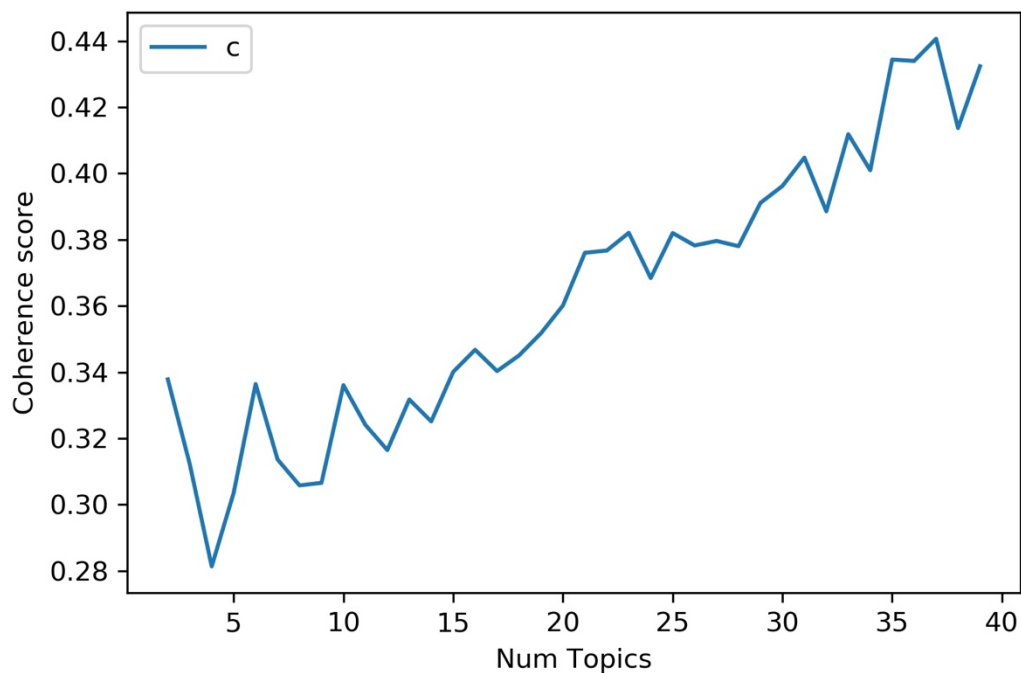
last pre-processing procedure was to lemmatize the text by reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form; the stemming was done following Porter's stemming algorithm (Porter, 1980). 

After the pre-processing procedures, I inspected the documents and also added 219 custom stemmed tokens as stop words and removed them from the text. The final pre-processed text for the documents has a total of 1,252 unique tokens.

3.2. Data Analysis

With the documents pre-processed as inputs, I generated a LDA model for each topic quantity, starting in 2 up to 40 topics. In figure 1, it is possible to see the coherence slowly creeping up as the number of topics increase. There is a tradeoff between coherence and topic numbers in a specific LDA topic Comodel. Particularly in this study, for the highest coherence value, I had two choices of topic models that have the same coherence value of 0.336. The first model has 6 topics and the second 10 topics. Since the coherence value is the same, I chose (following the Occam's razor principle) the model with the least number of topics.

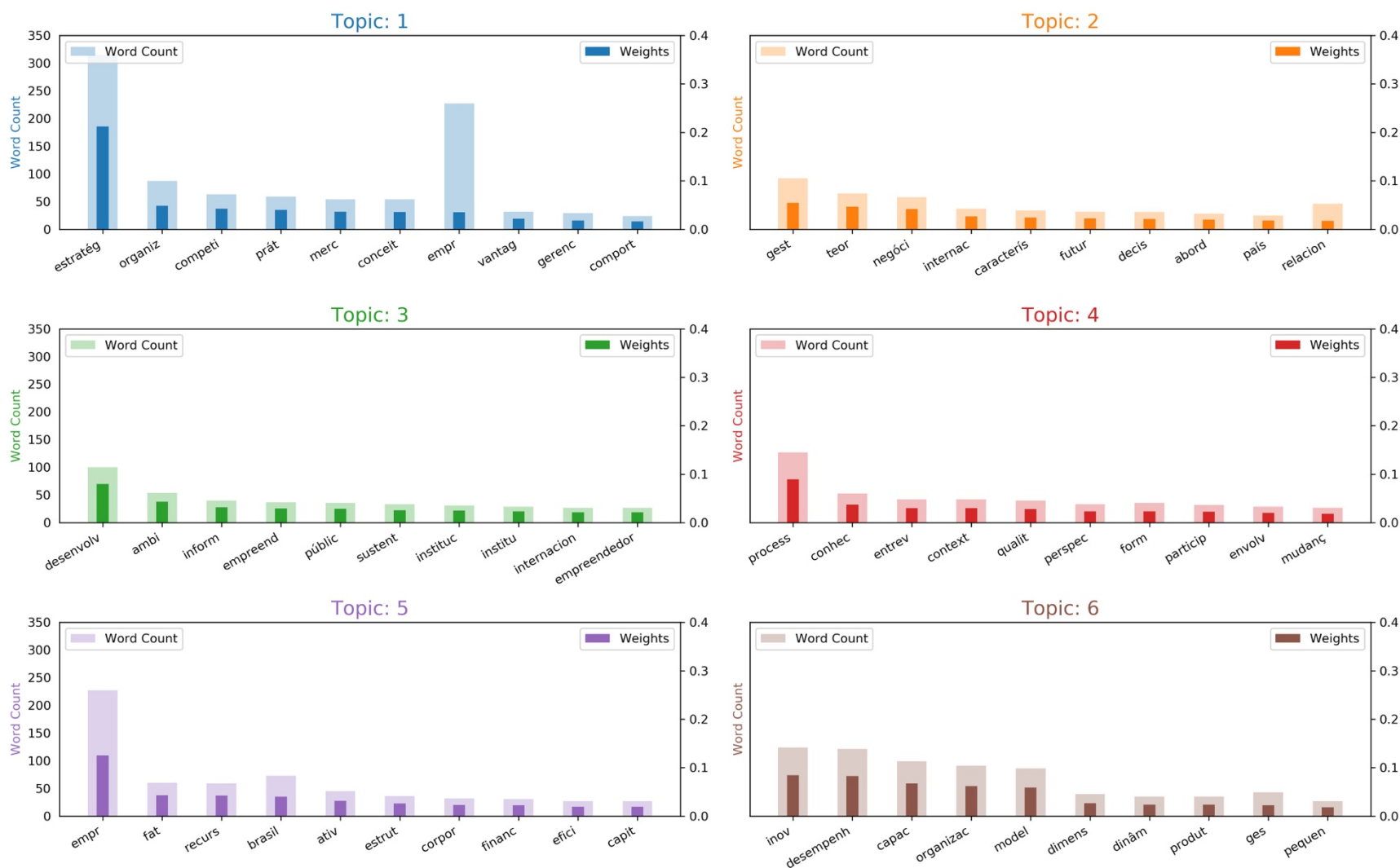
Figure 1. Number of topics



3.3. Results

The topic model has 6 topics. You can find each terms' weight for all the topics in Appendix I. Also, Figure 2 shows the topic's terms counts and weights for each one the of 6 topics. The first one is comprised of the terms: "estratég"; "organiz"; "competi"; "prát"; "merc"; "conceit"; "empr"; "vantag"; "gerenc"; and "comport". So topic 1 deals with the major strategy themes and competitive advantage. Thus, I name Topic 1 as "Strategy and Competitive Advantage". The most representative document in the sample for Topic 1 is titled "Comportamento Estratégico Organizacional e a Prática de Gerenciamento de Resultados nas Empresas Brasileiras".

Figure 2. Word count and importance of topic terms



Topic 2 has the following terms: "gest"; "teor"; "negóci"; "internac"; "caracterís"; "futur"; "decis"; "abord"; "país"; and "relacion". The most representative IJSM article in the sample has the following title: "Mulheres na gestão de topo: a problemática do GAP de gênero e salarial". Being most terms in Topic 2 related to International Business alongside with the most representative article dealing with gender inequality in the top management team, therefore Topic 2 is entitled "International Business and Top Management Team".

Topic 3 has the following terms: "desenvolv"; "ambi"; "inform"; "empreend"; "públic"; "sustent"; "instituc"; "institui"; "internacion"; and "empreendedor". The most representative document in the sample for this topic is "Do homo empreendedor ao empreendedor contemporâneo: evolução das características empreendedoras de 1848 a 2014". Being all terms and the most representative document regarding entrepreneurship, consequently, Topic 3 is named "Entrepreneurship".

Topic 4 has the following terms: "process"; "conhec"; "entrev"; "context"; "qualit"; "perspec"; "form"; "particip"; "envolv"; and "mudanç". The topic's most representative article in the sample is titled "Aprendizado de Rede no Contexto de Intercooperação e Fusão de Redes: A Opção de Não-Fusão". As noted by the topic terms', it is profoundly influenced by qualitative approach methods and techniques. The term "entrev" is the stemmed version of interview and the term "qualitativ" is the stem of qualitative. Due to the topic terms and the title of the topic's most representative document being learning and cooperation, Topic 4 is defined as "Learning and Cooperation".

Topic 5 has the following terms: "empr"; "fat"; "recurs"; "brasil"; "ativ"; "estrut"; "corpor"; "financ"; "efíci"; and "capit". The most representative document in the sample for the topic is "Determinantes da Estrutura de Capital de Empresas Brasileiras: Uma Análise Empírica das Teorias de Pecking Order e Trade-Off no Período de 2005 e 2014". This topic is mostly commanded by finance terms, thus it is named "Finance and Strategy".

The final topic, Topic 6, has the following terms: "inov"; "desempenh"; "capac"; "organizac"; "model"; "dimens"; "dinâm"; "produt"; "ges"; and "pequen". The most representative document for the topic is "Capacidades de Inovação em Serviços: Um Estudo nos Supermercados em Santa Catarina". The topic comprises the terms "capac" and "dinâm" which

means dynamic capability and also has the terms "pequen" meaning small firms. Therefore, Topic 6 is defined as "Dynamic Capabilities and Small Firms".

Now that the topics are present, we will delve into the relationship amongst the topics. By using the *Python* port of the *R* package 'LDAvis' (Sievert & Shirley, 2014), I was able to generate a multidimensional scaling (MDS) (Torgerson, 1958). In figure 3, there are two main axes of the MDS: the x-axis and y-axis. In this Cartesian space, all distances are assumed as Euclidean. We can see that there are clearly 3 groups. Topic 2 and 4 are grouped together. Furthermore, Topics 1, 3 and 6 are also grouped together opposite to Topic 2 and 4. The final group is composed of a single topic, Topic 5, and is the opposite position of both remaining groups. The MDS elucidates that "Topic 4 - Learning and Cooperation" is very similar to "Topic 2 - International Business and Top Management Team"; and that "Topic 1 - Strategy and Competitive Advantage", along with "Topic 3 - Entrepreneurship" and "Topic 6 - Dynamic Capabilities and Small Firms" share similarities. These two groups are opposed to each other in the main axis (the PC1 on the x-axis). Isolated without similarities to any topic is "Topic 5 - Finance and Strategy".

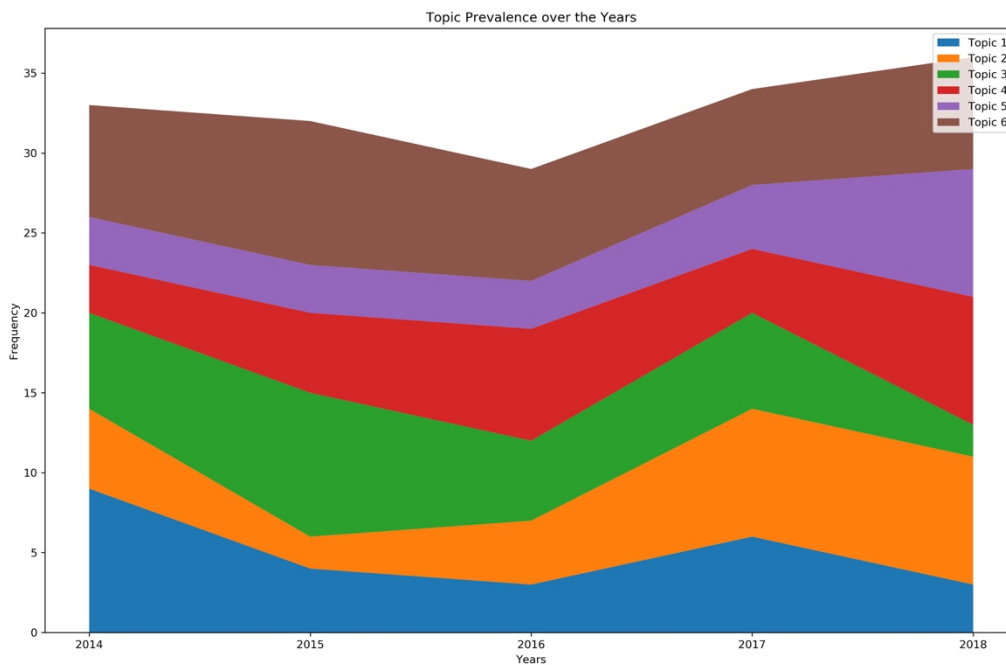
Figure 3. Intertopic Multidimensional Scaling



Finally, I present the topic predominance over the years in Figure 4. It has some trends to note. First, Topic 1 - Strategy and Competitive Advantage and Topic 3 - Entrepreneurship were

predominant in 2014 and declined towards 2018. Second, Topic 2 - International Business and Top Management Team and Topic 5 - Finance and Strategy are on the rise from 2016 onwards. Third, Topic 4 - Learning and Cooperation and Topic 6 - Dynamic Capabilities remained steady during the timeframe of analysis. Regarding the decline of Topic 1--which is a general topic, this can be interpreted as a specialization of the journal. As the years go by, IJSM started to publish more specialized content that was grouped into other topics. The rise of Topic 2 and Topic 5 can be seen as a trend towards more publications about finance, international business, and top management team.

Figure 4. Topic predominance per year



4. Conclusions

My objectives in this article were two. First, I introduce topic modeling as a social sciences research tool and map critical published studies in management and other social sciences that employed topic modeling in a proper manner. Second, I illustrate how to do topic modeling by applying topic modeling in an analysis of the last five years of published research in IJSM. Topic Modeling is a valuable toolkit for management researchers to use in their theory-building



process. It can shift the old paradigm that textual data belongs only to the qualitative realm by allowing textual data to be labeled and quantified in a reproducible manner that mitigates (or closely fully eliminates) researcher bias.

For further research, while addressing the limitations, I propose the use of CTM and STM instead of LDA to analyze the data and draw more insights from the results. CTM allows the topics to correlate with each other, and can give a better fit but harder to interpret topics (Steyvers & Griffiths, 2007). STM can benefit from the metadata that textual data carry alongside, bringing relationships between metadata and topics to the analysis (Robert et al., 2014). Also, mind that the median of the sample size from the 23 Topic Modeling articles in Table 1 is 8,000 documents. This could imply that Topic Modeling may benefit from a large sample.

References

- Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6), 1397–1410. <https://doi.org/10.1002/asi.23786>
- Bendle, N. T., & Wang, X. (2016). Uncovering the message from the mess of big data. *Business Horizons*, 59(1), 115–124. <https://doi.org/10.1016/j.bushor.2015.10.001>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(1), 993–1022.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1), 17–35. <https://doi.org/10.1214/07-aos136>
- Bordag, S. (2008). A comparison of co-occurrence and similarity measures as simulations of context. *Proceedings of the 9th international conference on computational linguistics and intelligent text processing*, 52–63. https://doi.org/10.1007/978-3-540-78135-6_5
- Chang, J. (2011). lda: Collapsed Gibbs sampling methods for topic models. R.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, 39(1), 110–135. <https://doi.org/10.17705/1CAIS.03907>

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, 26(2), 168–189. <https://doi.org/10.1017/pan.2017.44>

DiMaggio, P., Nag, M., & Blei, D. M. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606. <https://doi.org/10.1016/j.poetic.2013.08.004>

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2), 205395171560290. <https://doi.org/10.1177/2053951715602908>

Glaser, B., & Strauss, A. (1967). Grounded theory: The discovery of grounded theory. *Sociology the journal of the British Sociological Association*, 12(1), 27-49.

Hannigan, T., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V., Wang, M. & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*. <https://doi.org/10.5465/annals.2017.0099>

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hornik, K., & Grün, B. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software.*, 40(13), 1–30.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 23(2), 254–277. <https://doi.org/10.1093/pan/mpu019>

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A. & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and

Reliable Methodology. *Communication Methods and Measures*, 12(2–3), 93–118.
<https://doi.org/10.1080/19312458.2018.1430754>

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from <http://mallet.cs.umass.edu/index.php>

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

Mimno, D. (2013). *mallet*: A wrapper around the Java machine learning tool MALLET. Retrieved from <https://cran.r-project.org/package=mallet>

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, 41(6), 545–569. <https://doi.org/10.1016/j.poetic.2013.10.001>

Nelson, L. K. (2017). Computational Grounded Theory. *Sociological Methods & Research*, 1-40. <https://doi.org/10.1177/0049124117729703>

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2018). The Future of Coding. *Sociological Methods & Research*, 1-36. <https://doi.org/10.1177/0049124118769114>

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, 43(1), 88–102. <https://doi.org/10.1177/0165551515617393>

Ottolinger, P. (2019). *bib2df*: Parse a BibTeX File to a Data Frame. Retrieved from <https://cran.r-project.org/package=bib2df>

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137. <https://doi.org/10.1108/eb046814>

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

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K. & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4), 1064–1082. <https://doi.org/10.1111/ajps.12103>

Roberts, M. E., Stewart, B. M., & Tingley, D. (2014). stm: R package for structural topic models. *Journal of Statistical Software*, 10(2), 1–40.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Storopoli, J. (2019, July 22). Topic Modeling IJSM-RIAE. Retrieved from osf.io/97w6z

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: J. Wiley.

APPENDIX I - Topic terms weights

First Topic - 0.212*"estratég" + 0.049*"organiz" + 0.042*"competi" + 0.040*"prát" + 0.036*"merc" + 0.036*"conceit" + 0.035*"empr" + 0.022*"vantag" + 0.018*"gerenc" + 0.016*"comport"

Second Topic - 0.055*"gest" + 0.047*"teor" + 0.042*"negóci" + 0.027*"internac" + 0.024*"caracterís" + 0.022*"futur" + 0.021*"decis" + 0.020*"abord" + 0.018*"país" + 0.017*"relacion"

Third Topic - 0.080*"desenvolv" + 0.043*"ambi" + 0.032*"inform" + 0.030*"empreend" + 0.029*"públic" + 0.026*"sustent" + 0.025*"instituc" + 0.023*"institu" + 0.022*"internacion" + 0.022*"empreendedor"

Fourth Topic - $0.090 \cdot \text{"process"} + 0.037 \cdot \text{"conhec"} + 0.030 \cdot \text{"entrev"} + 0.030 \cdot \text{"context"} + 0.028 \cdot \text{"qualit"} + 0.023 \cdot \text{"perspec"} + 0.023 \cdot \text{"form"} + 0.023 \cdot \text{"particip"} + 0.020 \cdot \text{"envolv"} + 0.018 \cdot \text{"mudan\c{a}}"$

Fifth Topic - $0.125 \cdot \text{"empr"} + 0.043 \cdot \text{"fat"} + 0.042 \cdot \text{"recurs"} + 0.040 \cdot \text{"brasil"} + 0.032 \cdot \text{"ativ"} + 0.026 \cdot \text{"estrut"} + 0.023 \cdot \text{"corpor"} + 0.022 \cdot \text{"financ"} + 0.019 \cdot \text{"efici"} + 0.019 \cdot \text{"capit"}"$

Sixth Topic - $0.084 \cdot \text{"inov"} + 0.082 \cdot \text{"desempenh"} + 0.067 \cdot \text{"capac"} + 0.062 \cdot \text{"organizac"} + 0.059 \cdot \text{"model"} + 0.027 \cdot \text{"dimens"} + 0.024 \cdot \text{"din\^am"} + 0.024 \cdot \text{"produt"} + 0.022 \cdot \text{"ges"} + 0.018 \cdot \text{"pequen"}"$