How to Identify Hot Topics in Psychology Using Topic Modeling

André Bittermann[1] and Andreas Fischer[2]

[1]Leibniz Institute for Psychology Information (ZPID), Trier, Germany

[2]Forschungsinstitut Betriebliche Bildung (f-bb), Nuremberg, Germany

Correspondence address:

André Bittermann

Leibniz Institute for Psychology Information, (ZPID) Trier, Germany

Universitätsring 15

54296 Trier

Germany

abi@leibniz-psychology.org

Abstract

Latent topics and trends in psychological publications were examined to identify hotspots in psychology. Topic modeling was contrasted with a classification-based scientometric approach in order to demonstrate the benefits of the former. Specifically, the psychological publication output in the German-speaking countries containing German- and English-language publications from 1980 to 2016 documented in the PSYNDEX database was analyzed. Topic modeling based on latent Dirichlet allocation was applied to a corpus of 314,573 publications. Input for topic modeling was the controlled terms of the publications, that is, a standardized vocabulary of keywords in psychology. Based on these controlled terms, 500 topics were determined and trending topics were identified. Hot topics, indicated by the highest increasing trends in this data, were fascets of neuropsychology, online therapy, cross-cultural aspects, traumatization, and visual attention. In conclusion, the findings indicate that topics can reveal more detailed insights into research trends than standardized classifications. Possible applications of this method, limitations, and implications for research synthesis are discussed.

Keywords: topic modeling, hotspots, scientometrics, trends, controlled terms

How to Identify Hot Topics in Psychology Using Topic Modeling

Topics of particular significance in research-active fields have been referred to as "hotspots" (Erdfelder & Bošnjak, 2016). From a scientometric point of view, the occurrence of hotspots may reflect areas of current scientific discourse. On the other hand, hotspots may also derive from current needs of society, for example, consider the impact of topics such as digitalization, terrorism, or the German "refugee crisis" (beginning in 2015) on psychological research. Thus, addressing hotspots might help to deliver research results that are interesting to both the scientific community and/or the general public if the research is imparted comprehensibly (Friedman, 2008). Nevertheless, it is an open question how to identify the set of potentially hot topics in a domain of interest. In this paper, we will contrast two ways of identifying topics based on a corpus of scientific publications (manifest classifications vs. latent topics).

A comparatively simple and straightforward approach for identifying research topics is based on existing classification systems, such as the "Classification Codes"[1] outlined in the *Thesaurus of Psychological Index Terms* (Tuleya, 2007) published by the American Psychological Association (APA). Currently, this thesaurus provides 157 categories to describe the content included in the publication database, and each category may be considered a research topic. However, with regard to identifying hotspots, the apparent simplicity of this approach is burdened with multiple drawbacks: First, the approach is based on an established classification system, and thus some of the most recent (and hot) topics may not be represented in the analysis until the classification system is expanded accordingly; second, classifications may be too broad and abstract to capture the topics that are of particular significance in research-active fields (e.g., there is no classification code specific to "evaluation," even if some researchers may consider treatment evaluation an interesting topic); a third problem arises due to the fact that some publications address more than one

topic. Consider a study that examines the neuropsychological correlates of emotional lability in traumatized refugees. If only one classification is assigned (e.g., “Neuropsychology & Neurology”) the information on disorders and migration-related aspects remains hidden; instead, when using additional classifications to categorize these contents, the respective proportions remain unspecified (i.e., there may be equal or varying shares of each content).

A more complex approach to identifying topics is to derive latent topics from the manifest content addressed within a corpus of publications through methods such as topic modeling (e.g., Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004). The basic idea behind topic modeling is that every document can address different topics that are not known a priori. Thus, the goal is to identify these latent topics based on the documents’ manifest contents by employing algorithms that “analyze the words of the original texts to discover the themes that run through them” (Blei, 2012, p. 77). Since information on the level of the full text, abstract, or keywords can be used for topic modeling, the resulting topics have the potential to address specific subjects based on the corpus and independent from predefined classifications.

In topic modeling, each document is assumed to address each topic to varying degrees (0-100%). For example, a paper might comprise an evaluation topic with a share of 10% and other topics with a share of 90%. This means that in contrast to a dichotomous classification (this publication is assigned or is not assigned to the classification) or multiple dichotomous classifications, a probabilistic approach such as topic modeling can deal with heterogeneous topics of a publication in terms of topic proportions. In this study, such a probabilistic method is applied for topic modeling, namely latent Dirichlet allocation (LDA; Blei et al., 2003).

By applying statistical methods to the change of mean topic probabilities over time, rising and declining trends can be identified (Griffiths & Steyvers, 2004). Once trending topics are identified, scientific knowledge can be gathered from publications addressing these topics by conducting systematic reviews and meta-analyses to synthesize the results from related published research on a certain subject. The current study aims to deliver the

foundation for such research synthesis techniques in the context of hotspots in psychology: a data-driven bottom-up approach for the identification of latent topics and trends in psychological research.

Topic Modeling in Psychological Research and Scientometrics

Big data and topic modeling represent a relatively new approach of psychological research methods that can be applied to various research questions (e.g., Chen & Wojcik, 2016; Kosinski, Wang, Lakkaraju, & Leskovec, 2016). For example, Griffiths, Steyvers, and Tenenbaum (2007) used topic models for predicting word association and the effects of semantic association and ambiguity on a variety of language-processing and memory tasks. Steyvers and Griffiths (2008) showed that both human memory and information retrieval faces similar computational demands by employing topic models. Topic models have also been used for modeling couple and family text data (Atkins et al., 2012), improving the prediction of neuroticism and depression (Resnik, Garron, & Resnik, 2013), investigating mental health signals in Twitter (Coppersmith, Dredze, & Harman, 2014), analyzing the linguistic data of patient-therapist interactions (Imel, Steyvers & Atkins, 2015), and exploring differences in language use on Facebook across gender, affiliation, and assertiveness (Park et al., 2016).

In the field of scientometric analysis, which is highly relevant for the present study, Griffiths and Steyvers (2004) applied LDA topic models to a corpus of abstracts published in the *Proceedings of the National Academy of Sciences of the United States of America* and identified “hot” and “cold” topics. A topic was defined as hot if it showed an increasing linear trend in popularity and cold if it showed a decreasing linear trend in popularity. This approach was adapted in several other research fields, for example, to identify the major biological concepts from a corpus of protein-related publication titles and abstracts (Zheng, McLean, & Lu, 2006), to conduct a bibliometric analysis of aquaculture literature (Natale, Fiore, & Hofherr, 2012), to analyze the field of development studies (Thelwall & Thelwall, 2016), or

to explore hydropower research (Jiang, Qiang & Lin, 2016). The current study is the first to apply LDA-based topic modeling for a scientometric analysis of psychological research in the German-speaking countries.

A Brief Illustration of LDA-based Topic Modeling

In the following, a very brief and illustrative description of LDA and topic modeling is provided. Further details and more technical descriptions can be found in Blei (2012) and Blei et al. (2003). The underlying assumption of LDA is that a document represents a mixture of topics with different proportions (Blei et al., 2003). Using Bayesian probabilistic modeling, LDA aims to identify clusters of terms (i.e., topics) that tend to co-occur within documents (Park et al., 2016). Thus, topics are defined as a distribution over a fixed vocabulary (Blei, 2012). In a generative process, two kinds of probabilities are drawn from Dirichlet distributions over (1) the prior weight of a certain word in a topic (β) for the probabilities of terms occurring in a certain topic (φ), and (2) the prior weight of a certain topic in a document (α) for the probabilities of topics occurring in a certain document (θ) based on the terms within the document. “Prior” means that the α and β hyperparameters have to be set prior to the analysis. Lower values of α result in documents belonging to fewer topics, and lower values of β result in more separated topics.

For a simplified illustration of the main idea behind topic modeling in a scientometric context, imagine a corpus consisting of four documents and a model of four topics. For the sake of brevity, each document shall consist of 16 terms (see Table 1). In this idealized example, LDA reveals four topics by clustering co-occurring terms, of which the five most frequent terms are shown in Table 2. Note that each topic actually consists of all unique terms of the corpus, that is, of all four documents. The terms are sorted by frequency to best represent different topics. For the sake of illustration, the results presented in Table 2 can be considered ideal because these topics reflect optimal semantic differences. As a real LDA analysis for this very small sample corpus is based exclusively on term co-occurrences, one of

the resulting topics would be “intervention, parents, disgust, love, hate,” which includes different semantic meanings. Documents differ in topic proportions, and this is represented by the probability of a document belonging to a topic (θ). As shown in Table 3, Document 1 addresses all four topics with equal shares, whereas Document 2 mostly addresses Topic 2 and so on. The resulting mean document-topic probabilities by topic show that Topic 1 has a mean probability of 25% which corresponds to the expect proportion $1/k$ (with k being the number of topics). Topic 2, with a mean probability of 37.5%, can be considered as the most popular, whereas Topics 3 and 4 are less popular than average.

LDA is an unsupervised method, but the number of topics (k) must be defined a priori by the analyst. Griffiths and Steyvers (2004) examined different values of k and compared the resulting log-likelihoods. Yet another approach would be to test various values of k and determine the optimal k intellectually, that is, by expert judgment to decide whether the topics are in balance between too broad or too specific (Thelwall & Thelwall, 2016). In this study, we follow the first approach.

Using Controlled Terms for Topic Modeling

For a reliable identification of the representative topics within a research field, the information about the documents’ content must be of high quality. For instance, if abstracts give a mere introduction rather than an objective summary, the resulting topics will reflect the theoretical background or the studies’ raison d'être rather than their actual content. The same applies to keywords that, for instance, contain the statistical methods that were used and do not represent the actual topic of the study (e.g., in the case of “analysis of variance,” from a keyword point of view it remains unclear whether the method was simply applied, discussed, or further developed). To avoid latent semantic heterogeneity within a topic, all keywords should be chosen according to the same rules (e.g., keywords for statistical procedures are assigned only if they themselves are the focus of the study and not their mere application is referred to). In most studies, authors provide keywords that further summarize the document’s

content. If these keywords are uncontrolled (i.e., can be freely chosen), (1) it is not guaranteed that they actually represent the main concepts, ideas, and topics of the publication; (2) they sometimes are long phrases and not terms; and (3) keywords from different authors might be different terms for the same idea (e.g., adaptation vs. adaption vs. adjustment) or, conversely, the same words for different ideas. These are problematic aspects for topic modeling using LDA, since topics are identified according to word co-occurrences (Blei et al., 2003). Topic models aim to capture semantically related topics (Wallach, Mimno, & McCallum, 2009), but they do not generate topics based on the words' inherent semantic relations.

The PSYNDEX database is developed and hosted by the Leibniz Institute for Psychology Information (ZPID; Trier, Germany) and is a comprehensive database containing German- and English-language publications in psychology and closely related disciplines from the German-speaking countries. In early July 2017, there were more than 327,400 documents indexed in PSYNDEX (accessible at www.PubPsych.eu). The PSYNDEX editorial staff assigns controlled terms (CTs) from the aforementioned *Thesaurus of Psychological Index Terms* published by the APA (Tuleya, 2007; ZPID, 2016). In the context of topic modeling, this controlled vocabulary has several advantages: (1) The CTs correspond with the content of the publications. (2) The terms' standardized spelling avoids synonyms or variations in expressions. (3) The corpus for topic modeling consists of only those words that are relevant to the content. Stop words that contain little topical content (e.g., "the", "a", "and") have to be neither defined nor deleted. (4) All CTs are available in German and English; therefore, the whole corpus of publications can be used irrespective of the documents' language. (5) In contrast to abstract texts, the terms do not have to be stemmed with the resulting problem of word fragments. (6) Since the corpus contains fewer words, computation time decreases and fewer memory resources are needed. In a pretest with 3,846 documents, LDA based on CTs took less than 7% of the time needed for an abstract-based LDA while revealing comparable results. Thus, in contrast to prior research using abstracts as

primary data for topic modeling analysis (e.g., Griffiths & Steyvers, 2004; Jiang et al., 2016), the current study employs CTs for topic modeling.

Objectives

The objectives of the current study are twofold:

(1) to examine trends of latent topics, and

(2) to contrast latent topics with manifest classifications.

LDA-based topic modeling will be applied to a corpus of psychological publications from the German-speaking countries retrieved from PSYNDEX. Increasing and decreasing linear trends as well as nonlinear trends will be identified. Furthermore, the topics will be contrasted with classifications in terms of thematic specificity.

Method

Data

Data were extracted from the PSYNDEX database on July 3, 2017. A total of 316,996 of the indexed psychological articles, book chapters, reports, and dissertations were published between 1980 and 2016. Biographies or historical sources (reprints or selected readings) were excluded, since they usually address the topic retrospectively, resulting in $N = 314{,}573$ publications.

Software

Analyses were conducted in RStudio version 1.0.153 (RStudio Team, 2016) based on R version 3.4.2 (R Core Team, 2017). For text mining and topic modeling, the packages *tm* 0.7-1 (Feinerer, Hornik, & Meyer, 2008) and *topicmodels* 0.2-6 (Grün & Hornik, 2011) were used. Additional operations were conducted with packages *dplyr* 0.5.0, *readr* 1.1.0, *splitstackshape* 1.4.2, *Xmisc* 0.2.1, *lattice* 0.20-35, and *nnet* 7.3-12.

Topic Modeling

LDA was applied using Gibbs sampling with parameters as suggested by Awati (2015), that is, 4,000 omitted Gibbs iterations at beginning, 2,000 Gibbs iterations, 500

omitted in-between Gibbs iterations, and five repeated random starts. Parameters of the symmetric Dirichlet priors were set according to Tang, Meng, Nguyen, Mei, and Zhang (2014), that is, $\alpha = 0.1$ (resulting in documents belonging to fewer topics) and $\beta = 0.01$ (resulting in well-separated topics). Concerning the number of topics *k*, we inspected the log-likelihood estimates for various values of *k*, which is referred to as the commonly used approach (Kosinski et al., 2016). We ran models with 100, 150, 200, 300, 400, and 500 topics comparable to Griffiths & Steyvers (2004), who tested values of 50, 100, 200, 300, 400, 500, 600, and 1,000 topics. Values of *k* higher than 500 were discarded, since more topics decrease understanding and verifiability by experts (De Battisti, Ferrara, & Salini, 2015). Text input for the topic models were the publications' controlled keyword terms (CTs). They were prepared for LDA by removing spaces, parentheses, hyphens, slashes, and apostrophes.

Modeling Trends

Previous research employed linear regression models for identifying increasing and decreasing trends (Griffiths & Steyvers, 2004; Paul & Girju, 2009; Ponweiser, Grün, & Hornik, 2014). Hot topics were defined by the highest linear slopes. We extended this approach by taking nonlinearity into account to identify nonlinear trends. Specifically, we applied multilayer perceptrons (MLPs) with two hidden-units to model the average topic probability (mean of document-topic probabilities over all documents for each topic) as a nonlinear function of the year of publication. The MLPs applied provide nonlinear regression functions with a minimal sum of squared residuals for each topic, and thus provide an estimate of R^2, given an optimal nonlinear transformation of the year of publication (Fischer, 2015). Two hidden units were included to allow for nonmonotonic functions while at the same time minimizing the risk (and amount) of overfitting (Fischer, 2015). The difference between R^2_{MLP} and R^2_{linear} is applied as an indicator of the amount of nonlinearity that is not accounted for by the linear model. More specifically, nonlinearity is defined by $R^2_{\text{MLP}} > 2 \cdot R^2_{\text{linear}}$. Trends were estimated over a period of more than two years. Because of random

fluctuations—and considering the duration of a typical publication cycle—an estimation over a shorter time span implies severe overfitting (Fischer, 2015) and may not represent a topic's significance well. The complete R code used in the analyses is provided in the Electronic Supplementary Material (ESM 1).

Results

Model Selection

The corpus of $N = 314{,}573$ documents contained 6,073 unique terms. By comparing log-likelihoods of the resulting models (as shown in Table 4), $k = 500$ was determined as the optimal number of topics. A table containing the top 15 terms of all topics can be found in the Electronic Supplementary Material (ESM 2).

Trends in Topics

Linear trends in changes of mean document-topic probabilities (θ) over time were analyzed according to Griffiths and Steyvers (2004) with an additional examination of nonlinearity. Significantly increasing linear trends could be found for 128 of the topics, and significantly decreasing linear trends could be found for 135 of the topics, both at the $p = .0001$ level. The 10 topics with highest increasing linear trends (i.e., hot topics) are listed in Table 5. Figure 1 shows their mean document-topic probabilities (θ) by publication year. The major hot topics are neuropsychology and genetics, online therapy, human migration, traumatization, and visual attention. A closer look at the terms of these topics (Table 5) reveals that these major themes can be further specified. Traumatization, for example, can be further specified with three narrower topics: traumatization of refugees during war and torture (Topic 86), therapy of emotional trauma (Topic 344), and trauma-related disorders and processes (Topic 95).

Since the focus of the current study was on hot topics, additional trend analyses are reported briefly (see ESM 2 for topic terms and more information on the following topics). Strongly decreasing linear trends (i.e., cold topics) could be found in topics referring to

human-factors engineering (Topic 310), psychosomatic disorders (Topic 361), incarceration (Topic 472), social and political processes in West and East Germany (Topics 41, 393, and 186), experimental methodology (Topic 342), group psychotherapy (Topic 491), community mental health services (Topic 163), and infectious disorders (Topic 479).

The comparison of R^2_{linear} and R^2_{MLP} revealed topics with a considerable amount of nonlinearity that is not accounted for by the linear model. The largest difference between R^2_{linear} and R^2_{MLP} (i.e., nonlinear trends) could be found for topics referring to psychodiagnosis and testing (Topic 467, with peaks in 2006 and 2011), outpatient psychotherapy (Topic 334, with peak in 1998), family relations (Topic 259), prevention and health promotion (Topic 162, highest peak in 1991), Internet and information systems (Topic 481, with peak in 1999), organizational psychology (Topic 345), sexual relations with clients in psychotherapy (Topic 237, peaks in 1995 and 1998), racial and ethnic attitudes (Topic 130, peak in 1993), health behavior and dental health (Topic 44), and relations between socioeconomic background of the family and education (Topic 138).

Relationship Between Topics and Classifications

According to the second objective of the current study, we investigated whether topics can be allocated to a specific PSYNDEX subject classification. If this is the case, topics either match the classifications' content or they provide more detailed information within the classification. If this is not the case, topics cover themes that could only be matched by multiple classifications. For every document, the assigned classifications were compared to the documents' most probable topics in order to examine content similarities and differences. Similar to Griffiths and Steyvers' (2004) approach for identifying diagnostic topics, Figure 2 shows a level plot of mean document-topic probabilities (θ) by topics and main classifications for the hot topics. For creating the level plot, publications were grouped by classification (in case of multiple classifications, the document was assigned to each classification). Then, mean θ probabilities were determined by each classification. This allowed the investigation of

the extent to which a topic's semantic content (as reflected by its top terms) corresponds with the classification system. For the sake of clarity, only main classification categories are included in Figure 2 (as the complete APA classification system consists of 157 codes).

The darker the cells of the level plot, the higher the mean θ. If a topic column shows different colors, the θ values are not equally distributed over the classifications, that is, the topics' semantic content cannot be reflected by a single classification. Clearly, the topics do not match the classifications perfectly, but they do show correspondence with various classifications (in this case, only one dark cell is observed for each topic). For example, the highest mean θ for Topic 371 (referring to human migration and cross-cultural aspects) can be observed in "2900 Social Processes & Social Issues." Since it also shows a relatively high mean θ in various other classifications, this topic cannot be described by a single classification. The hot topics concerning neuropsychology (Topics 364, 249, and 323) and genetics (Topic 459) show their highest mean θ in "2500 Physiological Psychology & Neuroscience," but also in other classifications. No distinctively matching classification can be identified for Topics 86 (traumatization of refugees) and 95 (traumatization-related disorders).

Selecting Publications for Research Synthesis

The publications related to a topic can be filtered by (1) using the document-topic probabilities (θ) or (2) using the keywords that constitute the topic for literature search. We employed the first approach on the example of Hot Topic 386 (online therapy) and sorted documents by θ in decreasing order. This resulted in a list of all publications in the corpus, with the ones most likely addressing the topic ranking highest. The results were then filtered by selecting only empirical studies with values of θ higher than $1/k$ (i.e., the average document-topic probability). This means that Topic 386 occurs in these empirical studies with a probability above average. The distribution of θ values is shown in Figure 3. Inclusion criteria for subsequent research synthesis approaches can be applied to this subset of 1,083

documents. Since the documents are ranked by θ, a list can be generated that allows for an inspection of relevant documents in the order of their topic probabilities.

Discussion

The current study applied LDA-based topic modeling for a scientometric analysis of psychological research as a data-driven bottom-up approach for the identification of latent topics and trends. In a model with 500 topics, strongly increasing linear trends were found for topics addressing neuropsychology and genetics, online therapy, human migration and cross-cultural aspects, traumatization, and visual attention. These topics were referred to as hot topics in psychology. Additionally, it was shown how the resulting topics can be used for purposes of research synthesis.

The topics' contents corresponded with respective classifications, but as expected, they could not be matched to a single subject classification. Thus, the topics provided information beyond the scope of a predefined classification system. Prior scientometric research in psychology used classifications for determining trends (e.g., Krampen & Trierweiler, 2013; Krampen, 2016). From our results, it can be concluded that this approach is feasible as long as the classifications' specificity is satisfactory. Using topic modeling, we were able to find specific topics that would have not been easy to detect by a classification-based approach, for example, lifestyle of adolescents and popular culture (Topic 49), attitude change of public opinion (Topic 72), values in individualism versus collectivism (Topic 144), or traumatization of refugees because of war and torture (Topic 86). Most topics represent a mixture of classifications.

Methodological Limitations

Topic models were based on standardized keywords (controlled terms, CTs) of the publications. This approach resulted in a much smaller number of corpus terms than would have resulted from using abstracts. CTs reflect a document's content in a condensed manner and offer several advantages in the context of topic modeling: Computation times are shorter,

there is no need for stop words, they offer excellent readability, and since the topics consists of CTs, they can be used directly for subsequent literature searches.

A significant disadvantage of using CTs is the time of their first thesaurus inclusion as a potential artefact. For instance, recently added CTs such as “Political Asylum” or “Asylum Seeking” (both included in 2015) cannot describe a topic during the years before their addition. Nevertheless, if one is interested in recent topics, the following approach for defining hotspots besides considering trends over time could be employed: By building a corpus for the respective recent years (e.g., 2015-2016), popular topics could be defined by examining the highest mean document-topic probabilities. This represents a cross-sectional approach using all currently available CTs.

Similar to classifications, thesaurus-based CTs have limitations regarding their semantic detail. The uncontrolled keywords of the current study are “topic modeling, hotspots, scientometrics, trends, controlled terms,” with more or less corresponding CTs “Mathematical Modeling, Scientific Communication, Trends” (no matches for the quite specific keywords “hotspots” and “controlled terms”). The use of words in the abstract would overcome these shortcomings, since every word of the original text can be included. Downsides, on the other hand, are the problem of defining stop words (e.g., Schofield, Magnusson, & Mimno, 2017) and a much larger corpus vocabulary with higher computational demands that would require several days of calculation time or the use of a computer cluster.

The number of topics *k* was determined by computing models for values of *k* and inspecting the respective log-likelihoods, which is referred to as the commonly used approach (Kosinski et al., 2016). The log-likelihoods increased with higher values of *k*, indicating that a model with more topics could show an even better fit. However, a model with more topics is more difficult to be understood and verified by experts (De Battisti et al., 2015). Besides, inspecting the hot topics for the applied values of *k* in this study revealed stable themes of neuropsychology, online therapy, human migration, traumatization, and visual attention.

In this study, basic LDA was employed using the R programming language. Newer developments such as dynamic topic modeling with a focus on changes over time (Blei & Lafferty, 2006) or correlated topic models that aim to capture correlations between the occurrence of latent topics (Blei & Lafferty, 2007) could further improve the identification of hot topics in psychology. The analysis of abstracts from different languages by employing polylingual topic models (Mimno et al., 2009) or multilingual probabilistic topic modeling (Vulić, De Smet, Tang, & Moens, 2015) could be of interest for future research as well.

Implications for Research Synthesis

The topic modeling approach presented in this paper can be applied to the identification of hotspots in psychology. Erdfelder and Bošnjak (2016) related hotspots to the presence of a significant number of primary studies within a research-active field. We expanded the scope of hotspots by including all types of publications with the exception of historical studies and biographies in order to gain a comprehensive view on the topics that are addressed. For subsequent research synthesis purposes, primary studies which address hot topics can be easily identified in PSYNDEX by filtering the documents. This results in a list with the documents that show the highest document-topic probabilities at the highest ranks. A more common approach for selecting documents would be using the keywords (CTs) that constitute the topic.

In this paper, only the top 15 terms of each hot topic were reported. Since a topic consists of a long list of terms, with various frequencies and term-to-topic probabilities, we encourage readers to take a closer look at the topics of interest. A sophisticated method for visualizing and interpreting topics is provided by “LDAvis” (Sievert & Shirley, 2014), which defines the relevance for ranking terms within topics based on weight parameters and can be employed in R with the “LDAvis” package (Sievert & Shirley, 2015).

Other Possible Applications

A researcher who wants to develop a new area of interest can learn more about the subject's structure by looking at the underlying topics. A publication database could be explored more in depth using topics, as illustrated by topic-based web browsers of Wikipedia (Chaney & Blei, 2012) or the Signs journal (Goldstone, Galán, Lovin, Mazzaschi, & Whitmore, 2014). Moreover, for a better navigation through a model with many topics, scientific documents could be divided into several topic clusters (Yau, Porter, Newman, & Suominen, 2014). Such clusters could constitute empirically derived themes as an alternative to manifest classifications.

Since a document-topic probability is computed for every publication, those papers could be recommended that show the highest probabilities for the respective topic. A more sophisticated approach was presented by Wang and Blei (2011), who developed an algorithm for recommending scientific articles to users of an online community.

Authors usually indicate their fields of interest, for example, "psychotherapy research" or "cognitive processes." Here, topic modeling can be used to identify research topics based on the authors' publications (Lu & Wolfram, 2012; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004). This procedure results in a publication-based profile of authors which can be applied to find experts for specific topics, find authors with similar topics, or analyze authors' change of publication-based interests over time.

Conclusion

Topic modeling is a feasible method for an exploratory analysis of topics in psychological publications and for identifying hot research topics. The identification of specific topics in a large corpus of publications offers new possibilities of exploring research beyond predefined classifications. Furthermore, topics can be the starting point for subsequently applied research synthesis methods.

Acknowledgments

We thank Lisa Trierweiler and Katja Singleton for helpful comments and recommendations during the writing process, Jürgen Wiesenhütter and Veronika Kuhberg-Lasson for valuable input during early phases of this research, and Andreas Konz and Jannik Lorenz for hardware support.

References

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, *26*(5), 816-827. doi:10.1037/a0029607

Awati, K. (2015, September 29). A gentle introduction to topic modeling using R [Blog post]. Retrieved from https://eight2late.wordpress.com/2015/09/29/a-gentle-introduction-to-topic-modeling-using-r/

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. doi:10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113-120). New York, NY: ACM. doi:10.1145/1143844.1143859

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1,* 17-35. doi:10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022. doi:10.1162/jmlr.2003.3.4-5.993

Chaney, A. J. B., & Blei, D. M. (2012, March). Visualizing topic models. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (IWSCM).* Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4645/5021

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458-474. doi:10.1037/met0000111

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In P. Resnik, R. Resnik, & M. Mitchell (Eds.), *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical*

Reality (pp. 51-60). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W14-3207

De Battisti, F., Ferrara, A., & Salini, S. (2015). A decade of research in statistics: A topic model approach. *Scientometrics*, *103*(2), 413-433. doi:10.1007/s11192-015-1554-1

Erdfelder, E., & Bošnjak, M. (2016). „Hotspots in Psychology“: A new format for special issues of the *Zeitschrift für Psychologie*. *Zeitschrift für Psychologie*, *224*(3), 141-144. doi:10.1027/2151-2604/a000249

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software 25*(5), 1-54. doi:10.18637/jss.v025.i05

Fischer, A. (2015). How to determine the unique contributions of input-variables to the nonlinear regression function of a multilayer perceptron. *Ecological Modelling, 309,* 60-63. doi:10.1016/j.ecolmodel.2015.04.015

Friedman, D. P. (2008). Public outreach: A scientific imperative. *Journal of Neuroscience*, *28*(46), 11743-11745. doi:10.1523/JNEUROSCI.0005-08.2008

Goldstone, A., Galán, C., Lovin, C. L., Mazzaschi, A., & Whitmore, L. (2014). *An Interactive Topic Model of* Signs, edited by Andrew Goldstone. *Signs at 40*. Retrieved from http://signsat40.signsjournal.org/topic-model

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

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244. doi:10.1037/0033-295X.114.2.211

Grün, B. & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-30. doi:10.18637/jss.v040.i13

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, *52*(1), 19-30. doi:10.1037/a0036841

Jiang, H., Qiang, M., & Lin, P. (2016). A topic modeling based bibliometric exploration of hydropower research. *Renewable and Sustainable Energy Reviews*, *57*, 226-237. doi:10.1016/j.rser.2015.12.194

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493-506. doi:10.1037/met0000105

Krampen, G. & Trierweiler, L. (2013). Research on emotions in developmental psychology contexts: Hot topics, trends, and neglected research domains. In C. Mohiyeddini, M. Eysenck & S. Bauer (Eds.), *Handbook of psychology of emotions. Recent theoretical perspectives and novel empirical findings. Volume 1* (pp. 63-79). New York: Nova Science Publishers.

Krampen, G. (2016). Scientometric trend analyses of publications on the history of psychology: Is psychology becoming an unhistorical science? *Scientometrics*, *106*(3), 1217-1238. doi:10.1007/s11192-016-1834-4

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the Association for Information Science and Technology*, *63*(10), 1973-1986. doi:10.1002/asi.22628

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009, August). Polylingual topic models. In P. Koehn & R. Mihalcea (Eds.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2* (pp. 880-889). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/old_anthology/D/D09/D09-1.pdf#page=918

Natale, F., Fiore, G., & Hofherr, J. (2012). Mapping the research on aquaculture. A bibliometric analysis of aquaculture literature. *Scientometrics, 90*(3), 983-999. doi:10.1007/s11192-011-0562-z

Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., ... Seligman, M. E. (2016). Women are warmer but no less assertive than men: Gender and language on Facebook. *PloS One*, *11*(5), e0155885. doi:10.1371/journal.pone.0155885.t003

Paul, M. J., & Girju, R. (2009, September). Topic modeling of research fields: An interdisciplinary perspective. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference RANLP-2009* (pp. 337-342). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://www.anthology.aclweb.org/R/R09/R09-1.pdf#page=361

Ponweiser, M., Grün, B., & Hornik, K. (2014). Finding scientific topics revisited. In M. Carpita, E. Bentari, & E. Qannari (Eds.), *Advances in latent variables* (pp. 93-100). Cham, Switzerland: Springer International Publishing. doi:10.1007/10104_2014_11

R Core Team (2017). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Resnik, P., Garron, A., & Resnik, R. (2013, October). Using topic modeling to improve prediction of neuroticism and depression. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1348-1353). New York, NY: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D13-1133

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004, July). The author-topic model for authors and documents. In M. Chickering & J. Halpern (Eds.), *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487-494). Arlington, VA: AUAI Press. Retrieved from https://mimno.infosci.cornell.edu/info6150/readings/398.pdf

RStudio Team (2016). RStudio: Integrated development for R [Computer software]. RStudio, Inc., Boston, MA. Retrieved from http://www.rstudio.com/

Schofield, A., Magnusson, M., & Mimno, D. (2017). Understanding text pre-processing for latent Dirichlet allocation. In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 432-436). New York, NY: Association for Computational Linguistics. Retrieved from http://www.cs.cornell.edu/~xanda/winlp2017.pdf

Sievert, C., & Shirley, K. E. (2015). LDAvis: Interactive visualization of topic models. R package version 0.3.2 [Computer Software]. Retrieved from https://CRAN.R-project.org/package=LDAvis

Sievert, C., & Shirley, K. E. (2014, June). LDAvis: A method for visualizing and interpreting topics. In J. Chuang, S. Green, M. Hearst, J. Heer, & P. Koehn (Eds.), *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63-70). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W14-3110

Steyvers, M., & Griffiths, T. L. (2008). Rational analysis as a link between human memory and information retrieval. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for a Bayesian cognitive science* (pp. 329-350). Oxford: Oxford University Press.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014, January). Understanding the limiting factors of topic modeling via posterior contraction analysis. In E. P. Xing (Ed.), *31st International Conference on Machine Learning (ICML 2014)* (pp. 190-198). Stroudsburg, PA: International Machine Learning Society. Retrieved from http://proceedings.mlr.press/v32/tang14.pdf

Thelwall, M., & Thelwall, S. (2016). Development studies research 1975-2014 in academic journal articles: The end of economics? *El Profesional de la Información*, *25*(1), 47-58. doi:10.3145/epi.2016.ene.06

Tuleya, L. G. (Ed.). (2007). *Thesaurus of psychological index terms* (11th ed.). Washington, DC: American Psychological Association.

Vulić, I., De Smet, W., Tang, J., & Moens, M. F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, *51*(1), 111-147. doi:10.1016/j.ipm.2014.08.003

Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.). A*dvances in Neural Information Processing Systems 22 (NIPS 2009)* (pp. 1973-1981). LA Jolla, CA: Neural Information Processing Systems. Retrieved from http://dirichlet.net/pdf/wallach09rethinking.pdf

Wang, C., & Blei, D. M. (2011, August). Collaborative topic modeling for recommending scientific articles. In C. Apte, J. Ghosh, & P. Smyth (Eds.), *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 448-456). New York, NY: ACM. doi:10.1145/2020408.2020480

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

Zheng, B., McLean, D. C., & Lu, X. (2006). Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics*, *7*(1), 58. doi:10.1186/1471-2105-7-58

ZPID - Leibniz-Zentrum für Psychologische Information und Dokumentation (Eds.). (2016). *PSYNDEX terms* (10th ed.). Trier, Germany: ZPID. Retrieved from https://www.zpid.de/pub/info/PSYNDEXterms2016.pdf

Footnotes

[1]A list of all codes can be retrieved from http://www.apa.org/pubs/databases/training/class-codes.aspx. Each publication can be linked to one or more main classifications (e.g., "Psychometrics & Statistics & Methodology," "Human Experimental Psychology," or "Personality Psychology"), and/or respective subcategories (e.g., a publication classified as "Psychometrics & Statistics & Methodology" may be classified more specifically as "Sensory & Motor Testing" or "Clinical Psychology Testing").

Table 1

Example of Four Documents Consisting of 16 Terms Each

Document			
1	2	3	4
love	happiness	disgust	amazement
hate	joy	anger	surprise
fear	serenity	rage	joy
disgust	love	hate	happiness
intervention	therapy	psychoanalysis	psychotherapy
therapist	therapist	transference	counseling
client	client	client	disorder
disorder	treatment	disorder	treatment
mother	intervention	treatment	outcome
brother	disorder	intervention	exposition
sister	parents	mother	client
father	siblings	father	therapist
school	learning	parents	parents
learning	teacher	child	mother
grades	class	grades	college
class	college	achievement	university

Table 2

Five Most Common Terms of the Resulting Topics (Idealization)

Topic			
1	2	3	4
„emotions“	**„therapy“**	**„family“**	**„education“**
love	client	parents	class
joy	disorder	mother	college
happiness	therapist	father	grades
disgust	intervention	child	learning
amazement	treatment	brother	teacher

Note. The topic titles are descriptive terms provided by the authors and were not generated by the model.

Table 3

Illustration of Document-Topic Probabilities (θ)

	Topic				
Document	1	2	3	4	Sum
1	0.250	0.250	0.250	0.250	1
2	0.250	0.375	0.125	0.250	1
3	0.250	0.375	0.250	0.125	1
4	0.250	0.500	0.125	0.125	1
Mean	0.250	0.375	0.188	0.188	1

Note. Document 1 addresses all four topics with equal shares ($1/k$, with k = number of topics), whereas Documents 2 to 4 show different topic probabilities. By mean probabilities, Topic 2 is addressed with more than average probability and, thus, can be interpreted as the most popular.

Table 4

Log-Likelihoods (LL) of Topic Models by Different Numbers of Topics (k)

k	100	150	200	300	400	500
LL	-8234278	-7491948	-7032583	-6403997	-5978100	-5695993

Table 5

Top 15 Terms of the Ten Hottest Topics

Topic	Top 15 Terms
364	Functional Magnetic Resonance Imaging, Cerebral Blood Flow, Prefrontal Cortex, Amygdala, Neuroanatomy, Biological Neural Networks, Cingulate Cortex, Brain, Oxygenation, Insula, Rewards, Striatum, Hippocampus, Brain Connectivity, Cognitive Control
249	Functional Magnetic Resonance Imaging, Cerebral Blood Flow, Brain, Parietal Lobe, Prefrontal Cortex, Neuroanatomy, Frontal Lobe, Temporal Lobe, Oxygenation, Neuroimaging, Magnetic Resonance Imaging, Biological Neural Networks, Occipital Lobe, Visual Cortex, Spectroscopy
386	Internet, Computer Mediated Communication, Online Therapy, Online Social Networks, Internet Usage, Electronic Communication, Communications Media, Websites, Social Media, Virtual Reality, Computer Assisted Therapy, Cellular Phones, Privacy, Telemedicine, Information Technology
459	Genes, Polymorphism, Genetics, Serotonin, Genotypes, Dopamine, Alleles, Biological Markers, Phenotypes, Attention Deficit Disorder With Hyperactivity, Susceptibility (Disorders), Neurotransmission, Brain Derived Neurotrophic Factor, Neural Receptors, Tryptophan
371	Cross-Cultural Differences, Human Migration, Cross-Cultural Communication, Cultural Sensitivity, Cross-Cultural Treatment, Multiculturalism, Expatriates, Transcultural Psychiatry, International Organizations, Cross-Cultural Counseling, Globalization, Multicultural Education, Foreign Workers, Acculturation, Racial And Ethnic Differences
323	Magnetic Resonance Imaging, Brain, Neuroimaging, Neuroanatomy, Hippocampus, Gray Matter, Brain Size, Tomography, Prefrontal Cortex, White Matter, Amygdala, Cingulate Cortex, Cerebral Cortex, Temporal Lobe, Morphology
86	Posttraumatic Stress Disorder, Emotional Trauma, Refugees, Trauma, War, Victimization, Torture, Persecution, Survivors, Violence, Injuries, Asylum Seeking, Exposure Therapy, Human Migration, Transgenerational Patterns
344	Posttraumatic Stress Disorder, Emotional Trauma, Trauma, Eye Movement Desensitization Therapy, Stress Reactions, Intrusive Thoughts, Adjustment Disorders, Acute Stress Disorder, Traumatic Neurosis, Posttraumatic Growth, Complex PTSD, Exposure Therapy, Accidents, Medical Personnel, Metaphor
365	Attention, Visual Attention, Selective Attention, Visual Search, Distraction, Cues, Reaction Time, Stimulus Parameters, Eye Movements, Attentional Capture, Visual Perception, Stimulus Salience, Visual Stimulation, Attentional Bias, Divided Attention
95	Emotional Trauma, Posttraumatic Stress Disorder, Trauma, Dissociation, Dissociative Disorders, Early Experience, Dissociative Identity Disorder, Depersonalization, Borderline Personality Disorder, Neurobiology, Introjection, Dissociative Patterns, Amnesia, Psychodynamic Psychotherapy, Depersonalization/Derealization Disorder

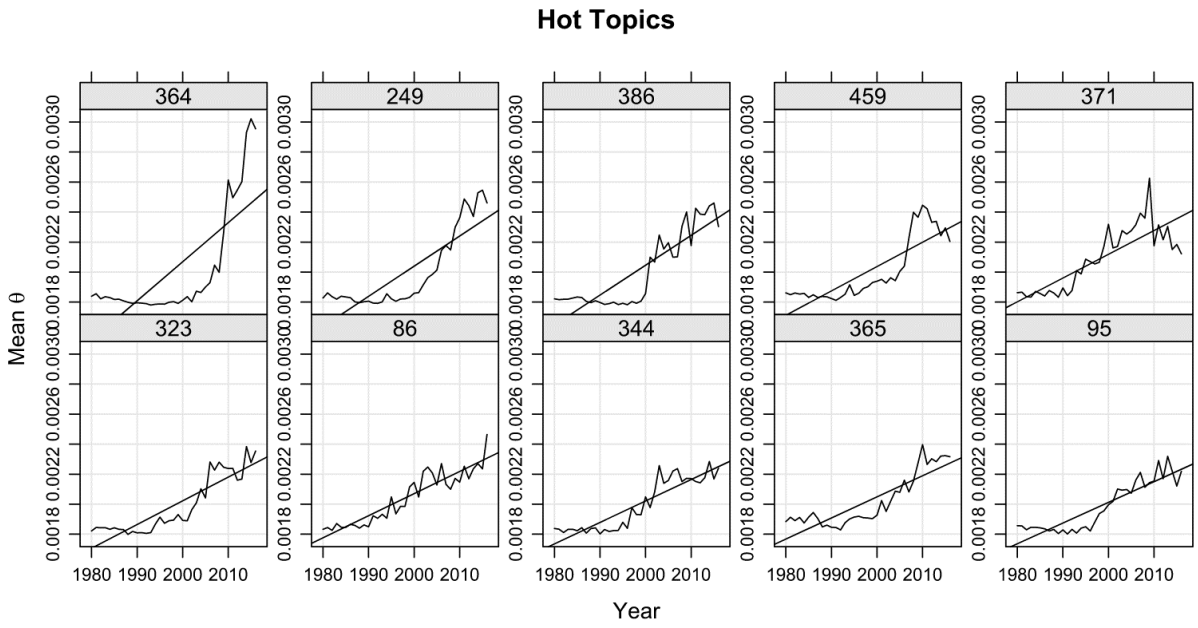

Figure 1. Mean values of document-topic probabilities θ by publication year for the 10 hottest topics with added linear regression line. The topics are described in Table 5.

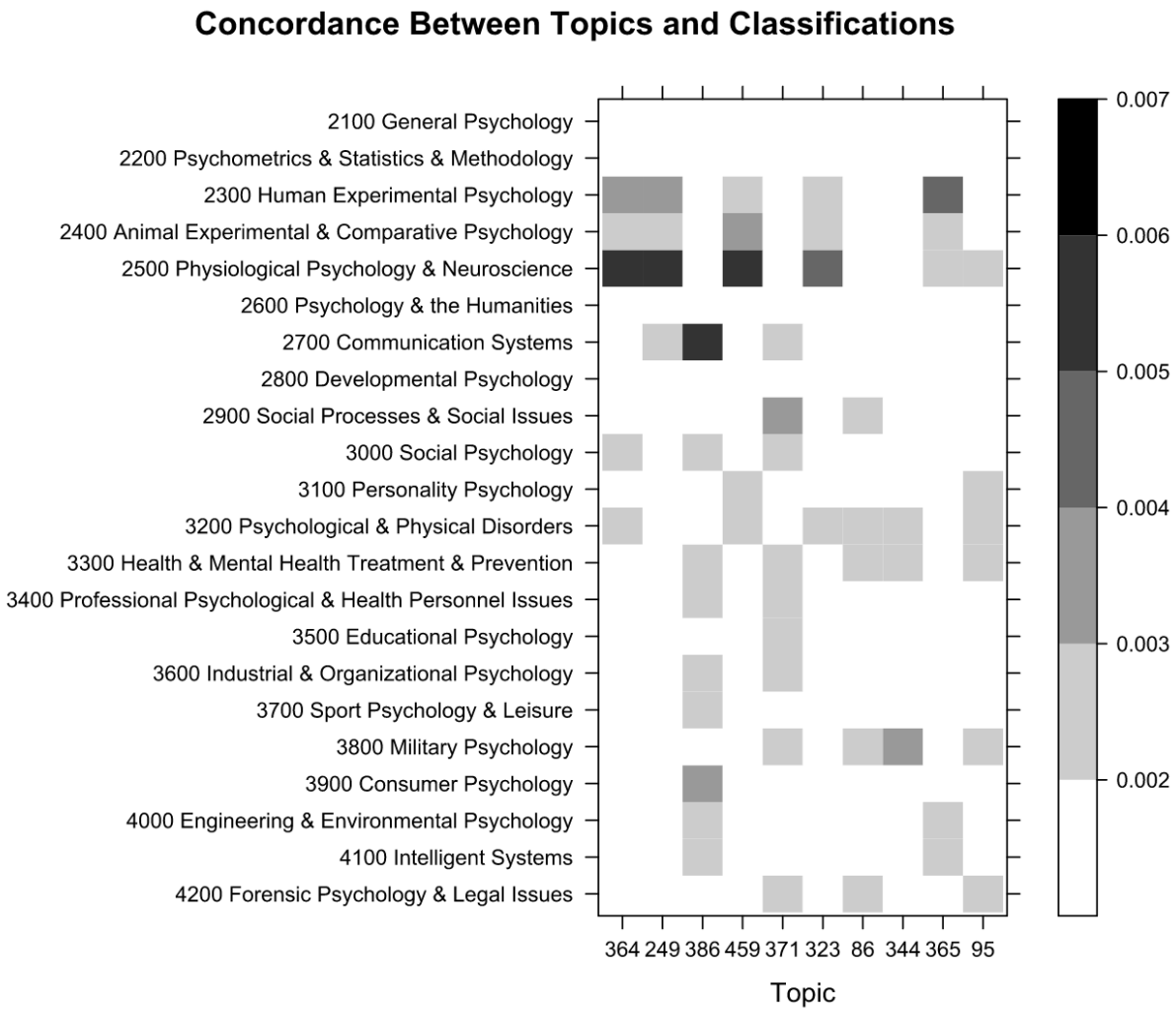

Figure 2. Level plot showing mean document-topic probabilities (θ) by topics and main classifications. Only the 10 hottest topics are displayed. Darker cells represent higher values of θ. For example, in the publications that were classified as "2500 Physiological Psychology & Neuroscience," the highest mean θ resulted for Topics 364, 249, 459, and 323. The respective topics are described in Table 5.

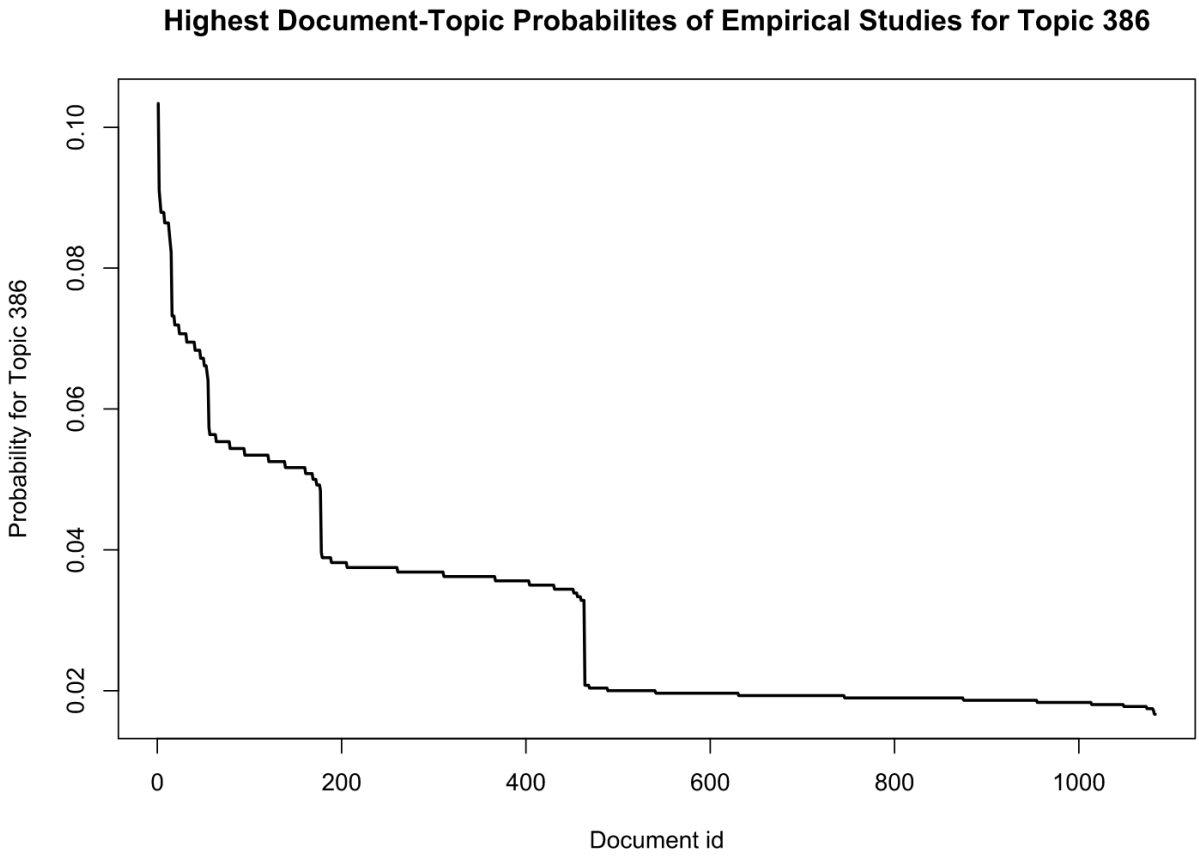

Figure 3. Document-topic probabilities (θ) of *n* = 1,083 empirical studies for Topic 386 (sorted by θ). Only documents with θ higher than average are shown.

Electronic Supplementary Material (ESM 1): R Code of the Analyses (ESM1_Code.R)

Electronic Supplementary Material (ESM 2): List of Topics for $k = 500$ (ESM2_Topics.csv)