Scientometrics (2020) 125:1229–1251
https://doi.org/10.1007/s11192-020-03640-0

A structural topic model approach to scientific reorientation of economics and chemistry after German reunification

Andreas Rehs[1]

Received: 15 December 2019 / Published online: 5 August 2020

Abstract
The detection of differences or similarities in large numbers of scientific publications is an open problem in scientometric research. In this paper we therefore develop and apply a machine learning approach based on structural topic modelling in combination with cosine similarity and a linear regression framework in order to identify differences in dissertation titles written at East and West German universities before and after German reunification. German reunification and its surrounding time period is used because it provides a structure with both minor and major differences in research topics that could be detected by our approach. Our dataset is based on dissertation titles in economics and business administration and chemistry from 1980 to 2010. We use university affiliation and year of the dissertation to train a structural topic model and then test the model on a set of unseen dissertation titles. Subsequently, we compare the resulting topic distribution of each title to every other title with cosine similarity. The cosine similarities and the regional and temporal origin of the dissertation titles they come from are then used in a linear regression approach. Our results on research topics in economics and business administration suggest substantial differences between East and West Germany before the reunification and a rapid conformation thereafter. In chemistry we observe minor differences between East and West before the reunification and a slightly increased similarity thereafter.

Keywords Topic modelling · German reunification · Dissertations · Structural topic modelling · Research field mapping

JEL Classification O33 · O52 · P30 · Z13

✉ Andreas Rehs
rehs@incher.uni-kassel.de

[1] International Centre for Higher Education Research, University of Kassel, Mönchebergstr. 17, 34127 Kassel, Germany

Introduction

Growth of science, growth of topical difference identification issues?

Classification systems of scientific literature play a central role in bibliometrics (Glänzel and Schubert 2003) and will become more and more important with the exponentially growing amount of scientific literature. From World War II to the early 2000s, the stock of scientific literature is estimated to have doubled about every 9 years (Bornmann and Mutz 2015) and in 2009 amounted to over 50 million publications (Jinha 2010). These growth rates and underlying numbers raise concerns that the large current and future stock of knowledge will become more and more difficult to structure for single scientists (Landhuis 2016) and established databases (Larsen and Von Ins 2010). Traditional classification systems rely on keyword assignments, expert-based classification of subjects, and forward and backward citations to embed a publication in the network of knowledge flows in scientific literature (De Bellis 2009). These methods include high levels of complexity reduction and therefore a loss of knowledge in the content of the scientific publications. Practically, subtle but often decisive differences between two papers on the same topic can therefore hardly be addressed without having expert-level knowledge in the respective scientific field. In the same manner, topical overlaps between loosely related papers cannot be detected without having expert knowledge in both papers' fields. The addition of more and more papers will eventually constrain the ability of experts to detect differences and similarities between papers. The large-scale quantification and detection of thematic differences in research topics is therefore an open problem in scientometric research. Advances in machine learning, especially in the statistical analysis of large text collections, alleviate these issues under certain circumstances. In this way precise difference detection between scientific texts can be feasible without having deep knowledge in the respective field.

The case of scientific reorientation in East and West Germany

In this paper we therefore develop and apply such a machine learning approach to difference detection based on the case study of dissertation titles written at East[1] and West[2] German universities in economics and business administration and chemistry before and after German reunification. German reunification is especially suited for investigating differences in research topics because the transition of the political system in East Germany went hand in hand with the transition of the scientific system. German reunification led to the dismantling in East Germany of a large number of chairs, institutes and research organizations, as well as a broad institutional restructuring in academia. Reasons included political motives, but in several instances also a mismatch between what had been researched under the old (socialist) system and what was considered interesting in the new one. This change affected social sciences more severely than natural sciences and therefore provided two different structures to investigate thematic differences and topical reorientation. In

[1] East Germany refers to the territory of the former German Democratic republic and today includes the German states of: Thuringia, Saxony, Saxony-Anhalt, Mecklenburg Western Pomerania and Brandenburg. The city of Berlin was separated in East- and West-Berlin during the German division.

[2] West Germany refers to the territory of the Federal Republic of Germany from 1949 to 1990. West Germany includes the states: Hesse, Lower-Saxony, Bavaria, Baden-Württemberg, Saarland, Rhineland-Palatinate, North-Rhine Westphalia, Bremen, Hamburg and Schleswig–Holstein.

these two structures, motives and incentives for individual scientists in the two disciplines and parts of Germany to change research topics differed substantially and may have manifested in minor and major thematic differences. The chapter "Historical background" will therefore elaborate on the disciplinary and general historical circumstances before and after the reunification.

Dissertation as a data source

Journal publications and their linked indicators, such as citations, are the main subject ofinvestigation in scientometric research and have contributed to substantial advances in the field (e.g., Garfield 1972; Hirsch 2005). However, under certain historical, institutional and disciplinary circumstances, such as in our case, journal articles are not the best means of inquiry.[3] We therefore use dissertation titles as an alternative source of information to identify and track the differences in the two disciplines in Germany before and after reunification. Dissertation titles offer several potential advantages for our approach and are, despite limited use in scientometrics (Morichika and Shibayama 2016), amply available in Germany. This is because every doctoral student is mandated to send in a copy of his or her dissertation to the German National Library (Deutsche Nationalbibliothek). The German National Library archives the dissertation and stores some basic author and dissertation information in its electronic catalogue. We have access to this catalogue, which provides us an almost complete list of dissertations that were submitted in both parts of Germany, since 1970. Thus, we have a good picture of the thematic landscape during our period of investigation in Germany. Our work is based on a number of presumptions: First, in Germany the doctoral advisor (often dubbed the "Doktorvater") has a strong influence on the doctoral student and their choice of research topic. Moreover, the advisor is usually required to have a chair at a university, as only they are entitled to award PhDs. Therefore, the dissertation topics most likely represent the research topics present at a chair. Second, the title of a dissertation represents its content in a very condensed form. Together, these assumptions lead to the conjecture that the research focus of a chair is reflected in the titles of dissertations submitted at an a university with which he or she is affiliated. This allows us to draw conclusions on the general thematic landscape of university research in Germany during our period of investigation.

A structural topic model approach to differences in dissertation titles

Our main effort was in applying a probabilistic text model ("structural topic model") to these dissertation titles, aggregating the outcomes and then incorporating them into a linear regression framework, which allows us to calculate the level of difference between

[3] At the most general historical level, journal publications have not been the dominant medium for scientific communication in disciplines where they are nowadays the standard form of publication. This especially applies to one of our subjects of investigation, namely economics and business administration in West Germany (Hicks 1999; Leininger 2009). We found no literature that reflects on the publication system and culture in economics and business administration in the German Democratic Republic. Regarding chemistry in West Germany, Weingart et al. (1991) indicate that journal publications were in the 80s and today still are the most popular means of publication (Hahn 2009). In the German Democratic Republic, publications in chemistry were common, but due to isolation the German Democratic Republic underperformed in comparison to West Germany in terms of relative publications per capita.

dissertation titles by regional and temporal origin of the dissertation. In this way our approach demonstrates how to identify and track differences between scientific work on the level of individual researchers, but also larger entities of the scientific system, such as different scientific disciplines or parts of a country. In our case study, we find in economics and business administration research topics considerable differences between East and West Germany before reunification. After reunification, we observe a strong and rapid conformation. In chemistry there are few differences between East and West before reunification. Afterwards, the results suggest a moderate thematic convergence.

Historical background

The scientific system and doctoral education in the German Democratic Republic

Since the birth of the two Germanies in 1949, the intra-German relationship has been characterized by a competition of political (and economic) systems. Walter Ulbricht, prominent veteran socialist politician of the German Democratic Republic (GDR) was renowned for his saying "overtaking without catching up". The early socialists strove to demonstrate the superiority of socialism over capitalism, with scientific and technological achievements playing a central role. Even the constitution of the GDR (§ 2, Abs 1) claimed that the foremost aim of a socialist society was to increase the effectiveness of scientific and technological development and labour productivity (Volkskammer der DDR 1976). This orientation of scientific advancement on aspects of productivity dated back at least to Lenin and had consequences for the academic landscape of the GDR. Industrial application of research findings was heavily emphasized. Basic research was carried out almost exclusively by universities, but free choice of the research subjects was increasingly restricted and almost non-existent beginning in the 1960s (Gruhn and Lauterbach 1977). PhD candidates had minimal freedom in choosing their research subjects. In the case of Humboldt University in Berlin shows that roughly two-thirds of dissertation topics followed the five-year research plans of the government (Wollgast 2001). Furthermore, international contact was more or less limited to other socialist states and access to Western world academics and their publications was difficult to gain (Mann 1979). Limited financial resources made internationally competitive research impossible in the majority of scientific fields. However, the conditions of career advancement in academia closely resembled those in West Germany. The average student in the GDR had to complete a basic and an advanced (or specialized) part of his study to earn a degree. Afterwards, a dissertation (Promotion A) had to be written to obtain the title "Dr." in a scientific field. In contrast to the Federal Republic of Germany, the GDR had universal requirements for the award of a PhD degree, which included a fair amount of ideology (Deutsche Demokratische Republik 1968, §5, Abs. 1). PhD degrees could be earned through research studies (2–3 years long, similar to a graduate school), employment at a university chair (usually four years' contract) or distinction in industrial and societal engagement (similar to an external PhD candidate) (Belitz-Demiriz et al. 1990; Guenther 1989).

Unlike in the GDR, the scientific system of West Germany during our period of investigation was (and still is) free of ideological constraints. The constitutional (basic law) "freedom of teaching and research" (§5, Abs. 3) guaranteed vast autonomy for university researchers. Regarding factors that could have implicitly constrained freedom of research in West Germany in the 1980s and early 1990s, Peisert and Framhein (1994) argue that, in

the case of third party funding, there was no strong influence from semi-public and public institutions on research topic choices. The systems of doctoral education in East and West Germany closely resembled each other; both countries doctoral students were predominately employed at the chairs directly; graduate schools played a minor role. However, the level of involvement of ideology in doctoral education clearly distinguished the two.

The transition and political change in Germany in 1990 had a deep impact on academic institutions, most notably in scientific fields that were heavily affected by socialist ideology. The prime example is economics and business administration, which was almost completely dismantled and rebuilt from the ground up, often involving new personnel, structures and research agendas. Kolloch (2001) reports that by 1994 90% of the economics and business administration chairs at the biggest East German university, HU Berlin,) were replaced with West Germans.

In chemistry the historical preconditions were quite different. In the GDR, the discipline was considered to be a crucial scientific productive force that would directly and indirectly increase economic output. Chemistry and other natural sciences were therefore oriented to the requirements of the local industry (Meske 2004), which led to a much greater focus on applied research in East Germany. GDR policymakers, for example, built a technical college in the centre of the East German chemistry cluster Leuna-Buna-Bitterfeld. The GDR chemical industry and, in consequence, the discipline of chemistry was dependent on crude oil deliveries from the Soviet Union to produce precursors and final chemical products. The GDR, however, used the dominant share of crude oil deliveries from the Soviet Union to refine petrol, which was to a large extent exported in order to bring in much-needed hard, foreign currency. This petrol-focused production caused a shortage in the production of other products based on crude oil (e.g., rubber and plastic). East German chemistry therefore researched non-oil-based ways of producing such goods. Lignite was a viable alternative, since East Germany had large lignite resources and existing processing facilities dating back to World War II. For the scientific discipline of chemistry this lignite based "business model" of the GDR resulted in a strong emphasis on related research problems. Chemistry as a discipline was therefore politically determined, applied and focused foremost on the special demands of East German chemical industry. For West Germany we find no indication of any profound specialization or a general focus on applied topics in chemistry. This may be a consequence of the constitutional right of freedom in teaching and research and a conservative industrial policy.

Data

The two disciplines, economics and business administration and chemistry, and their historical background before and after German reunification are therefore suited for our analysis of identifying research topic differences. They provide two structures: for economics and business administration, a structure with substantial topical heterogeneity before and after reunification; and for chemistry, one with relative topical homogeneity. In the following section, we will describe the processing steps used to obtain the final dataset of dissertation titles (Rehs 2020a).

We use the online catalogue of the German National Library as the basis for our analysis. The catalogue lists the vast majority of PhD dissertations submitted at German universities, including the GDR. There are entries for approximately one million PhD dissertations, which are classified by subject. We use this classification to distinguish between

economics and business administration and chemistry. Due to the peculiarities of German medical dissertations, we eliminate dissertations which are cross-listed in chemistry and medicine. Furthermore, we employ information on university location (cities, name of university or a combination of both) to separate East from West dissertations.[4] We assume that reorientation of research topics after the reunification continued until 2010. To obtain a picture of the thematic landscape before reunification, we consider the years 1980–1989. The years 1990–1994 are eliminated from our data, since the replacement of East German chairs took several years and the number of observations from East German university dissertations dropped significantly during this time period.

In the next step, we paste every dissertation title and subtitle into one string and standardize this string. Our pre-processing includes standard text-mining methodology: transformation to lowercase, removal of punctuation, language detection and removal of non-German titles, stemming, n-gram detection and removal of very frequent words, rare words, stopwords and short titles. Different languages in a text collection can considerably distort the outcomes of the topic modelling algorithm to be presented due to problems with (text-mining) token recognition. Although differently spelled words can have the exact same meaning in two languages, they are considered statistically as different tokens in text machines. Solutions based on translation cause more problems than they solve. Our approach is therefore to exclude all titles written in English. We are aware of the downsides of this procedure and might miss some important dissertations that are addressed to an international audience. Dissertations written in German might also differ in quality. Nevertheless, as our language identification algorithm (Ooms 2018) shows, English titles only account for roughly 10% of the dissertations. The small number of English titles would therefore distort the statistical inference based on topic modelling. All titles identified as neither German nor English are defaulted to German.

Mentioned n-grams are applied because some words are by nature bounded, like "United" and "States". To improve the performance of the topic model to be presented, we want the algorithm to treat these words as one character. Bigrams are two bounded words and trigrams three bounded words. In both corpora we count the most frequent bi- and trigrams. We assume that only the top bi- and trigrams add relevant context for the subsequent algorithm. For both economics and business administration and chemistry, we set the boundary for relevant n-grams at the top 1%. We proceed by searching these n-grams in every string. If they occur, we add them to the string and remove the words that composed them.

We remove very frequent and very rare words for reasons of complexity reduction and minor relevance for topic modelling. Very frequent have the same properties as stopwords, but are not included in standard stopword dictionaries since they are dataset specific. They don't add relevant context; rather, they are commonly used terms within a dataset and identically distributed across all documents (e.g., for dissertation titles, "investigation" or "method" may appear very frequently). We set the threshold for removal at the upper 0.1% limit of the most frequent words. The same holds for very rare words. Because of their low frequency, they don't add context, and are removed if they appear fewer than 3 times in total.

Finally, we delete very short titles from our data set. Since topic modelling infers the topic distributions by drawing words from each title numerous times, titles consisting of

[4] We exclude observations which are labelled "Uni Berlin", since it is uncertain whether the university is in East or West Berlin.

only few words can be problematic because there is less room for randomness in each title. We therefore exclude titles containing fewer than five words.

Topic modelling in large-scale text analysis

The latent Dirichlet allocation

To address our research question we use topic modelling, which is a family of probabilistic methods for analysing large text collections. Topic modelling has found various applications in scientometrics, such as in investigating the topics that construct scientific publications (Blei and Lafferty 2007). Any topic modelling algorithm is, in general, a generative model of word counts. In our case that means we define a data-generating process for each dissertation title and then use the data to find the most likely values for the parameters within the model.

The most common topic modelling algorithm is the latent Dirichlet allocation (short: LDA, Blei 2012; Blei et al. 2003). In the LDA algorithm our dissertation titles are represented as mixtures of topics. In these mixtures, each word within a given dissertation title belongs to exactly one topic. Single dissertation titles can therefore be considered as vectors of topic proportions, which indicate the percentage of words belonging to each topic. In the following section we will describe the statistical methodology and orient on the notation and description of Roberts et al. (2014a, b, 2016).

The generative process in LDA starts by considering each dissertation title (index: Diss) as a distribution over topics (θ_{Diss}), which is drawn from a global prior distribution. In the next step, for each word in the dissertation title (indexed by n), the LDA algorithm draws a topic (z) for that word from a multinomial distribution based on its distribution over topics ($z_{Diss,n} \sim \text{Mult}(\theta_{Diss})$). Depending on the topic selected, the observed word $w_{Diss,n}$ is drawn from a distribution over the vocabulary $z_{Diss,n} \sim \text{Mult}(\beta_{z_{Diss,n}})$, where $\beta_{k,v}$ is the probability of drawing the vth word in the vocabulary for topic k.

A hypothetical pre-1990 East German title in economics might therefore be represented as a mixture over 10 topics. Topics are, again, a distribution over words that are more or less likely to be related to that topic (e.g., "Marx", "worker", "class" might each have high probability in the same topic). The LDA is completed by assuming a Dirichlet prior for the topic proportions such that $\theta_{Diss} \sim \text{Dirichlet}(\alpha)$. However, there are disadvantages that come along with the application of LDA. The resulting posterior distributions can have many local modes. That means that different initializations can produce different solutions. In order to address this issue, we use the spectral initialization procedure described in Arora et al. (2013), which is also implemented in the R package on structural topic modelling (Roberts et al. 2014a).

Structural topic modelling

Structural topic modelling is an extension of the LDA process described above which allows covariates of interest (such as the temporal origin or university of the dissertation) to be included in the prior distributions for dissertation-topic proportions and topic-word distributions. Thus, the covariates offer a method of "structuring" the prior distributions in the topic model, including additional information in the statistical inference procedure. The topic prevalence (as described in the LDA section) can therefore

be influenced by some set of covariates X through a standard regression model with covariates $\theta \sim \text{LogisticNormal}\ (X\gamma, \Sigma)$. In contrast to the described LDA algorithm, we abolish the assumption that topical prevalence (how much a topic is discussed by a covariate) is constant across all dissertation titles. This is a major improvement in comparison to LDA and allows the parameters that generated the dissertation title to be reconstructed more precisely.

We use university and year dummies of the dissertation as topical prevalence variables in our structural topic models on chemistry and economics and business administration. We argue that these variables are best suited to capture temporal and university level variation in dissertation titles and are different from the main independent variables in the regression framework to be presented. Year dummies as topic prevalence variables should capture trends and temporarily popular topics in the 25-year span of our investigation. For universities, irrespective of their East or West German background, we presume that there are regionally bound topics. This is because the chairs at universities might have inherent topics that are reflected in the dissertations they produce. Therefore, we include university dummies as the second set of topical prevalence variables in our topic model. In structural topic models, proportions (θ) can also be correlated (see also Blei and Lafferty 2007); i.e., in a given dissertation title, the high proportion of a topic that is related to socialism might also increase the likelihood of high proportion of a related topic (e.g., a topic related to Leninism).

In our structural topic modelling, we stopped at the point where θ can be influenced by some set of covariates X through a standard regression model with covariates $\theta \sim \text{LogisticNormal}\ (X\gamma, \Sigma)$. The next step in the structural topic model algorithm is described as: "For each word (w) in the response, a topic (z) is drawn from the response-specific distribution, and, depending on the topic, a word is chosen from a multinomial distribution over words parameterized byβ, which is formed by deviations from the baseline word frequencies (m) in log space ($\beta_k \propto \exp\ (m + K_k)$)" (Roberts et al. 2014b, p. 4). This distribution can include a second set of covariates that can model how word frequencies between values of that covariate can differ. Within a "socialistic" topic, this allow GDR dissertations, as indicated by a variable, to use the word "Marx" more frequently than dissertations from West Germans (they might use "Engels" more often instead). Since the used version of the R package (Roberts et al. 2014a) allows the inclusion of only one variable for such "topical content" and our approach would require several other variables, we don't include any variable of such kind.

When it comes to finally fitting the structural topic model, the major problem is in the mathematically intractable posterior distribution. To solve such a problem, Roberts et al. (2014b, 2016) developed a method for approximate inference based on variational expectation–maximization algorithms (Blei et al. 2017; Dempster et al. 1977) that, upon convergence, give estimates of the model parameters. Convergence is achieved when the change in the approximate variational lower bound between the iterations becomes very small. We accordingly set the value for convergence to 1e−06.

In conclusion, there are two major improvements that structural topic modelling provides for our setting as compared to LDA. First, topics can be correlated, which much better reflects the "true" data-generating process behind dissertation titles and science in general. The second major improvement is that each dissertation title has its own prior distribution over topics defined by covariate X, rather than sharing a global mean.

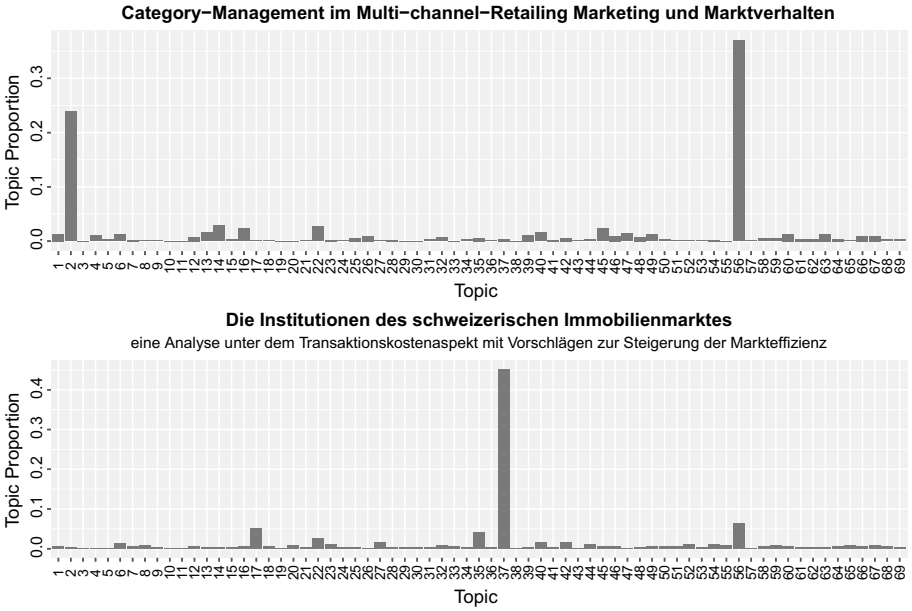

Fig. 1 Two title-topic distributions of the economics and business administration topic model

Topic model application and cosine similarity regression framework

In the next step, we estimate two separate structural topic models—one for economics and business administration and one for chemistry. For both we consider the whole period of investigation from 1980 to 1989 and 1995 to 2010. For each topic model we use 75% of the dissertations to estimate the model parameters. For the remaining 25%, we apply the topic models. This separation of training and test datasets is a standard procedure in machine learning and aims to detect overfitting of our models. Overfitting means that our topic model learns the data generating process of the underlying titles too well. In this way we lose model flexibility, which has negative impacts on the performance of the topic model on new, unseen dissertation titles. The final training and test set sizes in economics and business administration are a randomly sample of processed dissertation titles and include 6855 observations for the training and 1767 for the test set. In chemistry, sizes are 10,361 and 2580. East German test titles account for 317 dissertations in chemistry and 338 dissertations in economics and business administration (training and test set). In economics and business administration this broadly reflects the population size of East Germany (about 18% that of West Germany). In chemistry we find no explanation for the proportionally smaller number of dissertations in East Germany (9%).

When finally fitting our topic models, we arrive at 76 topics in chemistry and 69 in economics and business administration. One result of the two topic model applications is that we obtain a topic distribution for every title. Figure 1 illustrates the topic distribution

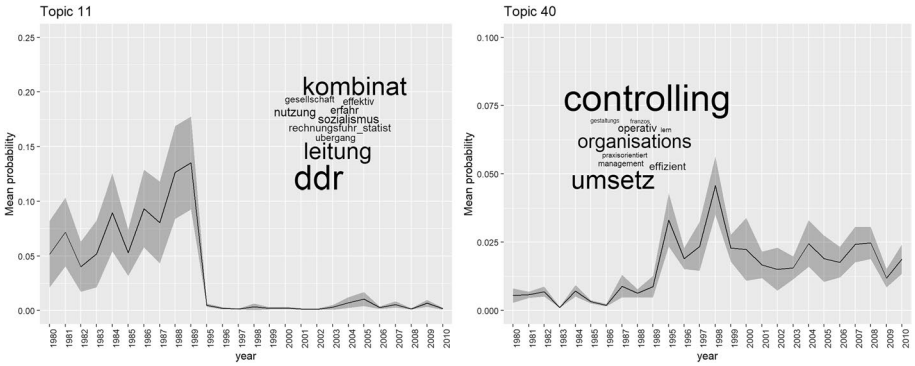

Fig. 2 Topical prevalence of two economics and business administration topics

Fig. 3 Mean topic prevalence before and after German reunification

of two titles in our topic model for economics and business administration.[5,6] Figure 2 now represents the top words with the highest β probability of two topics. We choose topics 11 and 40 in economics and show their yearly mean probability across all titles because they show how two, probably very antagonistic topics change in prominence over time. While

[5] To improve readability, we show the original title without the cleaning steps described in the previous chapter.

[6] Translation for title 1: Category management in multi-channel-retailing marketing and market behaviour. Translation for title 2: The institutions of the Swiss real estate market. An analysis under consideration of transaction costs with suggestions to market efficiency increase.

topic 11, which may indicate socialism, loses importance after 1990, topic 40, as a probable proponent of capitalism, on average gains importance. A list of words associated with other topics can be found in "Appendix". Since every topic is a probability distribution over words, top words may provide some indication of the underlying subject. However, interpretation should be done very cautiously, since the most probable words only represent a small fraction of the probability distribution. Moreover, most probable words are not necessarily the most exclusive words to a topic.

Figure 3 is similar to Fig. 2, but shows the distribution of the mean topic probability before and after the reunification for all topics. For economics and business, we can observe a high popularity of a small number of topics in East Germany before the reunification (such as topic 11). After reunification, high mean probability for single topics in one part of the country disappear. In direct comparison to economics and business, the mean probabilities for single topics in chemistry are small. However, there are still differences in some topics between East and West before the reunification. Remarkably, topics that weren't popular before the reunification in one part of the country became popular after the reunification. The popularity of topic 71, for example, increased considerably after the reunification in East Germany.

In order to compare the retrieved topic distribution of every title, we now use the cosine similarity measure, which has various applications in the comparison of topic model outcomes (see e.g., Ramage et al. 2010). The cosine similarity is a measure for the distance between two vectors and is defined between zero and one; values towards 1 indicate similarity. As topic proportions per dissertation title are vectors of the same length, the cosine similarity allows a comparison of the topic distribution between two documents. For the two exemplary dissertations we obtain a cosine similarity of 0.14.

In the next step, we calculate the cosine similarity between all topic-document distribution pairs (see dataset: Rehs 2020b). This means the topic distribution of title 1 is compared to title 2, title 3 and so on. We drop duplicate observations (e.g., when cosine similarity between 2 and 3 is the same as between 3 and 2). Since we know for every observation of the cosine similarity where both dissertations titles were written, we can employ this information in creating variables that can be attached to these similarity pairs (see Table 1 for an illustration of our dataset). We create a dummy *diff_part* that describes whether the two underlying dissertations for every similarity score are from different parts of Germany. The dummy variable *post95* indicates whether a dissertation was written after 1995.

Finally, we add university dummies to address differences in similarity scores arising at the university level. As the similarity score is calculated between two dissertations that were most often written at different universities, we consequently add dummies for both. The dummy *sameuni* indicates whether both titles in a pair are from the same university. In order to ease the interpretation of our dataset, we require both titles to be from the same year.

In the next step, we build different subsets of the data in order to address the peculiarities of our case study. For dissertations in chemistry, we build five data subsets: dissertations written before 1990 in both East and West Germany, dissertations written after 1990 in both Germanies, all dissertations written in West Germany and all dissertations written in East Germany for the time period studied. For economics and business administration, we proceed accordingly.

These subsets allow us to apply a linear regression framework, where the similarity score for each pair of dissertations is the dependent variable, and *diff_part, post95, same_uni* and the university dummies are the independent variables. This approach aims to aggregate the cosine similarities in order to demonstrate relationships between the

Table 1 Dataset structure. *Source*: Rehs (2020b)

Title 1	Title 2	*Cosine sim*	University 1 = *Univ_k*	University 2 = *Univ_k*	*same uni*	*same year*	*diff_ part*	*same part*	*post 95*
Die Institutionen des…	Catergory Management…	0.14	1	0	0	0	1	0	1
…	…	…	…	…	…	…	…	…	…

Table 2 Cosine similarities by subgroups

	n	Min	1st	Median	Mean	3rd	Max
Economics and business							
Cosine similarity all	62,586	0.0068	0.1462	0.2265	0.2578	0.3318	1
East = 1	10,458	0.0141	0.1460	0.2309	0.2843	0.3682	0.9960
West = 1	52,128	0.0068	0.1462	0.2258	0.2524	0.3266	1
diff_part = 1	18,607	0.0070	0.1235	0.1921	0.2213	0.2857	0.9951
diff_part = 0	43,979	0.0068	0.1589	0.2421	0.2732	0.3503	1
sameuni = 1	1308	0.0132	0.2377	0.3426	0.3844	0.4927	1
post95 = 1 and *diff_part* = 1	56,504	0.0068	0.1527	0.2342	0.2647	0.3398	1
post95 = 0 and *diff_part* = 1	6082	0.0121	0.1052	0.1631	0.1933	0.2406	0.8582
post95 = 0	12,784	0.0121	0.1433	0.2330	0.2819	0.3719	1
post95 = 1	49,802	0.0068	0.1470	0.2252	0.2561	0.3241	0.9974
Chemistry							
cosine similarity all	148,647	0.0027	0.0970	0.1691	0.2065	0.2729	1
East = 1	16,739	0.0034	0.1071	0.1769	0.2125	0.2765	0.9932
West = 1	131,908	0.0027	0.0968	0.1680	0.2057	0.2724	1
diff_part = 1	29,922	0.0034	0.1053	0.1751	0.2100	0.2740	0.9989
diff_part = 0	118,725	0.0027	0.0961	0.1675	0.2056	0.2726	1
same uni = 1	3744	0.0195	0.1606	0.2610	0.3023	0.3999	1
post95 = 1 and *diff_part* = 1	139,469	0.0027	0.0974	0.1684	0.2060	0.2724	1
post95 = 0 and *diff_part* = 1	9178	0.0034	0.1070	0.1800	0.2131	0.2800	0.9867
post95 = 0	47,329	0.0028	0.1046	0.1813	0.2164	0.2893	1
post95 = 1	101,318	0.0027	0.0953	0.1639	0.2018	0.2653	0.9998

Similarities of 1 are due to rounding

underlying groups of dissertation titles from East and West Germany and different periods. The regression formula is given by (1).

$$Cosine_{i,j} = \beta_0 + \beta_1 diffpart_{i,j} + \beta_2 post95 + \beta_3 diffpart * \beta_4 post95 + \beta_5 sameuni + \beta_n X_{i,k} + \beta_n X_{j,k} + \varepsilon_i. \quad (1)$$

j = dissertation 1 in pair, i = dissertation 2 in pair, k= university.

Table 2 and Fig. 4 show descriptive results for the cosine similarity by certain variables. In Fig. 4 we depict the mean similarity (with 95% conf. interval) of *diff_part* = 0 and *diff_part* = 1. The graph shows that the convergence in economics and business administration seems to have happened very quickly. In chemistry there was no convergence, as the average dissertation pair similarities by regional origins were never very different in our period of investigation.

Regarding the mean similarity of *diff_part* = 1 in Table 2, we observe in economics and business administration a significant increase from before to after 1990 and in chemistry a slight decrease. Chemistry topics were therefore, on average, more similar between East and West before the reunification than after the reunification. Nevertheless, the visual pattern of the mean by single years presented in Fig. 4 does not obviously support this finding. The results for *sameuni* in Table 2 also deliver interesting insights. Within a university, topics in both disciplines were considerably more similar than topics in different universities.

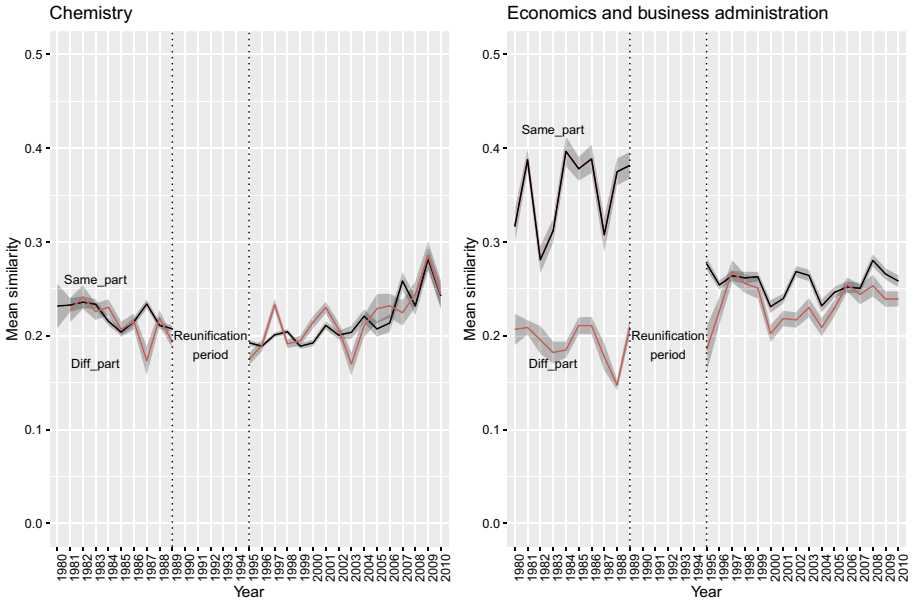

Fig. 4 Yearly mean cosine similarity between topic distributions in dissertation pairs

Table 3 aggregates our the chemistry cosine similarities in a linear regression framework. The pre models in both tables show the differences between East and West topics before reunification. Both pre models in Table 3 arrive at significantly negative coefficients of the variable *diff_part*. This indicates lower cosine similarity between two chemistry dissertations written in different parts of Germany. Full period model 1 in shows the differences between East and West Germany after reunification. The interaction of *diff_part* and *post95* in Table 3 is positive and statistically significant. This indicates increasing similarity between East and West German chemistry dissertations after the reunification. However, the effect diminishes after including university dummies and the variable *sameuni*, as shown in full period model 2. The last approach in chemistry concerns the thematic change within East or West German dissertations and is shown in models 5, 6, 7 and 8. The results suggest that there is no thematic change from before to after the reunification in East German chemistry dissertations. For West German chemistry dissertations, surprisingly, there is a negative change. This means that West German dissertations became more dissimilar while East ones didn't.

For economics and business administration, the results are presented in Table 4. Here, regression results of models 3 and 4 show a large decrease in cosine similarities for topic distributions of dissertation titles written in different parts of Germany before the reunification. Models 5 and 6 of Table 4 present the regression results of topics in economics and business administration before and after reunification in East Germany. In both models we reach significance and a substantial effect of -0.27 and -0.22, respectively. The last approach, which is presented in full model 1 and 2 of Table 4, shows the similarity between East and West after the reunification. The positive interaction term of *diff_part* and *post95* in full model 2 suggests that there is an increasing similarity, and the coefficient sizes of cosine similarity indicate that the effects observed in economics and business administration are of relevant magnitude. This could have been expected, as the discipline underwent

Table 3 Chemistry OLS regression

	Dependent variable:							
	Cosine similarity							
	Full period	Full period	Pre	Pre	East	East	West	West
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
diff_part	−0.004**	−0.009***	−0.004**	−0.007**				
	0.0017	(0.002)	(0.002)	(0.004)				
post95	−0.017**	−0.020***			0.009	−0.002	−0.018***	−0.020***
	0.009	(0.001)			(0.008)	(0.010)	(0.001)	(0.001)
*diff_part*post95*	0.013***	0.002						
	(0.002)	(0.002)						
Sameuni		0.098***		0.092***		0.119***		0.096***
		(0.002)		(0.004)		(0.011)		(0.003)
Constant	0.217***	0.266***	0.217***	0.182***	0.237***	0.228***	0.217***	0.260***
	(0.001)	(0.012)	(0.001)	(0.030)	(0.007)	(0.024)	(0.001)	(0.013)
Uni dummies	No	Yes	No	Yes	No	Yes	No	Yes
Observations	148,647	148,647	47,329	47,329	2,029	2,029	116,696	116,696
R^2	0.002	0.029	0.0001	0.041	0.001	0.080	0.003	0.029
Adjusted R^2	0.002	0.028	0.0001	0.039	0.0001	0.066	0.003	0.028

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 4 Economics and business administration OLS regression

	Dependent variable:							
	Cosine Similarity							
	Full period	Full period	Pre	Pre	East	East	West	West
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
diff_part	−0.169***	−0.187***	−0.169***	−0.202***				
	(0.002)	(0.003)	(0.003)	(0.003)				
post95	−0.105***	−0.069***			−0.272***	−0.222***	−0.042***	−0.041***
	(0.002)	(0.001)			(0.007)	(0.009)	(0.002)	(0.002)
*diff_part*post95*	0.147***	0.129***						
	(0.003)	(0.004)						
Sameuni		0.086***		0.086***		0.105***		0.074***
		(0.004)		(0.008)		(0.007)		(0.005)
Constant	0.3623***	0.358***	0.362***	0.383***	0.520***	0.546***	0.299***	0.323***
	(0.002)	(0.009)	(0.002)	(0.017)	(0.0004)	(0.042)	(0.002)	(0.009)
Uni dummies	No	Yes	No	Yes	No	Yes	No	Yes
Observations	62,586	62,586	12,784	12,784	3,104	3,104	40,875	40,875
R^2	0.069	0.123	0.202	0.416	0.332	0.419	0.008	0.037
Adjusted R^2	0.069	0.121	0.202	0.409	0.332	0.409	0.008	0.034

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

a drastic reorientation after German reunification. In chemistry, the statistically significant effects are much smaller. Chemistry may serve as an example of how even minor changes can be detected by our approach.

Discussion and conclusion

In this paper we have shown how scientists' research problem choices can be detected with a machine learning approach. For this purpose, we investigated the thematic change after an unexpected political transition. We used dissertation titles in the disciplines of economics and business administration and chemistry before and after German reunification in East and West Germany. We applied structural topic modelling combined with cosine similarity-based regression. We found differences between the two parts of Germany in both disciplines before the reunification. These differences decrease somewhat after the reunification. Our results suggest that East German dissertation title topics in the field of economics and business administration are significantly more different before reunification than thereafter.

The substantial differences in economics and business administration before the reunification are likely to be related to politics, and are in accordance with the historical circumstances that we described in the chapter "Historical background". Economics and business administration as a discipline was extremely important in the ideological framework of the GDR. The research of economists and business administrators, more so than in other disciplines, had to therefore be vetted and brought in line with socialist ideology. Topics related to capitalism, which were researched in western countries like West Germany, were therefore de facto impossible to research in the GDR.

Regarding our findings after the reunification, we again refer to chapter "Historical background". As described, massive personnel replacement, as well institutional redirection, took place in East German economics and business administration after the reunification. The free chairs were predominately filled with West Germans economists and business administration scholars (anecdotal evidence). Consequently, the dissertation topics picked by these new scientists would have been very different from the topics of the dismissed East German scientists and their predecessors. However, within the long time span we investigate after the reunification (15 years), other factors could have also led to declining differences within economics and business administration.

One potential explanation for the small differences in chemistry research topics between East and West Germany before reunification could be the industrial relevance of the discipline, which motivated the GDR government to directly and indirectly interfere with the topic choices of East German scientists. The prime example of direct influence was the official yearly plans for science and technology, which forced chemistry to meet the industry demands of East Germany (Gruhn and Lauterbach 1977). The economic and societal restrictions in the GDR also had an influence on topic choices and therefore on the topics and results that we can observe. Collaboration, for instance, was for East German scientists almost exclusively possible with researchers from other socialist countries (Weingart et al. 1991). This prevented thematic spread that could have resulted from collaboration with West German colleagues. The different characteristics of economic uncertainty of the GDR in comparison to West Germany may also have had an indirect impact on scientific topic choices. The academic field in the GDR was, for instance, fully employed at any point in time, albeit with a considerable hidden unemployment rate, as it was socialist state doctrine

to employ everyone (Gutmann 1979). Picking risky research problems was possibly not associated with risky labour market outcomes for East German chemists and scientists in general. Nevertheless, the choice of risky topics was contradicted by the aforementioned science and technology plans, which forced East German researchers to pick applied topics that met industry demands. Lastly, the small differences between East and West Germany in chemistry before the reunification could also be attributed to West German peculiarities.

The method presented and developed in this paper—structural topic modelling and a cosine similarity-based regression approach—are its main contributions, and aimed to detect differences in research topics of East and West German scientists before and after German reunification. As demonstrated, this turned out to be successful; our trained model detects reasonable differences in a set of unseen titles. The inclusion of dissertation level variables, like affiliation to single universities or dissertation year information, in training a topic model can be considered as a decisive advantage of our approach. Research problem choice is dependent on various factors, such as regional and temporal origin of the dissertation. In the topic modelling process, which tries to reconstruct the data-generating process behind the dissertation title, these factors should therefore not be considered constant across all dissertations in the training set (as done by the LDA topic model algorithm).

The incorporation of paired cosine similarities into a regression approach has, to our knowledge, never been used before and is therefore a methodical innovation of our paper. The regression framework presented in this paper provides not only an easily interpretable aggregation of the cosine similarities, but also a way to test hypothesis. In this sense, other contexts and datasets in scientometric research could be addressed by our approach, which may deliver new perspectives on thematic and, therefore, scientific change in general.

From the visual inspection of the most probable words in economics and business administration, we conclude that our model was able to discover meaningful relationships. The usage of short documents—in our case, dissertation titles—did not turn out to be a problem. In the application to the unseen documents, which were the basis for validation, our algorithm worked well. As topic modelling does not aim to label the detected topics, we can sometimes only guess what the found differences and their underlying topics most likely refer to. This is a major disadvantage of any sort of topic modelling. The foundation of this problem arises from language as a dynamic, complex and strongly context-related semantic system. Topic models can only find the relations in this system, but not understand and label them accordingly. It is therefore beyond the scope of our paper to find reasonable labels for topics we detected.

The linkage of our data to measures of scientific success and impact could provide interesting further research questions. The topical choices that are associated with academic rewards for PhD students, for example, could be investigated. Also, our method could be promising for the investigation of other types of documents; abstracts and scientific articles may contain document-level information which could shift topic proportions in the same way as the variables in our paper. Because of increased document length in these cases, the topic model algorithms would exponentially increase calculation time, but gain statistical properties and topic quality. Therefore, our method of structural topic modelling combined with a cosine similarity-based regression framework offers potential, generally, for applications in scientometrics and higher education research.

Acknowledgements Open Access funding provided by Projekt DEAL.

Availability of data and materials Rehs (2020a). Dataset: A structural topic model approach to scientific re-orientation of economics and chemistry after German reunification. [chemistry_raw_data.csv; economics_raw_data.csv]. Retrieved from 10.5281/zenodo.3874921.

Rehs (2020b). Dataset: A structural topic model approach to scientific re-orientation of economics and chemistry after German reunification. [cosine_distances_economics.csv; cosine_distances_chemistry.csv]. Retrieved from 10.5281/zenodo.3874921.

Appendix: Top 4 words highest β probability

Topic	Economics and business administration	Chemistry
1	'and,'the','portfolio','development','model'	'radikal','selektiv','alk','alkohol','additions'
2	'integration','kost','internationalisier','beruf', 'option'	'kohlenwasserstoff','konzept','oxidativ','methan ','methanol'
3	'prozess','wissenschaft_technisch','technisch_ fortschritt','rationalisier','wissenschaft_tech- nisch_fortschritt'	'funktionalisiert','baustein','verbruckt','chrom' ,'gold'
4	'schwerpunkt','konzentration','steuerpolit','ent wurf','steuerreform'	'oberflach','adsorption','wasserstoff','wechselw irk','ftir'
5	'bereich','energie','rationell','brd','konsumgut'	'the','complex','with','based','catalyst'
6	'einsatz','erfolgsfaktor','internet','onlin','medi'	'elektron','clust','zust','fest','spin'
7	'wandel','osterreich','qualitativ','organisator' ,'inner'	'olefin','einsatz','stabilisier','homog','ylid'
8	'mittelstand','intern','berat','modell','unterneh mensberat'	'basis','vorstuf','hinblick','para_phenyl','poly_ para'
9	'industri','effekt','branch','preis','west'	'platin','komplexbild','cis','stabilitat','phenyl'
10	'problem','sozialist','beding','volks','aufgab'	'stereoselektiv','enantioselektiv','enantiomerenre in','aminosaur','diastereoselektiv'
11	'ddr','kombinat','leitung','sozialismus','nutz ung'	'dynam','synthet','natur','membran','relaxation'
12	'strategi','ziel','orientiert','unternehmenskris' ,'bewalt'	'hoh','flussigkristall','niedermolekular','mesog' ,'nemat'
13	'industriell','aspekt','determinant','organisator', 'entwicklungsland'	'ubergangsmetall','phosphan','redox','cyclopenta dienyl','fragment'
14	'extern','prognos','bau','qualitatssicher','steu erungs'	'for','element','paramet','gallium_ indium','aluminium_gallium_indium'
15	'okonomi','geld','sozial','kritik','okologi'	'flussigkristallin','amphiphil','monom','phasenve rhalt','grenzflach'
16	'produkt','innovativ','finanz','backed_ securiti','rentenversicher'	'pfeil_recht','eis','typs','eta','mangan'
17	'polit','institution','quality','histor','islam'	'bzw','alkaloid','strukturaufklar','pyrrol','cycl isier'

Topic	Economics and business administration	Chemistry
18	'hintergrund','funktion','okonometr','verande r','jung'	'rhodium','carb','iridium','alkin','zweikern'
19	'unt','logist','bes_beruck','textil','sektor'	'elektro','uberbruckt','chromatographi','komple xier','sigma'
20	'beurteil','anhand','usa','wettbewerbspolit','b etracht'	'unt','phosphor','dimethylamino','phosphoran', 'nitro'
21	'forder','mittl','auswahl','massnahm','qualitats'	'verhalt','cycloaddition','dien','abfang','triazin'
22	'okolog','nachhalt','sozial','global','umwelt'	'diel_ald','neutral','hetero_diel','hetero_diel_ ald','selektivitat'
23	'rahm','komplex','nutzung','weiter','beitr'	'strukturell','kupf','praparativ','oxid','aspekt'
24	'basis','fuzzy','marktforsch','neuronal_ netz','einkaufsstattenwahl'	'situ','111′','adsorption','surfac','non'
25	'steu','ermittl','kapitalgesellschaft','finanzier',' grenzuberschreit'	'aromat','alkyl','phenol','aliphat','chloriert'
26	'forschung','betriebs','kost','sicher','kontroll'	'ausgewahlt','vergleich','ungesattigt','gegenub', 'substrat'
27	'bank','roll','kulturell','kund','unternehmens kultur'	'methyl','total','hydroxy','zugang','est'
28	'optimier','kommunikation','softwar','mittel', 'losung'	'stickstoff','phosphor','schwefel','kohlenstoff','s auerstoff'
29	'verbesser','qualitat','verwend','neuronal_ netz','kunstlich_neuronal'	'aufbau','messung','druck','temperatur','mpa'
30	'technisch','informations','rechnergestutzt','dar stell','betriebs'	'iii','oxo','tris','vanadium','chlor'
31	'struktur','japan','gesellschaft','dimension','al ternativ'	'analyt','modifiziert','hplc','biolog','trennung'
32	'information','integration','verteilt','heterog','w ertorientiert'	'thermisch','photochem','omega','isomerisier',' lamda'
33	'dynam','optimal','linear','investition','finanz plan'	'stereo','verwandt','tran','cis','grundlag'
34	'bezieh','zusammenarbeit','industrieland','nor d','kapitalist'	'modell','einfach','quantenchem','porphyrin','c hinon'
35	'schweiz','wandel','natur','welt','option'	'metall','modell','chelat','rhenium','haltig'
36	'uber','zentral','regel','gesetz','plan'	'dihydro','eta','kenntnis','lambda','sigma'
37	'sicht','institutionen','betracht','schweizer','we ttbewerbsfah'	'naturstoff','transformation','allyl_ substitution','beitr','biolog_aktiv'
38	'entscheid','computergestutzt','raum','werbun g','grenz'	'verwend','amorph','loslich','kohlenhydrat','m ateriali'
39	'risiko','risik','privat','ventur_capital','banking'	'peptid','konformation','modifizier','zyklisch','r acematspalt'
40	'controlling','umsetz','organisations','operativ ','effizient'	'las','ungewohn','immobilisiert','matrix','stud ium'
41	'wirkung','tourismus','mark','stadt','verhaltens wissenschaft'	'delta','trag','tetra','symmetr','kristallisation'
42	'markt','industrie','ausgewahlt','fallstudi','tran sformation'	'ubergangs','titan','rontgenstrukturanalys','koordi nation','semiempir'
43	'hilf','landlich','gebiet','technisch','kennzahl'	'hilf','infrarot','lichtinduziert','zeitaufgelost','b erechn'
44	'perspektiv','neu','system','regionalpolit','refo rmvorschlag'	'gas','massenspektrometr','nachweis','elementa r','partiell'

Topic	Economics and business administration	Chemistry
45	‘automobilindustri’,‘netzwerk’,‘kooperation’,‘virtuell’,‘interkulturell’	‘poly’,‘styrol’,‘polystyrol’,‘initiator’,‘copolymerisation’
46	‘servic’,‘financial’,‘engineering’,‘performanc’,‘integration’	‘analoga’,‘festphasen’,‘aufbau’,‘kombinator’,‘strategi’
47	‘arbeit’,‘einflussfaktor’,‘diagnos’,‘grundsatz_ordnungsmass’,‘grundsatz_ordnungs-mass_bilanzier’	‘molekul’,‘photo’,‘fluoreszenz’,‘raman’,‘induziert’
48	‘aspekt’,‘ergebnis’,‘land’,‘licht’,‘studi’	‘wassrig’,‘gamma’,‘sio2′,‘al2o3′,‘tio2’
49	‘personal’,‘syst’,‘evaluation’,‘fuhrungskraft’,‘fuhrung’	‘bindung’,‘aktivier’,‘funktionalisier’,‘aktiviert’,‘alkylier’
50	‘staatlich’,‘staat’,‘zusammenhang’,‘land’,‘gesellschaft’	‘optisch’,‘magnet’,‘farbstoff’,‘elektr’,‘schicht’
51	‘makro’,‘fundiert’,‘verhalt’,‘erklar’,‘arbeitsmarkt’	‘amin’,‘amino’,‘ring’,‘aryl’,‘substituent’
52	‘alternativ’,‘losung’,‘geldpolit’,‘finanziell’,‘entscheidungs’	‘molekul’,‘theoret’,‘ion’,‘zeolith’,‘umlager’
53	‘produktion’,‘effektivitat’,‘flexibl’,‘fertig’,‘vorbereit’	‘silicium’,‘kristall’,‘silan’,‘sol_gel’,‘silicat’
54	‘innovation’,‘erfolgreich’,‘fallbeispiel’,‘innovations’,‘organisational’	‘ternar’,‘kristall’,‘lithium’,‘alkali’,‘lanthanoid’
55	‘aufbau’,‘praktisch’,‘rahm’,‘unternehmensfuhr’,‘ansatzpunkt’	‘nickel’,‘koordinations’,‘zink’,‘cobalt’,‘silb’
56	‘marketing’,‘national’,‘einzelhandel’,‘determinant’,‘interaktion’	‘katalyt’,‘mono’,‘aufklar’,‘hydrier’,‘umwandl’
57	‘bestimm’,‘simulation’,‘hilf’,‘system’,‘eignung’	‘cyclisch’,‘umsetz’,‘nucleophil’,‘bzw_beziehungsweis’,‘elektrophil’
58	‘prozess’,‘modellier’,‘unternehmens’,‘dynam’,‘mittel’	‘grupp’,‘element’,‘amid’,‘nebengrupp’,‘moglich’
59	‘regional’,‘studi’,‘rechnungsleg’,‘ifr’,‘bilanzier’	‘oxidation’,‘mechanismus’,‘ruthenium’,‘reduktion’,‘gegenwart’
60	‘gross’,‘markt’,‘operationalisier’,‘bereitstell’,‘erfolgswirk’	‘katalysator’,‘palladium’,‘polymerisation’,‘eth’,‘katalys’
61	‘integriert’,‘unterstutz’,‘technologi’,‘ganzheit’,‘prozessorientiert’	‘optisch_aktiv’,‘pro’,‘baustein’,‘alkohol’,‘katalys’
62	‘einfuhr’,‘business’,‘gruppenarbeit’,‘organisator’,‘produktionsbereich’	‘analys’,‘gebund’,‘optimier’,‘spektr’,‘gaschromatograph’
63	‘dienstleist’,‘relevanz’,‘beschaff’,‘zusammenarbeit’,‘kooperation’	‘wass’,‘syst’,‘thermodynam’,‘mischung’,‘kritisch’
64	‘steuer’,‘konzeptionell’,‘handel’,‘dezentral’,‘handels’	‘addition’,‘versuch’,‘lithium’,‘aldehyd’,‘ungesattigt_ungesattigt’
65	‘rahmenbeding’,‘institutionell’,‘kommunal’,‘bundesland’,‘medizin’	‘wechselwirk’,‘festkorp’,‘hilf’,‘schwach’,‘saur’
66	‘basis’,‘verfahr’,‘entscheidungsorientiert’,‘krankenhaus’,‘energieversorgungs’	‘oligo’,‘sensor’,‘dendrim’,‘kunstlich’,‘potentiell’
67	‘bedeut’,‘entwicklungsland’,‘implikation’,‘wirtschaftspolit’,‘gegenwart’	‘linear’,‘anelliert’,‘thioph’,‘oligom’,‘nichtlinear_optisch’
68	‘region’,‘untersucht’,‘china’,‘strukturwandel’,‘berlin’	‘molekular’,‘anion’,‘experimentell’,‘supramolekular’,‘modellier’
69	‘einfluss’,‘zeitverwend’,‘ausgewahlt’,‘grenz’,‘faktor’	‘protein’,‘dna’,‘wechselwirk’,‘enzymat’,‘inhibitor’

Topic	Economics and business administration	Chemistry
70		'via','diel_ald','lewis_ saur','steroid','intramolekular_diel'
71		'massenspektrometri','kopplung','icp','prob','d irekt'
72		'struktur','organo','rontgenograph','schwingung s','alkali'
73		'typ','mechanism','extraktion','chemistry','imi dazolin'
74		'donor','biphenyl','wirt_gast','helical','axial'
75		'bildung','verfahr','effekt','zerfall','berucksicht'
76		'verschied','kation','voraussetz','carbonyl','ind uziert'

References

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., & Zhu, M. (2013). A practical algorithm for topic modelling with provable guarantees. In S. Dasgupta, & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning, Volume 28 of proceedings of machine learning research* (pp. 280–288). Atlanta: PLMR.

Belitz-Demiriz, H., Voigt, D., & Gries, S. (1990). *Die Sozialstruktur der promovierten Intelligenz in der DDR und in der Bundesrepublik Deutschland 1950–1982*. Brockmeyer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112*(518), 859–877.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(1), 993–1022.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215–2222.

Basic Law for the Federal Republic of Germany in the revised version published in the Federal Law Gazette Part III, classification number 100-1, as last amended by Article 1 of the Act of 28 March 2019 (Federal Law Gazette I p. 404).

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Landham: Scarecrow Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Deutsche Demokratische Republik. (1968). Verordnung über die akademischen Grade vom 06.11.1968. Berlin.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3), 357–367.

Gruhn, W., & Lauterbach, G. (1977), *Die Organisation der Forschung in der DDR* (127–213). In Institut für Gesellschaft und Wissenschaft, Erlangen: Campus Verlag.

Guenther, K.-H. (1989). *Das Bildungswesen der Deutschen Demokratischen Republik: Gemeinschaftsarbeit der Akademie der Pädagogischen Wissenschaften*. Berlin: Volk und Wissen.

Gutmann, G. (1979). Employment problems under socialism. *Intereconomics, 14*(2), 96–100.

Hahn, E. (2009). Publikationsverhalten in der Chemie. Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen: Beiträge zur Beurteilung von Forschungsleistungen, 104–107. Retrieved from: https://www.humboldt-foundation.de/pls/web/docs/F13905/12_disk_papier_publikationsverhalten2_kompr.pdf.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics, 44*(2), 193–215.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572.
Jinha, A. E. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing, 23*(3), 258–263.
Kolloch. (2001). Abwicklung und Neuaufbau der wirtschaftswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin zwischen November 1989 und Dezember 1993. In F. Theißen, (Ed.), *Zwischen Plan und Pleite. Erlebnisberichte aus der Arbeitswelt der DDR*. Bühlau Verlag.
Landhuis, E. (2016). Scientific literature: Information overload. *Nature, 535*(7612), 457–458.
Larsen, P., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575–603.
Leininger, W. (2009). *Publikationsverhalten in den Wirtschaftswissenschaften*. Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen: Beiträge zur Beurteilung von Forschungsleistungen, 39–40. Retrieved from: https://www.humboldt-foundation.de/pls/web/docs/F13905/12_disk_papier_publikationsverhalten2_kompr.pdf.
Mann, R. (1979). Internationale Wissenschaftsbeziehungen. In Institut für Gesellschaft und Wissenschaft (Ed.), *Das Wissenschaftssystem in der DDR*. Berlin: Campus Verlag.
Meske, W. (2004). *From system transformation to European integration. Science and technology in Central and Eastern Europe at the beginning of the 21st century*. Münster: LIT Verlag.
Morichika, N., & Shibayama, S. (2016). Use of dissertation data in science policy research. *Scientometrics, 108*(1), 221–241.
Ooms, J. (2018). cld3: Google's Compact Language Detector 3. Retrieved Feburary 7, 2019, from https://cran.r-project.org/web/packages/cld3/cld3.pdf, version 1.1.0.
Peisert, H., & Framhein, G. (1994). *Das Hochschulsystem in der Bundesrepublik Deutschland: Struktur und Entwicklungstendenzen*. Bad Honnef: Bock.
Ramage, D., Dumais, S. T., & Liebling, D. J. (2010). Characterizing microblogs with topic models. In *Proceedings of the fourth international AAAI conference on weblogs and social media* (pp. 130–137). Menlo Park: The AAAI Press
Rehs. (2020a). *Dataset: A structural topic model approach to scientific re-orientation of economics and chemistry after German reunification*. [chemistry_raw_data.csv; economics_raw_data.csv]. Retrieved from https://doi.org/10.5281/zenodo.3895119
Rehs. (2020b). *Dataset: A structural topic model approach to scientific re-orientation of economics and chemistry after German reunification*. [cosine_distances_economics.csv; cosine_distances_chemistry.csv]. Retrieved from https://doi.org/10.5281/zenodo.3895119
Roberts, M. E., Stewart, B. M., & Tingley, D. (2014a). stm: R package for structural topic models. *Journal of Statistical Software, 10*(2), 1–40.
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., et al. (2014b). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.
Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association, 111*(515), 988–1003.
Volkskammer der DDR. (1976). Verfassung der Deutschen Demokratischen Republik vom 6. April 1968 in der Fassung des Gesetzes zur Ergänzung und Änderung der Verfassung der Deutschen Demokratischen Republik vom 7. Oktober 1974. Berlin: Staatsverlag der Deutschen Demokratischen Republik.
Weingart, P., Strate, J., & Winterhager, M. (1991). *Bibliometrisches Profil der DDR*. Bericht an den Stifterverband für die Deutsche Wissenschaft und den Wissenschaftsrat: Universitätsschwerpunkt Wissenschaftsforschung, University of Bielefeld.
Wollgast, S. (2001). *Zur Geschichte des Promotionswesens in Deutschland*. Bergisch Gladbach: Dr. Frank Graetz Verlag.