UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Ott-Kaarel Martens

Evolution of Topics in the Psychology Domain

Bachelor's Thesis (9 ECTS)

Supervisor: Eduard Barbu, PhD

Tartu 2020

Evolution of Topics in the Psychology Domain

Abstract: Topic modeling is a set of statistical methods for modeling collections of discrete data such as text corpora. It is used as a text-mining tool to discover the hidden semantic structures in a text body. Latent Dirichlet Allocation, a particular method for topic modeling is a generative probabilistic model that models texts as a mixture of underlying topics. In this thesis, Latent Dirichlet Allocation is used on a large corpus of texts from the domain of psychology. A model with 100 topics is generated, and the resulting topics are labeled. The occurrence of the topics is analysed over a time span of 40 years. The resulting model could be used for further psychology domain specific research.

Keywords:
Topic models, psychology, Latent Dirichlet Allocation, semantic models

CERCS: P175 Informatics, systems theory

Teemade Evolutsioon Psühholoogias

Lühikokkuvõte: Teemade mudeldamine on kogum statistilisi meetodeid diskreetsete andmekogumite (näiteks tekstikorpuste) modelleerimiseks. Sellist mudeldamist kasutatakse tekstikaeves teksti varjatud semantiliste struktuuride leidmiseks. Varjatud Dirichlet Jaotamine (ing *Latent Dirichlet Allocation*) on generatiivne teemamudeldamise meetod, mis modelleerib teksti kui mikstuuri teemadest. Käesolevas töös rakendatakse Varjatud Dirichlet Jaotamist suurele psüholoogiaalaste tekstide korpusele. Töös genereeritakse 100 teemaga teemamudel, ning saadud teemad pealkirjastatakse. Teemade esinemist analüüsitakse 40-aastase perioodi vältel. Saadud mudelit on võimalik rakendata järgnevates uurimustes.

Võtmesõnad:
Teemamudelid, psühholoogia, semantilised mudelid

CERCS: P175 Informaatika, süsteemiteooria

Contents

1 Introduction **5**

1.1 The subject and goal of the thesis 5

2 Theoretical background **7**

2.1 Latent Dirichlet Allocation 7

2.1.1 The generative process of LDA 8

3 Assembling the corpus **10**

3.1 Criteria for corpus resource selection 10

3.2 Assessed candidates for corpus resources 11

3.2.1 APA PsycArticles® database 11

3.2.2 Annual Reviews 11

3.2.3 PubMed 11

3.3 Generating the corpus 12

4 Preprocessing **13**

4.1 Used preprocessing tools and libraries 13

4.1.1 Gensim 13

4.1.2 NLTK 14

4.1.3 spaCy 14

4.2 Preprocessing steps 14

4.2.1 Tokenization and stopword elimination 14

4.2.2 Multiword expression identification 15

4.2.3 Lemmatization, PoS-Tagging and Named Entity Recognition . . 15

4.2.4 Dictionary and bag-of-words representation 16

5 Generating the topic model **17**

5.1 Introduction to MALLET 17

5.2 Comparison of Gensim and MALLET LDA implementations 17

5.3 Topic count 18

6 Analysis of the topics **19**

6.1 Labeling the topics 20

6.1.1 Labeling with WordNet hypernyms 20

6.1.2 Human labeling 21

6.2 Topic popularities over time 21

6.2.1 Biggest increases in topic occurrence over the timespan 22

6.2.2 Biggest decreases in the topic occurrence over the timespan . . 23

6.2.3 Most variance in topic occurrence 24

6.2.4 Most stable topic occurrences (least variance) 24

7 Conclusion **26**

A Topics **29**

B Source code **50**

C Licence **51**

1 Introduction

Since the inception of digital storage of information, the volume of the stored information has been rapidly increasing. Digital storage has become the norm of how information is preserved. Although this medium has significant advantages in terms of cost, flexibility and scalability of preserving information, the rapid growth in the volume gives rise to many problems in terms of handling the information. "As our collective knowledge continues to be digitized and stored, ... , it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search, and understand these vast amounts of information."[5]

To illustrate this domain of problems, we might consider the manner by which we search for information online. For a more concrete example, we can examine how an online publication, such as a newspaper, can be interacted with. The primary medium for rapidly conveying the contents of a newspaper article is the title. Some publications might also provide keywords by which we can retrieve relevant articles. However, there is a lack of tools and functionality for discovering the relations between articles, such as the "common denominator" in terms of the themes of the articles. The labour of figuring out how different documents fit into a larger context and relate to one another is offloaded to the processing power of the consumer - the reader.

This set of problems has been addressed by researchers in fields such as natural language processing and probabilistic machine learning. In the last few decades, novel methods have arised for solving these problems, one of them being topic modeling.

The process of topic modeling can be described as employing statistical methods that "analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time."[5] Purposes for this kind of processing are manifold: one might want to annotate a collection of documents according to their contents, discover outlying works, emerging themes, or other changes in trends. The topic model can also be used for recommending thematically similar content to an item. It has been used to group behavioural user personas[20], summarize meetings[12] and much more. Even modelling of chemical compounds has been proven to yield utility with this method.[21]

1.1 The subject and goal of the thesis

The subject of this thesis is to apply topic modeling to investigate the topics in the domain of psychology. Psychology is a field with a rich history. There exist a plethora of subdisciplines inside this domain, and the field has undergone many paradigm shifts. The forefront of the evolution of this field, as is the case for many, if not all practices,

is reflected in the content of its academic literature. Employing topic modeling to analyze this literature provides insights to the evolution of the contents of the field. Previously, A. Bitermann and A. Fischer have identified hot topics - themes with the highest increasing trend in a period of time - in psychological literature, using topic modeling. In this research, the authors concluded that topic modeling is a feasible method for an exploratory analysis of topics in psychological publications, stating that "The identification of specific topics in a large corpus of publications offers new possibilities of exploring research beyond predefined classifications."[4]

In this thesis, topic modeling, more precisely Latent Dirichlet Allocation, is applied to a large corpus of abstracts from academic literature in the domain of psychology. A LDA topic model is generated from the corpus and the resulting topics are labeled and analyzed.

The thesis is organized as follows. In chapter 2, an overview is given of topic modeling, focusing primarily on Latent Dirichlet Allocation. In chapter 3, the assembly of a corpus is discussed. Chapter 4 gives an overview of the preprocessing of the data. Chapter 5 focuses on the generation of the topic model. In chapter 6, the resulting model and topics are analysed.

2 Theoretical background

The overarching goal of modeling large collections of discrete data, such as text documents, is to find short descriptions of the individual items in the collection. The descriptions should reduce the dimensionality of the original items while preserving information about the semantic contents of the items and the statistical relations between them. This enables efficient processing of the collection while preserving the ability to do operations such as classification, novelty detection, summarization and so forth.[7]

One of the earlier notable solutions for this problem is the so-called term frequency-inverse document frequency (tf-idf) method, which weighs the occurrence of terms in a document offset by the terms occurrence in the whole corpus. While being a useful method for applications such as ranking keyword search results, it does not result in a major dimensionality reduction and provides little use for modeling the statistical structure inside the documents, or the relations between them.[7]

A useful step forward, a method called latent semantic indexing (LSI) was put forth by information retrieval researchers in 1990. LSI uses Singular Value Decomposition on the tf-idf matrix to identify a linear subspace in the matrix that captures the most variance between the documents. This results in a significant dimensionality reduction, while preserving the majority of the information about variability.[8]

The probabilistic latent semantic indexing, or pLSI, was developed by Hoffmann in 1999 as an alternative to LSI. pLSI is a generative method that models each word in a document as a sample from a mixture model. The components of the mixture model are multinomial random variables that can be interpreted as topics.[11] According to this method, each word in a document is generated from a single topic and a document can exhibit multiple topics. This means that a document can be represented by a probability distribution over the topics. pLSI has many advantages for modeling collections of discrete data over the formerly mentioned ones, most notably the reduction in dimensionality, and the ability for a document to exhibit multiple topics, which resembles how people generally interpret topics. However, pLSi is incomplete in the sense that it provides no generative model for the document topic proportions. This means that the number of variables grows linearly with the increase in corpus size, and that there is no way to assign topic probabilities to a document outside the corpus.[7]

2.1 Latent Dirichlet Allocation

To overcome these limitations, Latent Dirichlet Allocation (LDA) was described by David M. Blei, Andrew Y. Ng and Michael I. Jordan in 2003. LDA aims to model all hierarchical layers of the collection of documents. The resulting model is a three-level

hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. In Latent Dirichlet Allocation, a topic is formally defined as a distribution over a fixed vocabulary. A document can exhibit multiple topics, because a topic is sampled repeatedly for each word inside a document.[7]

2.1.1 The generative process of LDA

It is useful to describe LDA by its generative process - "the imaginary random process by which the model assumes the documents arose."[5] This process happens as follows. Internally, LDA utilises two Dirichlet distributions (which are commonly used prior distributions in Bayesian statistics), one for associating documents to topics and the other for topics to words. For generating a document, a topic distribution is picked from the first Dirichlet distribution. Then, for each of the words of the document, a topic is picked from the selected distribution. As a third step, a topic-to-word multinomial distribution is picked from the second Dirichlet distribution for each of the topics. As a final step, a specific word is picked for each word in the document (the topic of which is already decided at this point) from the corresponding multinomial distribution of words.

This process is described as a graphical model on Figure 1. More formally, the generation of a document **w** in a corpus by this imaginary process happens by the following steps:

1. Choose $N \sim$ Poisson(ξ)
2. Choose $\theta \sim$ Dir(α)
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim$ Multinomial(θ)
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n

Where:

- N, a document, has a sequence of words denoted by **w**=(w_1, w_2, ..., w_n)
- θ is a k-dimensional Dirichlet random variable
- α is the parameter of the Dirichlet prior on the per-document topic distribution
- β is s the parameter of the Dirichlet prior on the per-topic word distribution[7]

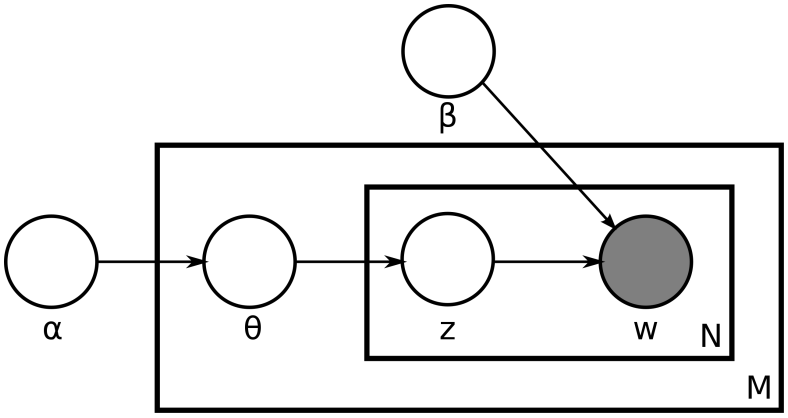


Figure 1. Plate notation representing the LDA model. The rectangles represent replication. The outer rectangle represents a document, and the inner rectangle represents the repeated choice of topics and words from the topic, replicated N times (the amount of words in a document)[7]

The process of generating a LDA model from a corpus is done by posterior inference. Although the posterior distribution cannot be inferred exactly, different approximate inference algorithms can be utilised for the inference, such as Laplace approximation, variational approximation, and Markov chain Monte Carlo.[13] Commonly, Gibbs sampling is utilised for sampling-based approaches.[5]

There have been many extensions of LDA developed by subsequent research, such as Pachinko Allocation, which in addition models the relations between the topics themselves;[14] and Dynamic Topic Modelling, which also models the change of topics over time.[6]

3 Assembling the corpus

For the task of assembling the corpus for this kind of modelling, the primary resources evaluated for assembling the corpus were databases that aggregate many academic publications.

3.1 Criteria for corpus resource selection

In order to add a temporal dimension to the resulting model, the corpus should cover as extensive of a timespan as possible. A collection spanning multiple decades likely provides a wide overview of the field which is suitable for this kind of modelling approach. This was one of the main criteria for the corpus selection.

The amount of sub-resources (amount of journals and publications) and the total volume of individual documents in the collection is a metric that is highly relevant, since a high sample size provides a more accurate model. A collection that houses many sub-resources is likely to have a more uniform coverage of the field, since the individual publications might be highly specialized to a certain sub-domain. To capture the topic changes for each year (as done in Chapter 6) it is especially important that each year has a high volume of items in order to minimize the distortion.

Another metric for assessing the suitability of a resource is the relevance of its contents. For a resource that assembles multiple journals, the Impact Factor (a metric that measures the relevance of the publication by the amount of citations the publication receives[3]) of the individual journals might indicate a good quality resource. However, a low impact factor does not imply that the publication is not representative of the domain (and vice versa), and thus should not be heavily discriminated against, although it does indicate lower contribution to the domain overall.

A secondary criterion that was assessed but also not primarily discriminated against is the ease of retrieving the contents of the collection. Some collections offer dedicated file transfer protocol solutions for retrieving the files, which is convenient for handling large volumes of data. For the resources assessed for this thesis, the relative ease of accessing the contents did not vary significantly.

In the context of this thesis, it was also a criterion that the resource should provide unpaid access to the contents.

3.2 Assessed candidates for corpus resources

For this thesis, multiple resources were evaluated by the aforementioned criteria for assembling the corpus. Among which, the ones evaluated more suitable will be briefly described here.

3.2.1 APA PsycArticles® database

The American Psychology Association hosts the PsycArticles® database, that stores all resources published by APA and affiliated journals.[18] The notable positive qualities of this resource are the high impact factors of the contained publications and the long-spanning history of the entries. The negative qualities are its near exclusivity to american publications, which likely results in an offset between the themes of corpus' contents and those of the psychology field overall, and the fact that it is a paid resource.

3.2.2 Annual Reviews

The Annual Reviews is a publishing organization "dedicated to synthesizing and integrating knowledge for the progress of science and the benefit of society."[2] Among its publications is the "Annual Review of Psychology", which has been publishing major advances in psychology since 1950.[1] The mission of this publication aligns nicely with the goal of this thesis, and the journal does have a high impact factor. However, this resource consists only of a single publication, and the overall volume of the contents is too small for broad-based analysis by methods such as topic modelling.

3.2.3 PubMed

PubMed is a search engine maintained by the United States National Library of Medicine and National Institutes of Health, that contains over 30 million article citations with abstracts. It's primary data source is the MEDLINE database, which includes literature published from 1966 to present from more than 5000 journals. PubMed citations and abstracts include the fields of biomedicine and health, and cover portions of the life sciences, behavioral sciences, etc.[19]

This resource fits most of the criteria described above very well. A key distinction about this resource is that the full text bodies of the individual articles are not present, only abstracts are accessible.

3.3 Generating the corpus

After consideration, PubMed was evaluated as the resource best fit for the corpus generation. The overall volume and coverage of journals in this resource is excellent compared to alternatives. While the abstract form in general contains only a fraction of text compared to a full article, it distills the crucial information about an article into a much more compact format. There is evidently a lack of specificity when it comes to the information that an abstract presents compared to a full-text article, but in the context of this resource and the alternatives, the large total volume of texts makes up for the small dimensions of individual texts. This factor also allows more documents to be processed with the same amount of processing power in a given time period. Thus, the tradeoff was deemed justified.

The PubMed resource covers many domains in addition to psychology, but since it features a search engine, a desired subset of the whole resource can be obtained. A keyword search with the keyword "psychology", along with the condition that an abstract is present, was run on the resource, which yielded a result of 1159324 articles. Figure 2 displays the articles retrieved from the resource by publication year.

The PubMed search engine offers a variety of formats the results can be downloaded in. The results were downloaded in an XML-format which includes all the metadata related to the article. The link to the downloadable search results can be seen in Appendix B readme file.

To better prepare the data for subsequent steps, the XML file was parsed and stored in a local relational database (A MySQL database hosted inside a Docker container, see Appendix B for source code) to conveniently access the abstracts' metadata in later steps, such as document ids and publication years. Since the result file was relatively large (14 GB), the file was processed as a stream to avoid having to hold the entire file in working memory. A minor proportion of the processed documents exhibited structural inconsistencies (such as an abstract being missing although explicitly specified during the search query) which required skipping of those documents. After parsing, 1128319 documents were stored in the relational database. The abstracts of the documents were then exported from the database as a csv file for further steps such as preprocessing and model generation.

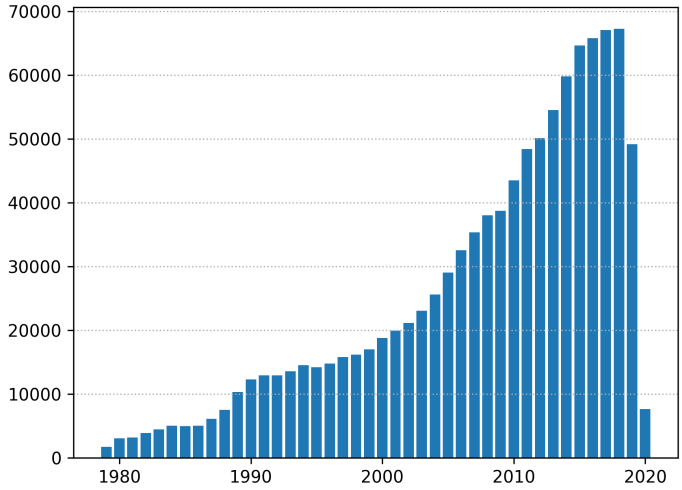


Figure 2. The assembled corpus - article count by year

4 Preprocessing

As a necessary step for model generation, a preprocessing pipeline was built to transform the original texts into an appropriate format. The preprocessing script was written in the python programming language using various natural language processing libraries. The tools used for preprocessing and the sequential steps of the process will be briefly described in this chapter. The produced script is included in the source code (see Appendix B).

4.1 Used preprocessing tools and libraries

4.1.1 Gensim

Gensim is an open-source Python library designed for natural language processing tasks, with the main focus on topic modelling. It includes implementations - usually parallelized ones - of widely used natural language processing algorithms, including Latent Semantic Analysis and Latent Dirichlet Allocation. It also provides wrappers to execute various tools outside the Gensim and Python ecosystem, such as the Mallet LDA wrapper.[10]

4.1.2 NLTK

The NLTK, short for Natural Language Toolkit, is a widely used open-source Python library for natural language processing tasks. It provides a wide array of algorithms and utility functions, as well as interfaces to lexical resources, such as WordNet, and wrappers to other domain-related libraries.[17]

4.1.3 spaCy

spaCy is another open source Python library for language processing, with a strong emphasis on efficient implementation and suitability for usage in production environments. As the name implies, the library is mainly written in Cython. Thinc, the backend of spaCy, provides convolutional neural network implementations for tasks such as part-of-speech tagging and named entity recognition, both of which were utilised in this thesis.[22]

4.2 Preprocessing steps

This chapter describes the various stages of the preprocessing pipeline, which is also outlined in Figure 3.

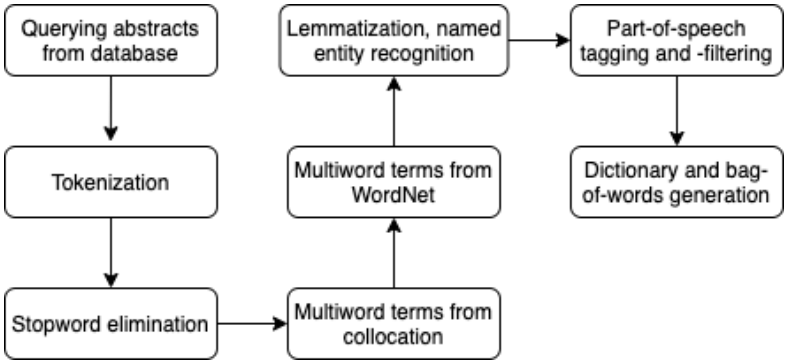


Figure 3. The preprocessing pipeline

4.2.1 Tokenization and stopword elimination

As the first step, the documents in the corpus were split into tokens using the simple_preprocess function from the Gensim library. This step also included transforming the words into lowercase. Simultaneously, the stopwords were eliminated using the

standard set of stopwords from the NLTK package, along with the minimal word length condition of 3 characters.

4.2.2 Multiword expression identification

Multiword expressions are expressions which are made up of at least 2 words, that act as a single unit at some level of linguistic analysis.[16] As an example, expressions such as "long term" and "body mass index" lose their semantic meaning if split into one word tokens and interpreted separately. Hence, it is useful to join the sequential tokens splitted in the previous step that could be considered multiword expressions.

There are different approaches to identifying these expressions in a corpus, two of which were utilised in the process of this thesis. The first approach is relying on statistically significant colocation, meaning that if two or more words occur sequentially in the corpus regularly and the count of these occurrences is proportionally high to the individual occurrences of the words, then this combination of words should probably be identified as a multiword expression. In this thesis, the Phrases model provided by Gensim was utilised for extracting multiword expressions in this manner, in a configuration that could yield a maximum expression length of six words. The second approach that was used for extracting multiword terms was matching all the multiword expressions from the WordNet vocabulary. Since the WordNet features a controlled vocabulary, this kind of matching will result in semantically accurate multiword expressions. At the time of executing this matching, the wordnet corpus featured a total of $\sim$62 000 multiword expressions. The found terms were joined into single tokens with underscores, as is usually done, and examples of the found terms can be seen in the presented topics in Appendix A.

4.2.3 Lemmatization, PoS-Tagging and Named Entity Recognition

As the next step of the preprocessing pipeline, the spaCy library was utilised to perform a variety of sequential steps for each document. The spaCy model configuration used was an English language pipeline with all the supported standard components except dependency parsing and text categorization (these components were disabled to increase efficiency). Each document was loaded into the spaCy model, which results in part-of-speech tagging, named entity recognition and lemmatization of the tokens. The named entities were joined in the same way as the multiword expressions in the previous step. The lemmatized tokens were then filtered based on the selected part-of-speech tags. Multiple configurations were experimented with for this, notably nouns-plus-verbs and nouns-only part-of-speech filtering. The nouns-only configuration, more specifically nouns and proper nouns only, was selected as the final configuration.

4.2.4 Dictionary and bag-of-words representation

Before the model generation, the corpus needs to be converted into more efficient data structures - a dictionary and a bag-of-words representation of the corpus. The dictionary is essentially a mapping between words occurring in the corpus and integers. The bag-of-words corpus is a representation of the corpus where each document is represented as an array of integer pairs (m, n), where m is the id of a word in the dictionary and n the count of this particular word in the document. To transform the corpus into these structures, the corresponding Dictionary module from the Gensim library was used.

5 Generating the topic model

There are multiple published libraries and software projects providing implementations of topic modeling algorithms. The different implementations vary in different dimensions, including efficiency. More sophisticated implementations make use of parallelised computation, efficient internal data structures and other performance improving mechanics. Also, the method and specific parameters for inference differ for various implementations. For this thesis, two different implementations of Latent Dirichlet Allocation were evaluated: the native LDA implementation from the Gensim library and the version from MALLET, a software project which will be briefly described next.

5.1 Introduction to MALLET

MALLET is a Java-based library, designed for statistical natural language processing, that is developed primarily by Andrew McCallum in the University of Massachusetts, and published under an open source license. In addition to other tools, it contains efficient, sampling-based implementations of various topic modeling algorithms, including LDA.[15] MALLET's implementation of LDA is efficient, but requires O(corpus_words) of memory.[9]

5.2 Comparison of Gensim and MALLET LDA implementations

The Gensim library provides a wrapper function for the MALLET implementation of LDA, which was used in this thesis. Hence, both LDA implementations were invoked through the same library. The method signatures for both implementations are roughly similar, requiring parameters for the corpus and the dictionary as well as the number of topics, and optional tweaking of hyperparameters used for the model generation, but the Gensim implementation allows for a bit more flexibility in terms of configurable low-level parameters. Both MALLET and Gensim offer a parallelised implementation of the algorithm, although for Gensim, the LdaMultiCore class needs to be used instead of generic LdaModel to allow parallel processing.

Both of the implementations were run, and the comparison results for the same configuration (100 topics) are as follows. The generation of a model took 13 min and 50 min, for Gensim and MALLET implementations respectively. The Mallet model generation includes some pre- and post-generation steps such as transforming the corpus to a more optimised data structure, and this processing takes extra time, therefore resulting in a longer training time. The coherence for these models, reported by the Gensim CoherenceModel was 0.4072 and 0.5677 respectively.

To decide on which model to proceed with, the coherence values were taken into account, but equally importantly, the coherence of the topics was estimated empirically. The MALLET model by manual investigation seemed to produce more higher-quality topics – the clusters were more similar to how people would cluster words to topics compared to the Gensim model output. Considering these factors, MALLET's implementation was selected for further work.

5.3 Topic count

An important parameter to decide for generating a topic model is the number of topics to be generated. Multiple factors have to be taken into account, such as the scope of the corpus, and the desired specificity of the individual topics. A low topic count can introduce topics composed of themes that do not have close semantic similarity, but too high of a topic count on the other hand introduces fragmentation of topics where it is not desired. The goal is to minimise both of these deviations.

For this thesis, models were generated with 2, 5, 10, 20, 30, 50, 100 and 200 topics. Figure 4 shows the coherences reported by Gensim's coherence model for each of the models, using the c_v coherence metric. The model picked for further analysis was a model with 100 topics. The coherence value was higher than all the models with lower topic counts and only marginally smaller than the 200 topic model coherence. Also, the topics seemed to be of high quality by empirical investigation. 100 topics is also suggested as a generally appropriate number of topics by the authors of this method.[7]

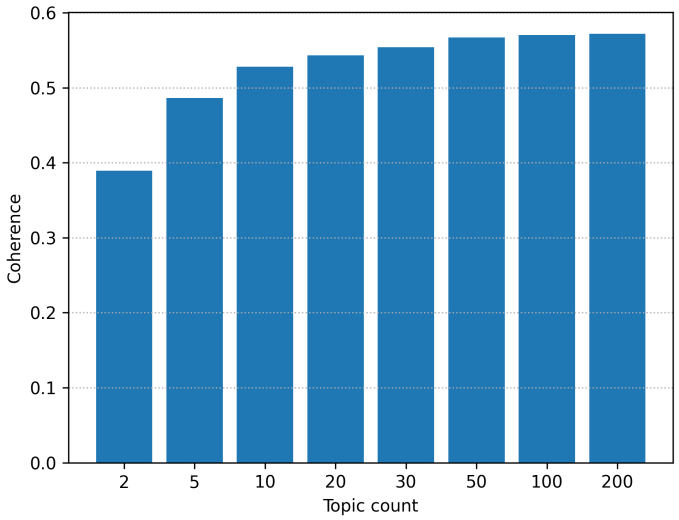


Figure 4. The coherence of n-topic models reported by Gensim's CoherenceModel

6 Analysis of the topics

The topics outputted by the model are words with probabilities belonging to that specific topic. Table 1 shows three sample topics from the final model, with top 10 highest probability words for both of these topics. The probability indicates the percentage of choosing that particular word when choosing a word from that topic, or in other words, how big of a proportion of the topic does that word occupy. We can see that for the first two of these topics (topics 2 and 18) – and the same is true for the majority of the topics – one word has a relatively high probability of being drawn from the topic, and with subsequent words, the probability decreases rapidly.

Based on this, it is also quite easy to make sense of the topic - topic number 2 clearly has words about substances that are prone to being abused, and words about substance abuse in general, such as dependence and addiction. Same is true for topic number 18 - we see that the dominant word in the topic is memory, and the subsequent words are either related to memory or experimentation, so it is reasonable to say that this topic is about research related to memory.

Based on these factors, we could say that these are good quality topics. This is now true for all topics though - there occur some, like topic 16 in Table 1, which do not have words with distinctively high probabilities, and making sense of the semantic coherence of the words is difficult. Fortunately, the number of such topics is a relatively small minority.

Table 1. Three sample topics – top 10 words with the associated probabilities

Topic no. 2		Topic no. 18		Topic no. 16	
word	**prob %**	**word**	**prob %**	**word**	**prob %**
alcohol	18.059	memory	24.5233	body	8.3926
drug	11.49	recognition	4.0119	form	7.2845
substance	6.2274	test	3.5491	order	5.5624
user	4.5384	recall	2.8567	part	4.5107
dependence	3.989	retention	2.5932	representation	4.4806
drinking	3.4883	item	2.4305	integration	3.0521
addiction	2.4297	retrieval	2.422	reference	2.631
cannabis	2.1811	experiment	2.3545	result	2.4552
alcohol_consumption	1.9942	list	1.7655	space	2.3281
marijuana	1.6846	short_term	1.4908	place	2.3214

6.1 Labeling the topics

The model's representation of a topic is in the form of a probability distribution over the vocabulary. While this is useful in the computational process, this is not how people generally represent abstract topics – a more human-like way is to label the topic with a term that represents the topic as a whole. To label the topics generated by the model, two approaches were employed, and both will be described as follows.

6.1.1 Labeling with WordNet hypernyms

The WordNet resource provides information about the semantic relations of words in the form of hypernyms and hyponyms. Hypernym is a higher level term in a taxonomy of labels (for example, child is a hypernym to boy and girl), meaning a common hypernym of multiple words acts as an umbrella term for the words and can be used as a label.

For this process of finding the common hypernym, a small subset of words from the topic have to be selected, since the addition of each new word rapidly decreases the specificity of the label and the resulting common hypernym becomes a very general term (such as *entity*). Multiple strategies were assessed for selecting the words from the topic for this process. The configuration that yielded the best results was selecting two to three words from the top 10 words of a topic, one of which is the top word. For each of such combinations, the common hypernyms were found, which were evaluated by its' depth in the hypernym tree (ie how many hypernym levels it takes to reach the *entity* node, which is the root node for nouns), picking the term with the biggest depth. The resulting labels obtained with this method can be seen in Appendix A and the source code for

generating these in Appendix B.

For assessing the suitability of this method for labelling in this context, it can be said that it provides some utility, but is lacking in some key aspects. That being mainly because of the fact that not all words – that do fit in one topic – have a semantically meaningful hypernym. For example, in topic 2 in Table 1, the words alcohol and user have the lowest common hypernyms *causal_agent* and *physical_entity*. These terms are too abstract to provide a meaningful label, but to a human labeler it is clear that these two words could be labeled under alcohol usage or something similar. This is because the closest relation between semantically tied words is not always hypernymy, but another kind of relation. It can be argued that a very detailed lexical database that models a larger variety of relations could yield better results for this process. Nevertheless, this method provided a useful alternative perspective compared to the labeling described next.

6.1.2 Human labeling

Since labeling the word-distribution topics is, as mentioned, a human way of dealing with these abstract topics, it can be assumed that humans are capable of coming up with good quality labels. Even more so, if the persons doing the labeling are acquainted to the domain the topics are about.

For this thesis, three people labeled the topics manually, out of whom one person is currently pursuing a Doctoral degree in psychology, the second has obtained a Master's degree in psychology and the third is a researcher in the field of natural language processing and has a PhD in cognitive sciences. The labelers were presented with individual sheets which showed the top 10 words for all the topics along with the associated probabilities. The labelers were asked to come up with labels that would describe the clusters of words as best as they could. The resulting labels can be seen in Appendix A.

All of the labelers reported difficulty labeling a minor proportion of the topics. 15% topics were reported by one of the three labelers as hard to label, 6% by 2 out of 3 labelers and 3% by all three labelers. This was mainly due to loose semantic similarity of the words of these topics. Overall, as can be seen in Appendix A, the labels assigned by the labelers generally overlap or are semantically similar.

6.2 Topic popularities over time

The generated topic model enables extraction of topic proportions per each document. Because the publication year of each document is known, the topic proportions can be aggregated for each year. From this point, the topic's relative occurrence can be plotted

to visualize the change in a given topic's popularity over time. An example of this visualisation can be seen in Figure 5 . On the graph, the horizontal axis represents a year, from 1979 to 2019, which gives us 40 years of changes. The vertical axis represents the proportional occurrence of the topic relative to all topics. The numeric values on this axis are less relevant themselves, but are primarily added to better capture the magnitude of the changes.

Hereafter, the topic's occurrences are ranked by a set of measures with the top ranking topics illustrated and briefly commented on. The full set of graphs displaying the proportion of each topic over time can be seen in Appendix A.

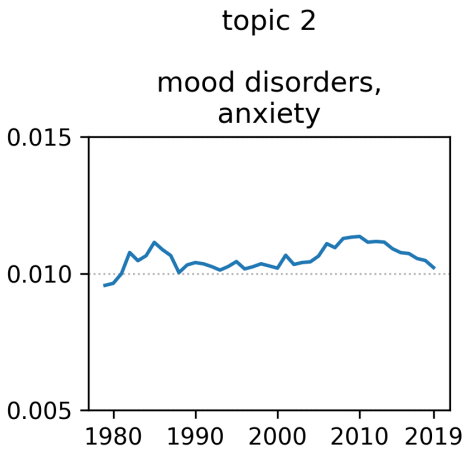


Figure 5. A sample topic's occurrence over time

6.2.1 Biggest increases in topic occurrence over the timespan

Figure 6 shows the graphs, along with selected labels, for topics which proportional increase has been the highest from 1979 to 2019. Topic 56, which features words related to neuroscience and fMRI, experienced a 92% increase in its proportional occurrence. Topic 28, that centers around interventions and the effectiveness of interventions, had a 81% increase over the course of this timespan. Topic number 8, which seems to signify reviews as a methodology for research, increased 62% over the selected period.

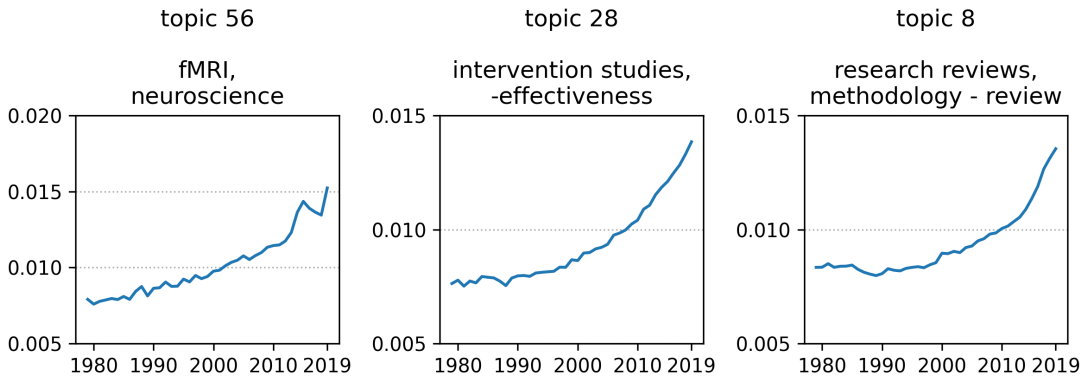


Figure 6. Topics with the biggest increase in occurrence from 1979 to 2019

6.2.2 Biggest decreases in the topic occurrence over the timespan

Figure 7 shows the changes in popularity for topics for which the popularity has decreased the most. Topic 23's proportional occurrence decreased 43% from 1979 to 2019, and even more (63%) when considering the period from its peak in 1988 to 2019. The topic features words about animal testing (rat, administration, injection) and drugs (cocaine, ethanol). The other two topics that exhibited the biggest decline, topic 98 and topic 53, were both one of the few that the labelers reported difficulty labeling. Topic 98, which received cognitive research as a label, had a 37% decrease over the timespan, whereas topic 53, which the labelers failed to label, but featured the top words concept and identity, exhibited a 34% decrease over the time period.

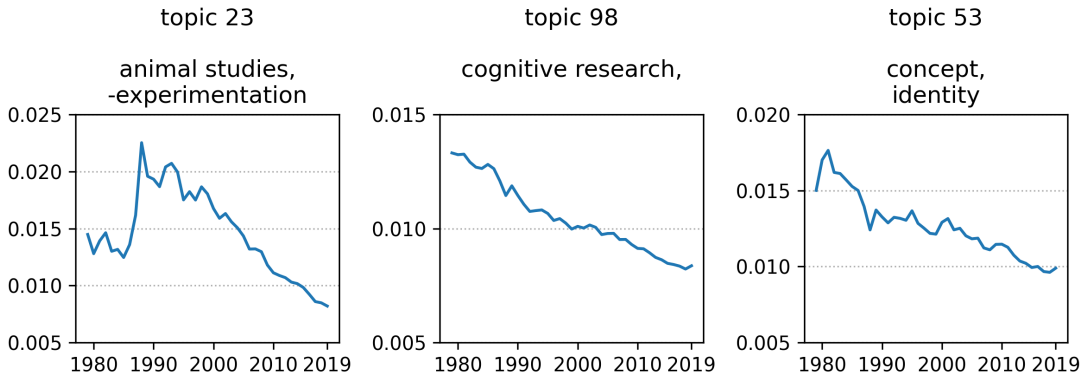


Figure 7. Topics with the biggest decrease in occurrence from 1979 to 2019

6.2.3 Most variance in topic occurrence

Figure 8 shows the topics for which the occurrence varied the most over the selected period of time. Topic number 23, discussed also in the previous subchapter, showed the most variance in occurrence from all the topics. The second-ranking topic for the amount of variance is topic number 5, which was labeled as studies related to neural activity.

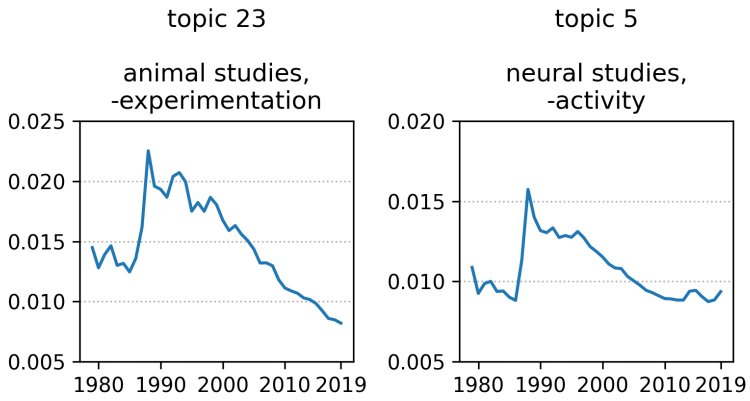


Figure 8. Topics with the largest variance over time in terms of occurrence from 1979 to 2019

6.2.4 Most stable topic occurrences (least variance)

The topics that showed the least amount of variance from all the topics are displayed in Figure 9. Topic 26, which showed the most stability by this metric, was labeled as representing paper abstract and experiment design. It seems feasible that this kind of topic is represented in most, if not all abstracts and its occurrence does not vary heavily across time. Another topic that showed a very low level of variance is topic number 33, which was labeled as representing visual perception studies. From the graph, a slow decline can be seen but the overall change across the time period is very small.

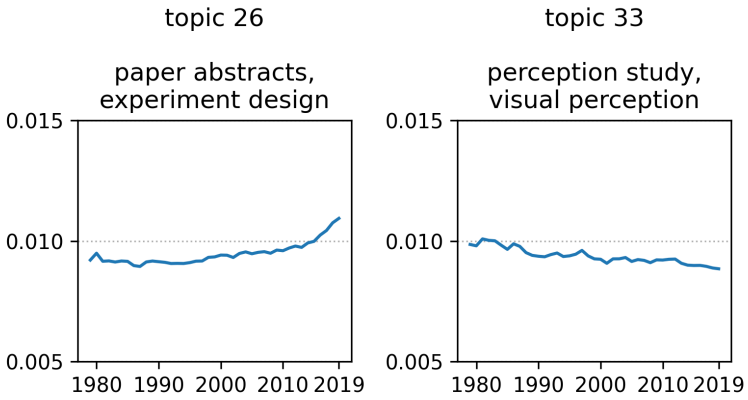


Figure 9. Topics with the least amount of variance over time in terms of occurrence from 1979 to 2019

7 Conclusion

In this thesis, a topic model was generated on a large collection of abstracts from the domain of psychology.

The goal of the thesis was to generate a topic model and analyse the resulting topics in the temporal dimension. This goal was met, through the following stages. The corpus for model generation was assembled from the PubMed resource and featured around 1.12 million abstracts. The texts were preprocessed for the model generation. The topic model was generated with MALLET's implementation of LDA, with 100 as the count of topics. The resulting topics were labeled with two strategies: manual labeling by three people well-aquainted to the domain, and by programmatic labeling using WordNet hypernyms. The topic occurrences were plotted in a time period of 40 years, from 1979 to 2019. The full set of resulting topics and corresponding graphs can be seen in Appendix A.

The occurrences of the topics were analysed from the statistical viewpoint, highlighting the topics outlying by some measure. The resulting model would likely provide an interesting model for analysis in some subsequent psychology research.

References

[1] *Annual Review of Psychology*. URL: https://www.annualreviews.org/journal/psych.

[2] *Annual Reviews - What We Do*. URL: https://www.annualreviews.org/about/what-we-do.

[3] Svetla Baykoucheva. "From the Science Citation Index to the Journal Impact Factor and Web of Science: interview with Eugene Garfield". In: *Managing Scientific Information and Research Data*. Ed. by Svetla Baykoucheva. Chandos Publishing, 2015, pp. 115–121. ISBN: 978-0-08-100195-0. DOI: https://doi.org/10.1016/B978-0-08-100195-0.00012-3.

[4] André Bittermann and Andreas Fischer. "How to Identify Hot Topics in Psychology Using Topic Modeling". In: *Zeitschrift für Psychologie* 226 (Feb. 2018), pp. 3–13. DOI: 10.1027/2151-2604/a000318.

[5] David M. Blei. "Probabilistic Topic Models". In: *Commun. ACM* 55.4 (Apr. 2012), pp. 77–84. ISSN: 0001-0782. DOI: 10.1145/2133806.2133826. URL: https://doi.org/10.1145/2133806.2133826.

[6] David Blei and John Lafferty. "Dynamic Topic Models". In: vol. 2006. Jan. 2006, pp. 113–120. DOI: 10.1145/1143844.1143859.

[7] David Blei, Andrew Ng, and Michael Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (May 2003), pp. 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

[8] S. C. Deerwester et al. "Indexing by Latent Semantic Analysis". In: *Journal of the American Society of Information Science* 41.6 (1990), pp. 391–407.

[9] *Gensim - LdaMallet*. URL: https://radimrehurek.com/gensim/models/wrappers/ldamallet.html.

[10] *Gensim: topic modelling for humans*. URL: https://radimrehurek.com/gensim/.

[11] Thomas Hofmann. "Probabilistic Latent Semantic Indexing". In: *the 22nd International Conference on Research and Development in Information Retrieval (SIGIR 99:1999* (Apr. 2004).

[12] Tai Chia Huang, Chia Hsuan Hsieh, and Hei Chia Wang. "Automatic meeting summarization and topic detection system". English. In: *Data Technologies and Applications* 52.3 (July 2018), pp. 351–365. ISSN: 2514-9288. DOI: 10.1108/DTA-09-2017-0062.

[13] Michael I. Jordan, ed. *Learning in Graphical Models*. Cambridge, MA, USA: MIT Press, 1999. ISBN: 0262600323.

[14] Wei Li and Andrew Mccallum. “Pachinko allocation: DAG-structured mixture models of topic correlations”. In: June 2006, pp. 577–584. DOI: 10.1145/1143844.1143917.

[15] Andrew McCallum. *MALLET: A Machine Learning for Language Toolkit.* URL: http://mallet.cs.umass.edu/.

[16] *Multiword Expressions*. URL: https://aclweb.org/aclwiki/Multiword_Expressions.

[17] *Natural Language Toolkit*. URL: https://www.nltk.org/.

[18] *PsycArticles database*. URL: https://www.apa.org/pubs/databases/psycarticles.

[19] *PubMed Help*. URL: https://www.ncbi.nlm.nih.gov/books/NBK3827/.

[20] S. Pyo, E. Kim, and M. kim. “LDA-Based Unified Topic Modeling for Similar TV User Grouping and TV Program Recommendation”. In: *IEEE Transactions on Cybernetics* 45.8 (2015), pp. 1476–1490.

[21] Nadine Schneider et al. “Chemical Topic Modeling: Exploring Molecular Data Sets Using a Common Text-Mining Approach”. In: *Journal of Chemical Information and Modeling* 57.8 (2017). PMID: 28715190, pp. 1816–1831. DOI: 10.1021/acs.jcim.7b00249. URL: https://doi.org/10.1021/acs.jcim.7b00249.

[22] *spaCy - Industrial-strength Natural Language Processing in Python*. URL: https://spacy.io/.

A Topics

In this section, the resulting topics from the generated model are presented along with the labels and the graphs describing the change in a topics occurrence over time. The topics are described in the format shown in Table 2.

Words 1-10 are the top words of a given topic ranked by the corresponding probability. The probability indicates the percentage of drawing that particular word then picking a word from this topic (ie how big of a proprtion of the topic does this word make up).

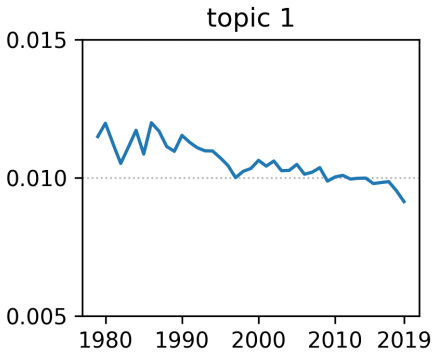
Labels 1-3 are labels assigned by the human labelers. Label 4 is the label assigned by the WordNet labeling.

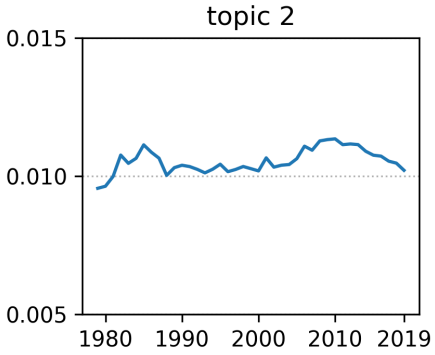
On the right side of the topic's table is the corresponding temporal ocurrence graph.

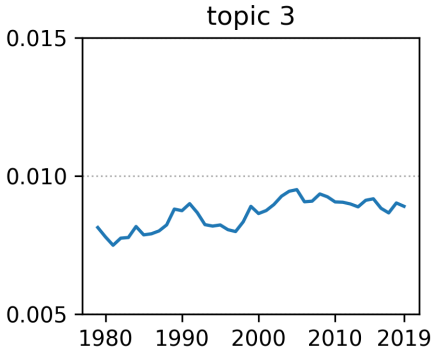
Table 2. The format of the topics shown in this section

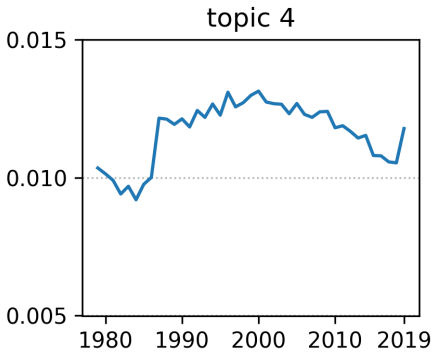
Topic number		
word 1	probability %	label 1
word 2	probability %	label 2
word 3	probability %	label 3
word 4	probability %	label 4
word 5	probability %	
word 6	probability %	
word 7	probability %	
word 8	probability %	
word 9	probability %	
word 10	probability %	

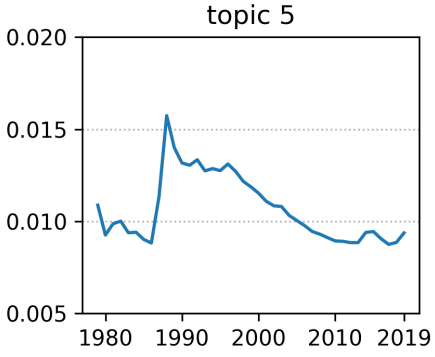
1		
alcohol	18.059	substance abuse
drug	11.49	substance use
substance	6.2274	substance dependence
user	4.5384	drug of abuse
dependence	3.989	
drinking	3.4883	
addiction	2.4297	
cannabis	2.1811	
alcohol_consumption	1.9942	
marijuana	1.6846	
2		
depression	40.9001	anxiety and mood disorders
anxiety	29.5724	mood disorder assessment
symptom	5.2716	mental disorders
scale	2.201	mental disorder
inventory	1.4201	
severity	1.1697	
beck_depression	1.06	
gad	0.9094	
bdi	0.7474	
anxiety_disorder	0.7039	
3		
smoking	11.017	smoking
smoker	6.3129	smoking
tobacco	4.3614	smoking behaviour
study	3.5668	tobacco
cigarette	3.1652	
nicotine	2.8531	
smoking_cessation	2.3951	
initiation	2.2617	
rate	1.7897	
abstinence	1.7242	
4		
target	6.9361	visual perception research
object	6.077	sensory/experimental psychology
experiment	4.7906	psychological experiment
location	3.5465	aim
motion	2.3949	
color	1.9075	
direction	1.8916	
observer	1.8527	
feature	1.6404	
stimulus	1.4902	
5		
area	4.425	neuroanatomy research
stimulation	4.2308	neural studies
lesion	4.1023	neural activity
neuron	3.0395	area
site	2.687	
pathway	1.7027	
cell	1.5802	
activity	1.3773	
response	1.3143	
system	1.2345	



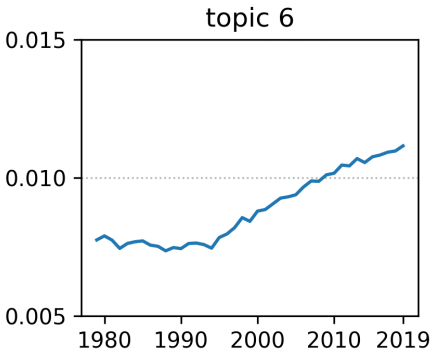


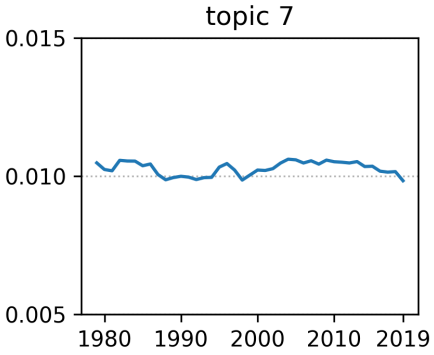


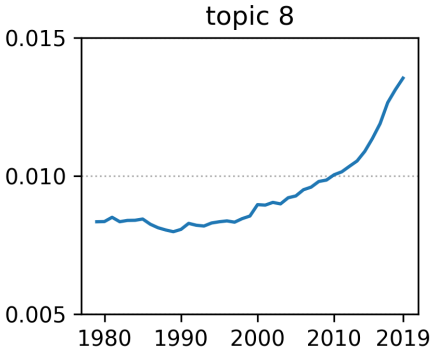


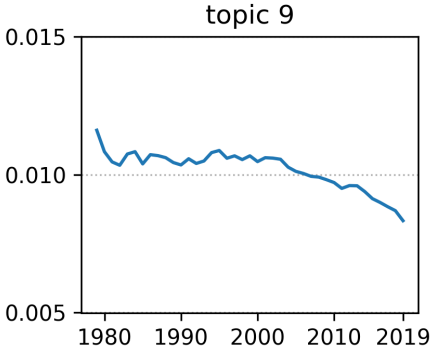


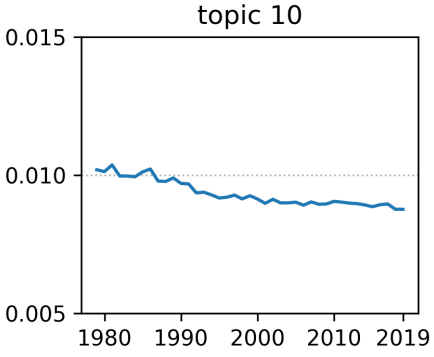
6		
participant	49.1592	studies
people	25.043	participants
person	12.4691	human subjects
study	5.25	organism
individual	2.3426	
research	0.7357	
intellectual_disability	0.4412	
finding	0.3677	
half	0.2409	
took_part	0.2236	
7		
score	40.0867	psychometry
scale	13.6458	instrument description
index	4.8377	statistics
correlation	4.4941	appraisal
questionnaire	4.1798	
subscale	2.8521	
measure	1.9237	
inventory	1.6445	
assess	1.3433	
assessment	1.3346	
8		
study	17.2162	research reviews
evidence	9.6112	methodology - review
review	7.5107	review
literature	7.2114	examination
article	3.7485	
search	2.895	
outcome	2.3373	
research	2.217	
systematic_review	1.3932	
author	1.3931	
9		
case	15.5041	domestic violence
history	8.4603	case history - violence
violence	6.7685	sexual abuse
abuse	4.9827	happening
victim	3.1864	
sexual_abuse	1.818	
offender	1.7578	
incident	1.7187	
victimization	1.6321	
report	1.6096	
10		
ability	16.6731	studies of ability
study	14.5109	-
difficulty	9.9147	study significance
result	7.4047	ability
degree	4.0571	
limitation	2.211	
volunteer	2.1146	
test	1.5941	
finding	1.4913	
mobility	1.4188	



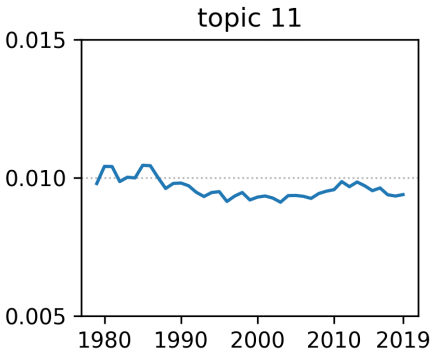


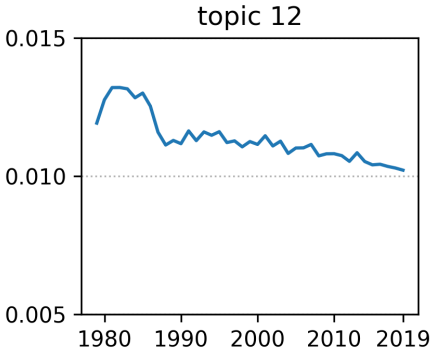


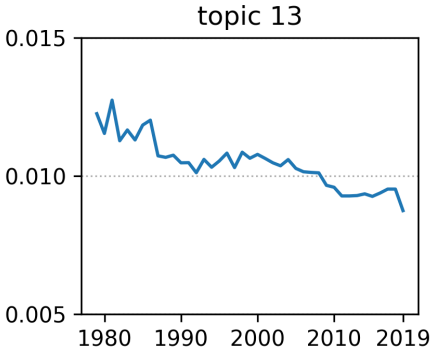


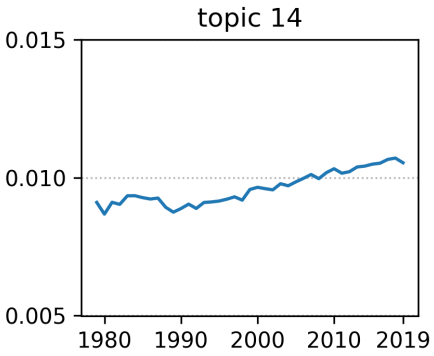


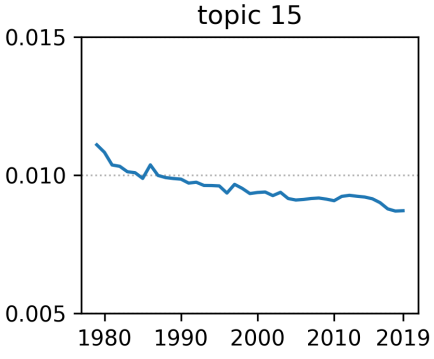
11		
relationship	35.2873	couples' relationship studies
partner	6.9971	couple studies
conflict	5.3994	relationship dynamic
study	4.5575	relationship
couple	3.4571	
style	3.2534	
finding	2.9792	
trust	2.1505	
spouse	1.9795	
research	1.9465	
12		
treatment	52.5954	psychotherapy efficacy research
therapy	12.3042	psychotherapy intervention
cbt	2.4859	psychotherapy
therapist	2.4062	treatment
compliance	2.4046	
outcome	2.1166	
psychotherapy	1.9143	
improvement	1.7516	
session	1.2392	
outpatient	1.0943	
13		
patient	11.818	studies involving surgery
surgery	7.6231	medical intervention
procedure	3.699	surgery protocol
complication	2.0074	procedure
case	1.8206	
operation	1.6029	
surgeon	1.398	
technique	1.1266	
biofeedback	0.6766	
questionnaire	0.5954	
14		
month	27.3761	intervention efficacy
outcome	21.5312	longitudinal study
baseline	9.9003	standardized assessment
follow	6.7622	time period
long_term	6.2968	
improvement	4.4822	
period	2.5313	
assessment	1.8963	
outcomes	1.5213	
measure	1.4147	
15		
behavior	54.9374	studies of aggression
study	5.9781	behavioural study
aggression	5.7165	study outcome
result	3.8374	behavior
finding	3.1077	
influence	2.1107	
research	1.1396	
implication	1.1237	
interaction	0.8154	
addition	0.796	



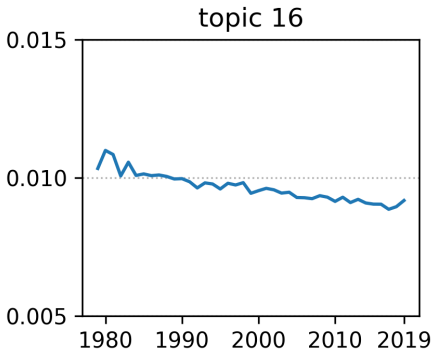


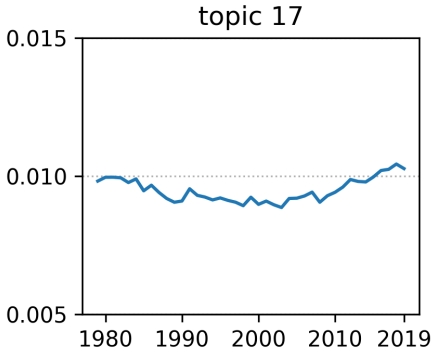


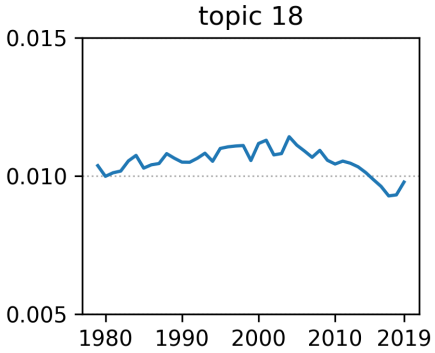


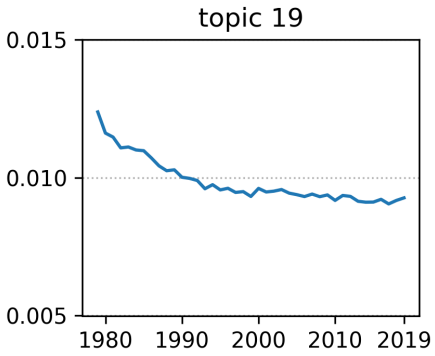


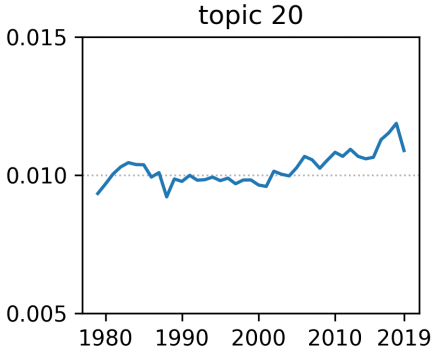
16		
body	8.3926	-
form	7.2845	-
order	5.5624	-
part	4.5107	body
representation	4.4806	
integration	3.0521	
reference	2.631	
result	2.4552	
space	2.3281	
place	2.3214	
17		
program	19.8763	sport psychology
training	17.1222	training program (physical)
skill	8.4367	sport for health
participation	6.3021	plan
confidence	3.2896	
team	2.9068	
sport	2.2989	
improvement	1.7517	
game	1.6121	
athlete	1.3057	
18		
memory	24.5233	memory research
recognition	4.0119	experimental - memory
test	3.5491	memory
recall	2.8567	memory
retention	2.5932	
item	2.4305	
retrieval	2.422	
experiment	2.3545	
list	1.7655	
short_term	1.4908	
19		
strategy	14.6117	Motor learning studies
error	6.6483	maybe experimental
feedback	4.8281	task analysis
procedure	4.6933	idea
learning	4.2125	
result	3.6933	
technique	3.415	
delay	3.298	
failure	2.9742	
instruction	2.6795	
20		
student	27.6613	educational psychology
education	4.8746	educational (psychology)
university	4.1309	education
nursing	3.4688	whole
college	1.8585	
learning	1.7883	
teaching	1.5062	
faculty	1.4182	
career	1.3692	
curriculum	1.2934	



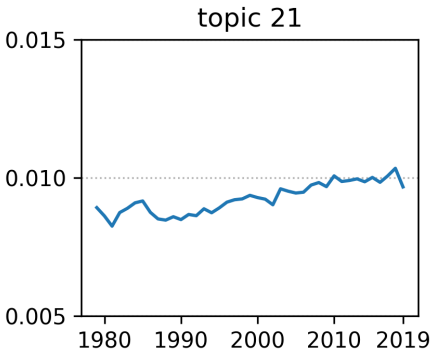


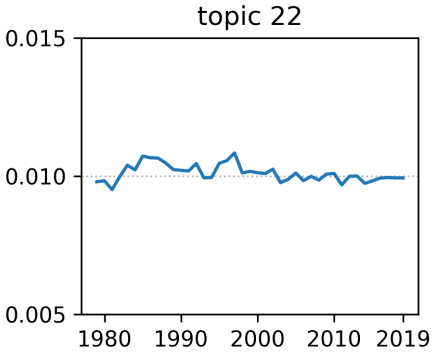


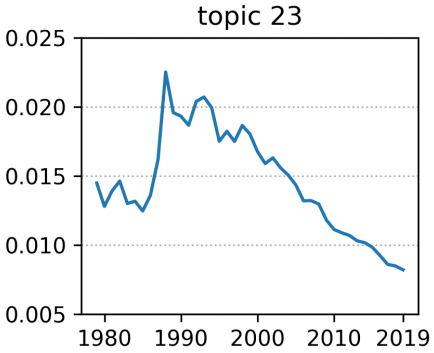


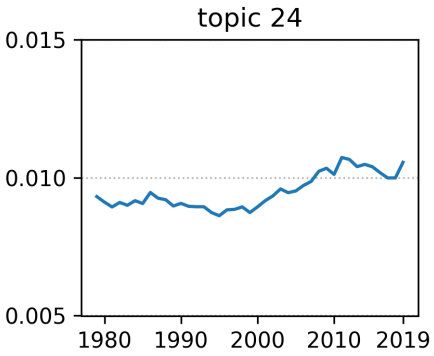


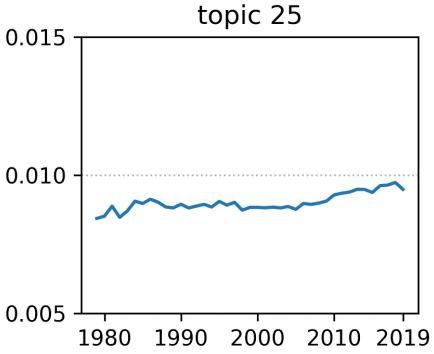
21		
cancer	16.9004	cancer research
patient	6.0353	medical - cancer, stroke
survivor	6.0196	cancer cure
stroke	5.3672	cancer
diagnosis	4.9055	
breast_cancer	3.5215	
treatment	2.985	
survival	2.7227	
disease	2.1777	
chemotherapy	1.5531	
22		
disease	12.331	neurodegenerative disease research
dementia	8.6198	cognitive impairment - dementia, alzheimer
impairment	3.607	symptoms
alzheimer_disease	2.7491	ill health
decline	2.6245	
patient	2.1744	
diagnosis	1.9131	
headache	1.8023	
progression	1.7651	
mci	1.5603	
23		
rat	7.1553	animal studies
administration	3.579	animal experiments
effect	3.5422	animal experimentation
animal	2.3773	animal
cocaine	2.2542	
injection	1.9345	
ethanol	1.7985	
dose	1.7151	
experiment	1.5924	
drug	1.5768	
24		
emotion	10.3717	affect research
category	7.3976	affect recognition
face	5.8756	emotion recognition
image	3.7533	feeling
recognition	3.1451	
expression	3.0289	
picture	2.9799	
affect	2.6964	
judgment	2.6426	
participant	2.421	
25		
support	32.0064	social support
social	7.387	social support
study	6.7224	-
resource	5.5651	device
network	4.1936	
contact	4.0232	
impact	3.8608	
finding	2.6247	
isolation	2.0683	
datum	1.9597	



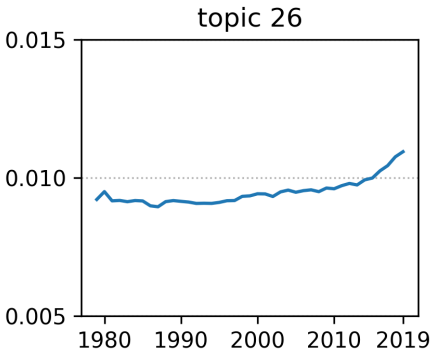


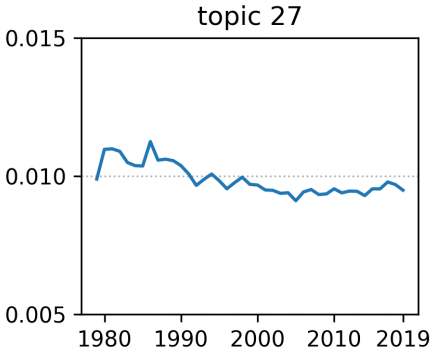


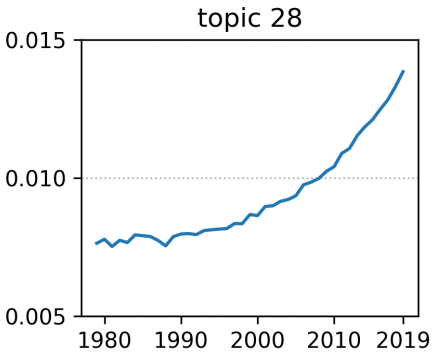


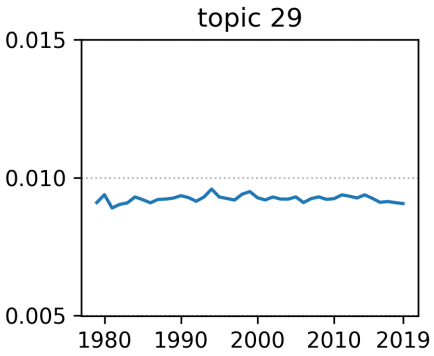


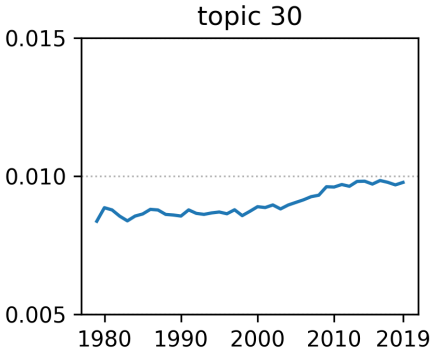
26		
datum	22.6329	paper abstracts
study	16.4803	methodology - time
method	15.202	experiment design
analysis	14.8905	datum
data	6.825	
design	2.8765	
collection	2.2004	
purpose	1.3266	
conclusion	1.0793	
aim	0.8318	
27		
mother	14.5371	pregnancy and birt
infant	9.7793	maternity study
pregnancy	7.6518	giving birth
birth	4.3677	parent
postpartum	2.8579	
maternal	2.6709	
father	2.6638	
month	2.5466	
delivery	1.9875	
development	1.4805	
28		
intervention	37.244	intervention studies
efficacy	7.288	intervention effectiveness
trial	6.8784	outcome assessment
effectiveness	3.4093	proceeding
programme	3.2095	
outcome	3.1278	
mindfulness	1.3892	
design	1.3746	
reduction	1.3707	
pilot	1.2435	
29		
time	41.7533	research study characteristics
exposure	9.6357	methodology - time/longitudinal
point	7.0308	time as variable
study	5.7194	datum
stability	1.6281	
period	1.4794	
impact	1.4344	
dentist	1.3091	
timing	1.2232	
datum	1.0061	
30		
adult	30.8961	conditions through life
childhood	7.0931	lifespan or age
study	6.4474	adolescent development
onset	6.4268	organism
adhd	6.2867	
age	5.638	
adulthood	3.7414	
adolescence	2.5803	
sample	2.4772	
finding	2.4218	



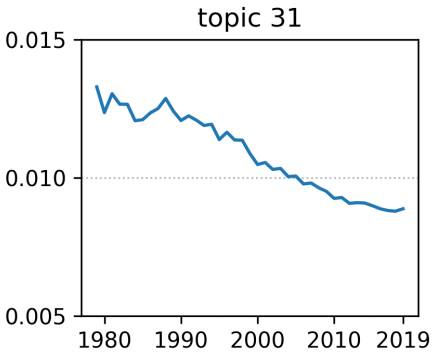


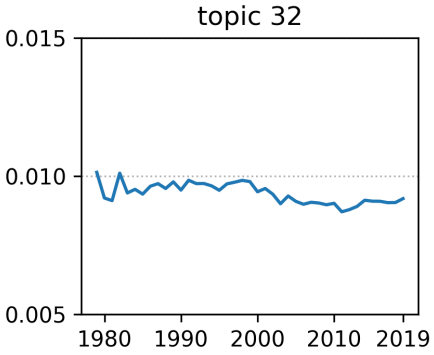


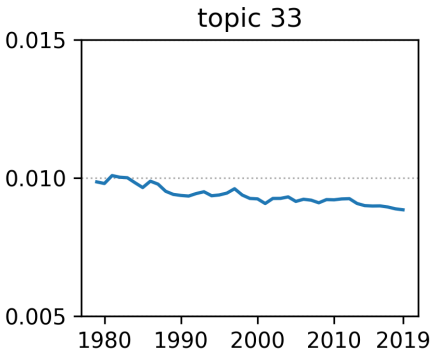


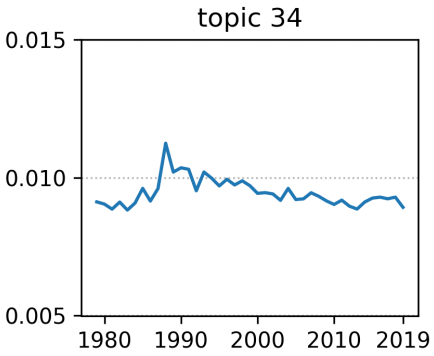


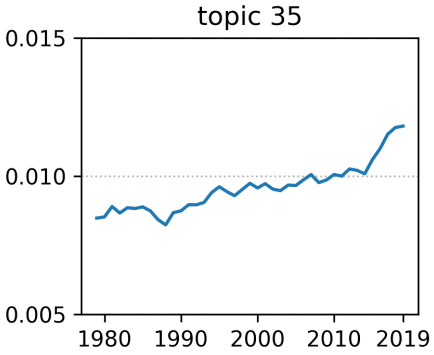
31		
subject	25.7702	motor/movement studies
motor	8.4929	-
phase	6.8808	development of movement
movement	5.9092	content
hand	4.8455	
balance	1.4017	
muscle	1.1291	
force	1.0046	
arm	0.9473	
limb	0.845	
32		
sleep	12.3207	sleep disorders studies
duration	8.0501	sleep research
mood	7.2415	sleep study
fatigue	6.9246	physical condition
insomnia	2.7348	
complaint	2.5406	
study	2.527	
shift	2.1784	
night	1.9533	
disturbance	1.4878	
33		
perception	24.5811	visual perception studies
study	11.636	sensory/perception
rating	9.8671	perception study
class	4.5294	activity
result	3.6008	
orientation	3.4531	
agreement	3.0389	
finding	2.7318	
voice	2.412	
report	2.003	
34		
behaviour	14.6348	diet behaviour studies
food	10.6155	eating behaviour
diet	2.8619	nutrition study
consumption	2.6211	activity
intake	2.0474	
nutrition	1.7492	
meal	1.5541	
energy	1.3603	
consumer	1.2808	
deprivation	1.182	
35		
practice	16.3438	-
client	4.4482	treatment engagement
implementation	3.6005	-
guideline	3.2563	activity
engagement	3.1633	
recommendation	2.7942	
evidence	2.2786	
project	2.2009	
approach	2.0869	
expert	1.9246	



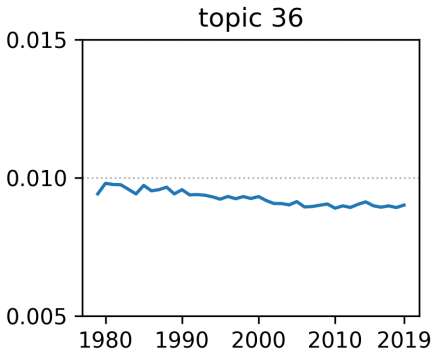


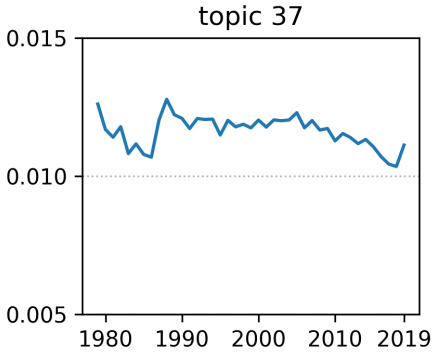


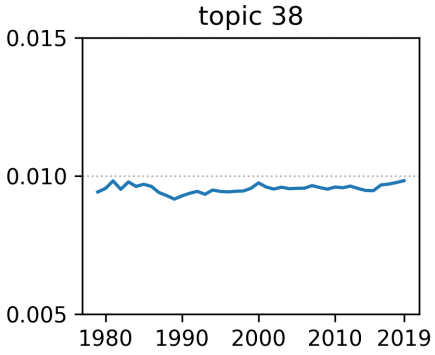


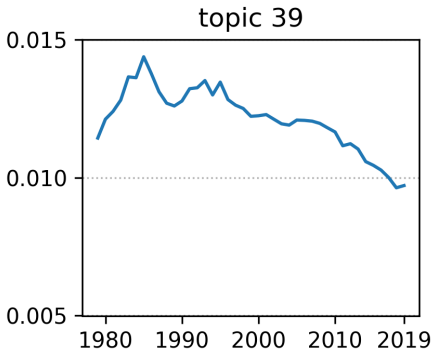


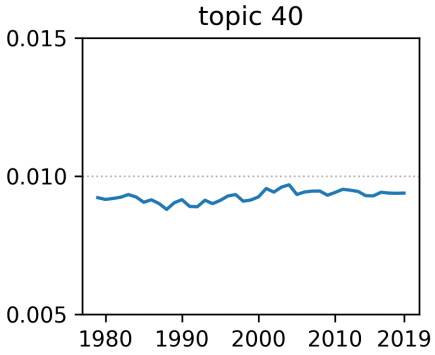
36		
function	19.0866	strength training research
component	10.2966	prediction
study	8.7299	recovery of function
loss	8.2524	function
recovery	7.7574	
capacity	5.3191	
strength	3.5938	
result	2.9256	
finding	2.1741	
core	2.0468	
37		
response	18.2599	auditory perception studies
stimulus	10.3756	experimental
cue	5.1179	psychological experiment
experiment	3.8874	experiment
trial	3.0591	
auditory	3.0482	
stimuli	2.86	
presentation	2.2772	
processing	2.1386	
sequence	1.8646	
38		
assessment	21.2289	assessment
evaluation	10.4908	assessment
tool	7.662	evaluation procedure
screening	5.4783	appraisal
criterion	4.4121	
examination	2.8584	
measurement	2.4481	
indicator	1.9415	
protocol	1.9313	
screen	1.8901	
39		
disorder	29.3719	psychopathology research
diagnosis	7.6477	disorder assessment
psychosis	2.762	psychopathology
dsm	2.1488	disorder
ocd	1.9855	
comorbidity	1.878	
criterion	1.7847	
bipolar_disorder	1.7474	
psychopathology	1.6813	
episode	1.6122	
40		
information	33.6898	-
question	6.859	-
source	6.5918	online psychological study
expectation	4.8542	pleading
content	4.7097	
internet	3.2102	
material	2.0427	
communication	1.6542	
format	1.3169	
answer	1.1593	



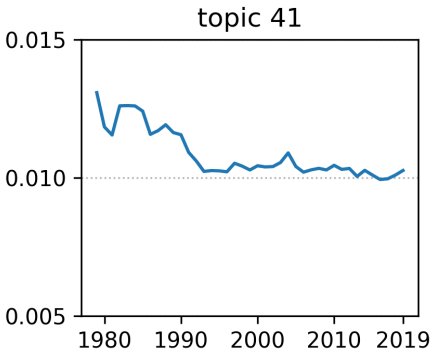


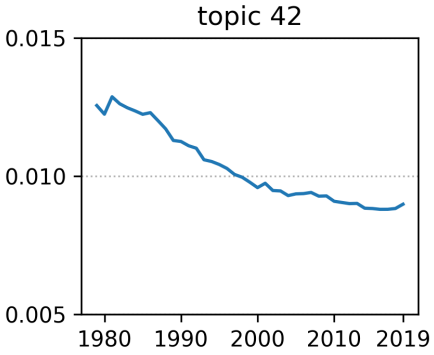


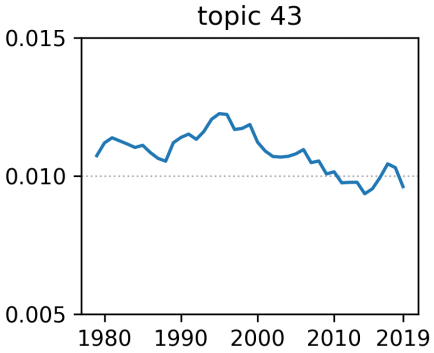


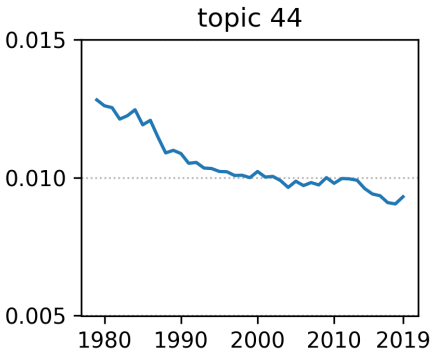


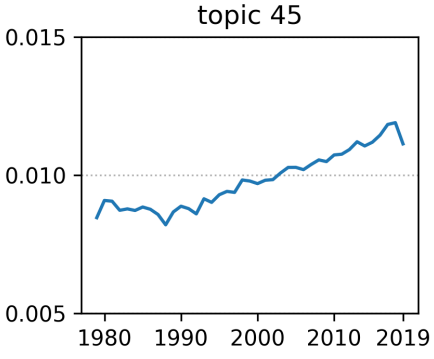
41		
word	10.0163	language processing research
language	8.6826	language processing
speech	4.3501	language
processing	2.5901	speech act
experiment	1.9973	
reading	1.9907	
letter	1.6954	
sentence	1.6613	
comprehension	1.4525	
result	1.3953	
42		
type	19.4865	autism spectrum research
pattern	14.2315	disorder specification
profile	5.3592	statistical analysis
characteristic	5.2929	concept
analysis	4.9153	
subgroup	4.0934	
syndrome	3.9994	
cluster	3.7881	
autism	3.0842	
insight	2.3855	
43		
nurse	14.0235	nurses
hospital	11.8521	nursing
care	8.5538	health care
staff	6.5173	health professional
nursing	5.1667	
unit	4.3915	
resident	3.9736	
admission	2.5174	
discharge	2.1667	
inpatient	1.7761	
44		
personality	8.7357	borderline personality disorder research
dimension	7.3525	personality
trait	5.3645	personality disorder
attachment	2.9486	whole
sample	2.9337	
anger	2.6183	
bpd	2.3293	
impulsivity	2.263	
psychopathology	2.0295	
relationship	1.7748	
45		
experience	20.5807	qualitative research
interview	8.1922	methodology - qualitative
theme	4.8541	interview
focus	3.3004	content
understanding	1.9403	
narrative	1.7913	
challenge	1.7353	
explore	1.7224	
analysis	1.5988	
content	1.5863	



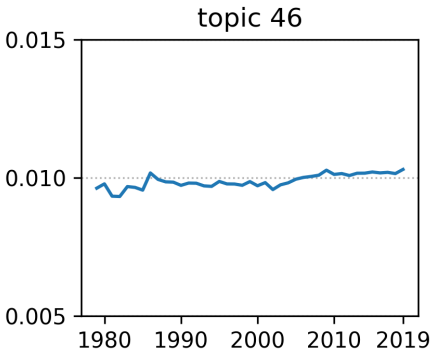


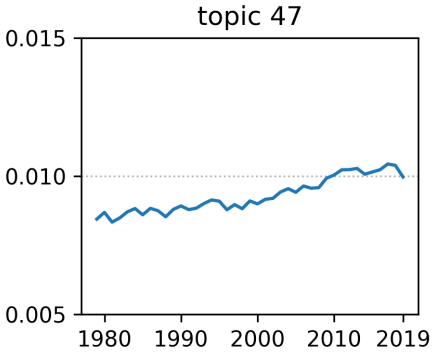


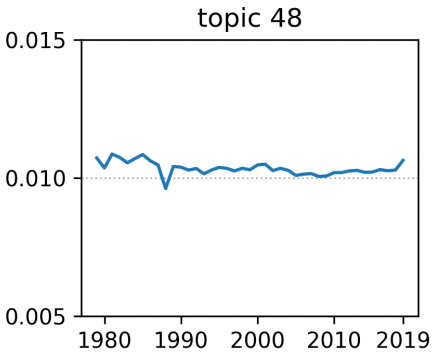


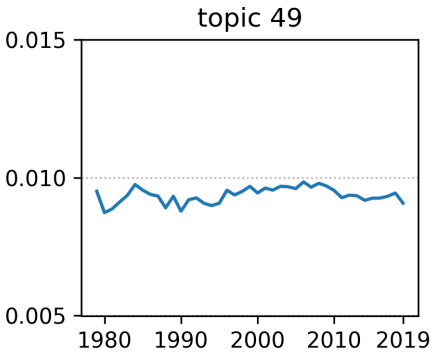


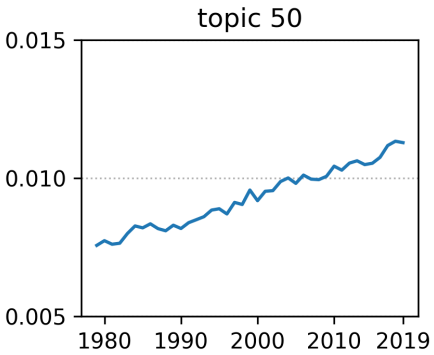
46		
age	42.3593	characteristics of study participants
year	34.7975	sample description
cohort	3.3474	group comparison design
education	2.8065	time period
range	1.5474	
decline	1.3617	
gender	1.0126	
age_group	0.9326	
population	0.8692	
sex	0.7796	
47		
knowledge	18.5485	studies on knowledge and attitudes
attitude	9.0529	-
belief	8.7123	-
awareness	6.0416	cognition
intention	4.5955	
education	3.413	
norm	2.2824	
message	1.9204	
questionnaire	1.9027	
medium	1.8323	
48		
process	13.1242	theory
approach	10.6821	theoretical perspective
context	9.9131	-
theory	9.2586	activity
perspective	3.3135	
framework	3.0786	
understanding	2.223	
contexts	2.2115	
element	1.9928	
complexity	1.8793	
49		
pain	34.5109	pain research
disability	8.8666	pain
intensity	5.2787	pain management
chronic	2.7256	pain
study	1.4139	
management	1.365	
vas	1.202	
relief	1.0797	
discomfort	0.9414	
tolerance	0.9357	
50		
health	53.8243	health status and its determinants
status	8.491	health research
population	2.7599	-
survey	2.1701	condition
education	2.0685	
promotion	1.9058	
literacy	1.2769	
household	0.9914	
determinant	0.9002	
professional	0.809	



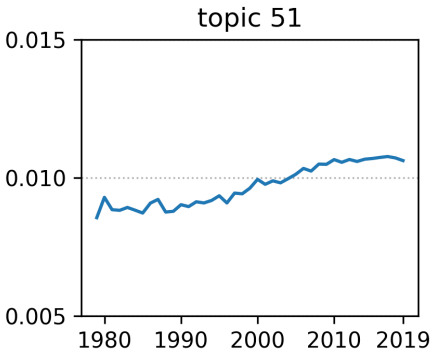


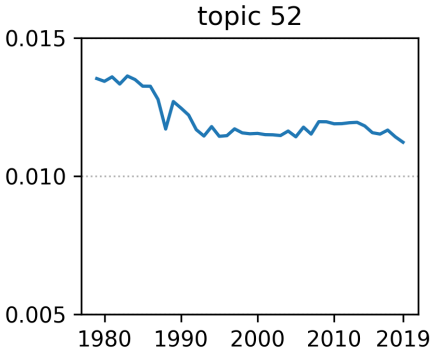


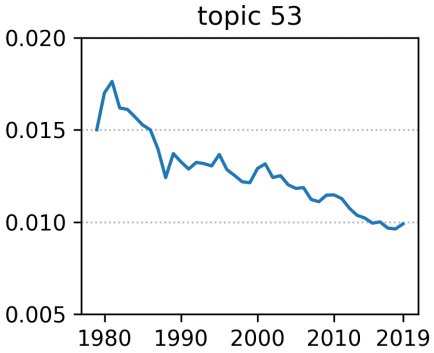


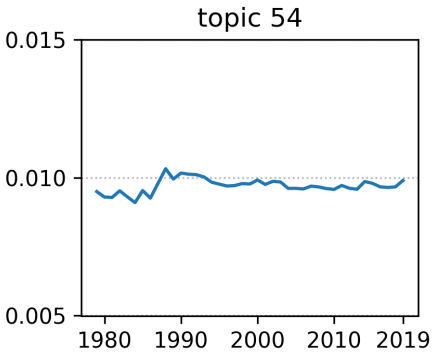


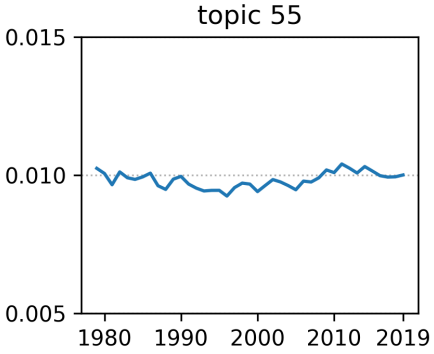
51		
factor	44.473	cardiovascular disease research
risk	42.2491	-
prevention	1.3944	prevention research
cvd	0.7715	integer
vulnerability	0.7411	
cardiovascular_disease	0.6748	
logistic_regression	0.5149	
population	0.4497	
history	0.3641	
myocardial_infarction	0.2445	
52		
child	62.1794	study population in developmental studies
parent	20.1051	family/parenting research
family	2.057	parenting
parenting	1.8447	organism
year	0.7052	
childhood	0.5131	
toddler	0.4049	
mother	0.3898	
preschooler	0.3807	
father	0.3417	
53		
concept	5.3208	-
identity	3.2748	-
psychology	2.8872	-
author	1.975	concept
paper	1.4125	
article	1.3272	
idea	1.265	
theory	1.0733	
psychiatry	1.0644	
world	1.0051	
54		
effect	74.9758	words used in describing results
result	4.6382	prediction/effect
side	3.3074	psychological experiment
influence	1.6002	evidence
interaction	1.5465	
evidence	0.9931	
addition	0.7422	
mechanism	0.7351	
enhancement	0.7146	
size	0.6548	
55		
adolescent	18.5774	adolescent/youth as study population
school	12.676	sample - adolescent
youth	9.1434	adolescence
peer	6.0038	organism
girl	4.2691	
teacher	3.7582	
grade	3.4501	
boy	3.1675	
year	2.0386	
adolescence	1.3696	



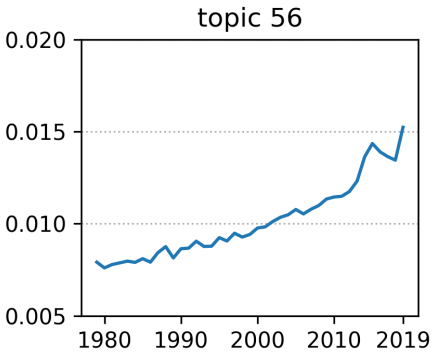


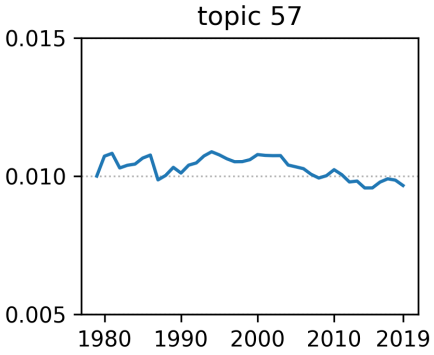


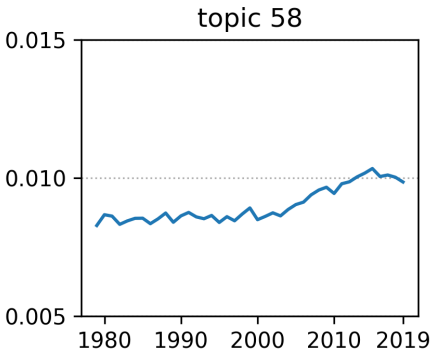


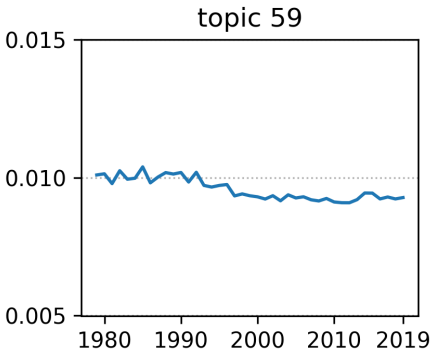


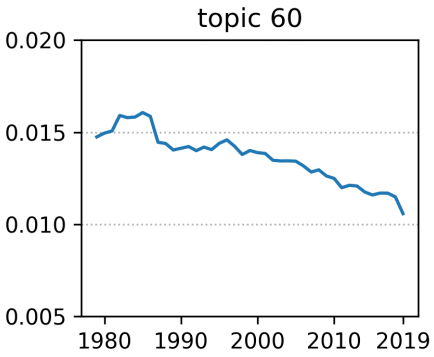
56		
brain	12.5156	FMRi studies
region	7.2315	neurological research
activation	4.9558	neuroscience
network	3.9173	neural structure
area	2.7491	
volume	2.1301	
fmri	1.9251	
amygdala	1.7211	
connectivity	1.4744	
structure	1.072	
57		
issue	6.0756	social policy studies
policy	3.9742	practical outcome - policy
discussion	2.6939	legislation
view	2.162	activity
autonomy	1.8202	
society	1.8195	
responsibility	1.7379	
statement	1.3716	
concern	1.3659	
law	1.2252	
58		
activity	32.8049	weight and physical activity research
weight	7.6394	health - physical activity
exercise	7.4905	obesity
obesity	4.6595	activity
bmi	2.6186	
lifestyle	2.4609	
study	1.7248	
fitness	0.9908	
cam	0.9794	
body_mass_index	0.9277	
59		
level	60.7663	biomarker levels as a variable
study	4.2812	methodology - medical testing
concentration	2.5877	physiology
blood	1.6855	device
plasma	1.4772	
serum	1.241	
result	1.1559	
testosterone	0.9128	
lead	0.8366	
increase	0.8121	
60		
patient	95.4081	patient
healthy_controls	0.5168	patients
outpatient	0.4648	heart disease
patients	0.4538	affected role
disease	0.2466	
prognosis	0.2362	
heart_failure	0.2241	
myocardial_infarction	0.1493	
angina	0.1174	
pci	0.1142	



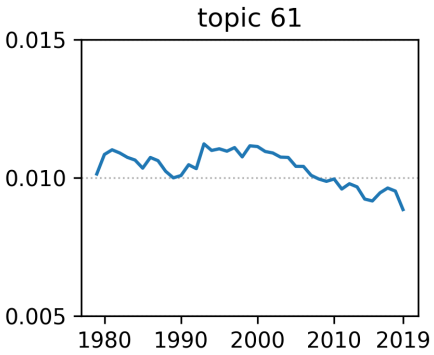


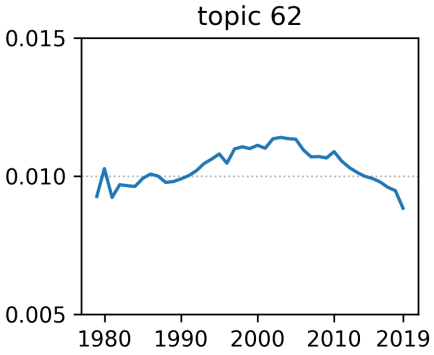


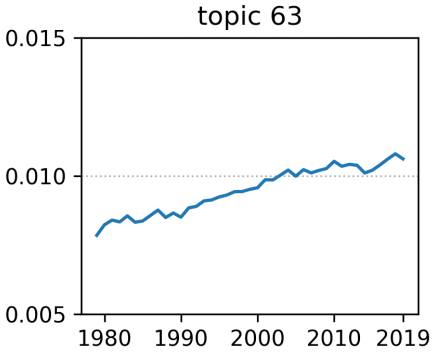


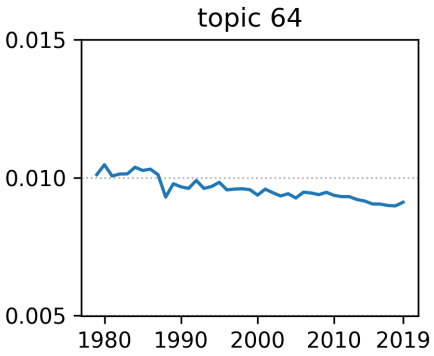


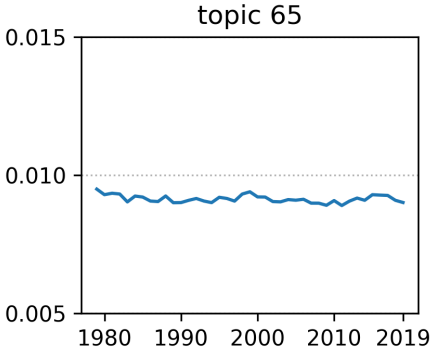
61		
physician	11.5485	-
communication	7.5203	-
clinic	4.7507	primary care
doctor	4.1962	doctor
visit	3.5309	
primary_care	3.4053	
percent	3.3016	
consultation	2.786	
asthma	2.7831	
medicine	2.2022	
62		
woman	52.5437	-
man	17.2632	sample - gender
sexuality	1.356	sexual behavior
abortion	1.3264	organism
fertility	0.9687	
infertility	0.8519	
menopause	0.7154	
sexual_activity	0.6234	
contraception	0.5602	
desire	0.5397	
63		
survey	11.9595	survey/questionnaire as a method
satisfaction	11.1955	methodology - instrument and sample
respondent	6.4134	survey
questionnaire	5.5617	act
country	4.2383	
cross_sectional	2.0613	
population	1.6875	
majority	1.574	
area	1.3867	
question	1.2418	
64		
problem	35.8313	mental health research
mental_health	19.9528	mental health research
adjustment	5.093	adjustment
study	3.6579	difficulty
mental_illness	3.1367	
population	2.1941	
difficulty	1.9253	
solution	1.3858	
sample	1.2636	
conduct	1.1809	
65		
change	35.9581	traumatic brain injury research
injury	9.1282	physical health - intervention
stage	6.9367	rehabilitation program
rehabilitation	6.4313	change
post	4.1922	
study	3.6145	
increase	2.5292	
tbi	2.1857	
impact	1.5724	
decrease	1.2828	



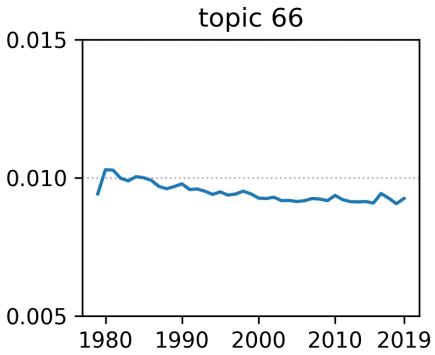


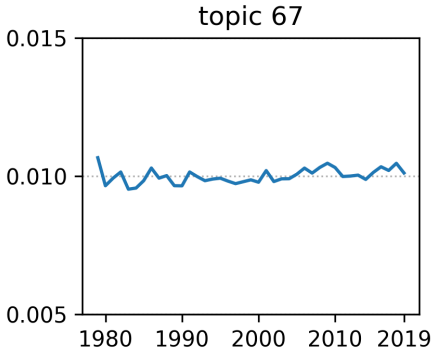


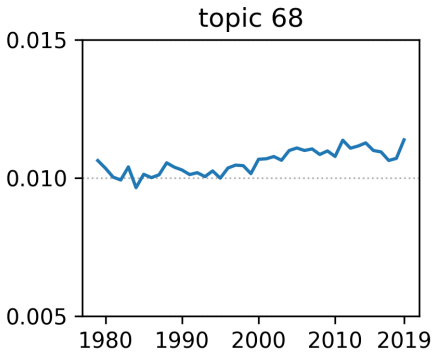


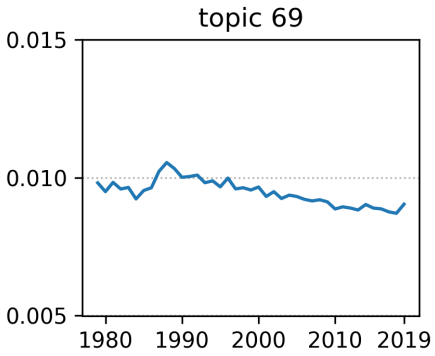


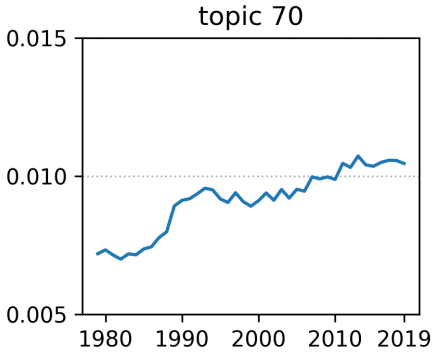
66		
role	23.5869	development role
development	19.488	role
relation	7.6392	-
involvement	5.9348	activity
finding	4.0946	
aspect	3.3897	
importance	3.1325	
growth	2.5258	
influence	2.107	
contribution	2.0975	
67		
rate	19.3357	suicide research
prevalence	9.6387	mapping prevalence
population	9.3361	suicide
suicide	6.9032	change of state
mortality	3.0346	
incidence	2.9692	
death	2.5541	
risk	2.0125	
suicide_attempt	1.7942	
harm	1.7119	
68		
task	30.5309	working memory research
performance	19.5796	experimental psychology (working memory)
attention	11.3508	psychological experiment
inhibition	2.3473	work
load	1.6598	
accuracy	1.6147	
interference	1.5599	
processing	1.1999	
speed	1.1863	
working_memory	1.1377	
69		
number	18.4435	comparing numeric differences
sensitivity	7.967	experimental psychology (perception)
size	6.7947	numerical analysis
comparison	4.1002	merchandise
result	3.9728	
line	3.0593	
ratio	2.5217	
length	2.2381	
distance	2.1837	
difference	2.1734	
70		
hiv	12.0946	HIV research
stigma	3.5562	HIV/AIDS research
prevention	3.1742	AIDS
sex	2.841	ill health
partner	2.4556	
condom	2.3832	
aids	2.2637	
risk	2.2502	
disclosure	2.0944	
art	1.9058	



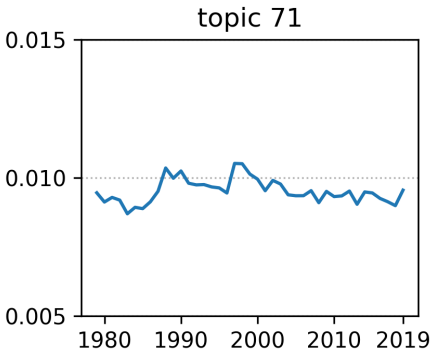


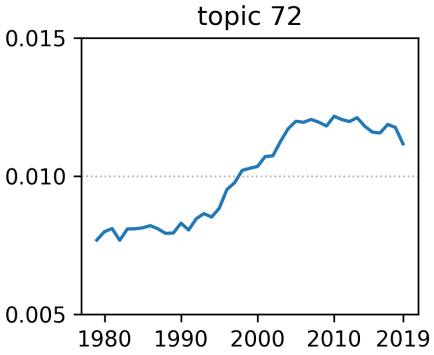


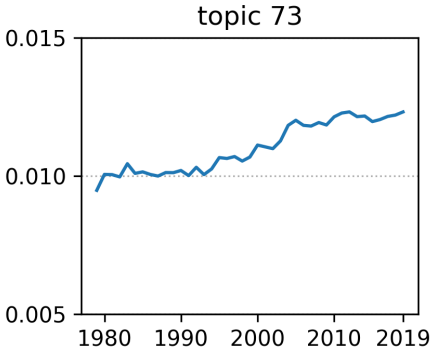


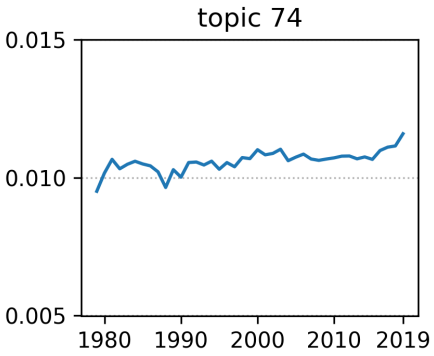


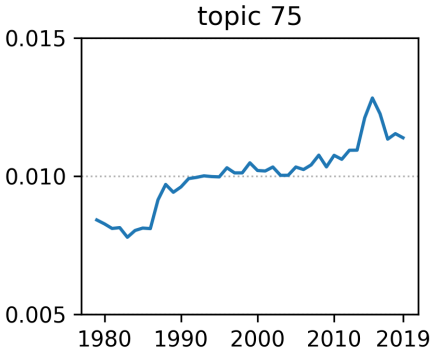
71		
animal	4.2799	evolutionarily comparative studies
adaptation	3.9455	adaption research with animals
human	3.6985	adaptation
specie	1.9378	animal
environment	1.3276	
evolution	1.0335	
laboratory	1.0211	
olfactory	0.9684	
monkey	0.9582	
call	0.8869	
72		
life	32.4904	health and life quality research
quality	25.0209	life quality
qol	8.1876	health related quality of life
hrqol	4.3075	content
domain	3.1821	
impact	3.0203	
questionnaire	2.7703	
health	2.4619	
disease	2.4243	
hrql	0.8074	
73		
item	12.2697	psychometry
scale	8.0631	methodology - instrument
instrument	4.662	measurement
measure	4.0157	measuring instrument
questionnaire	3.5536	
validity	3.4549	
version	3.3308	
reliability	3.1273	
sample	2.7729	
structure	2.5713	
74		
research	30.1253	research literature
literature	3.6485	-
researcher	3.473	research
field	3.1596	investigation
area	2.9657	
issue	2.8583	
review	2.5928	
article	2.0941	
paper	1.7244	
challenge	1.4377	
75		
mouse	4.055	neuroanatomy research in mice
expression	2.6086	animal studies
brain	2.4234	animal neuroscience
cell	1.6408	device
protein	1.2435	
mechanism	1.2398	
hippocampus	1.084	
alteration	1.0115	
receptor	0.961	
role	0.9332	



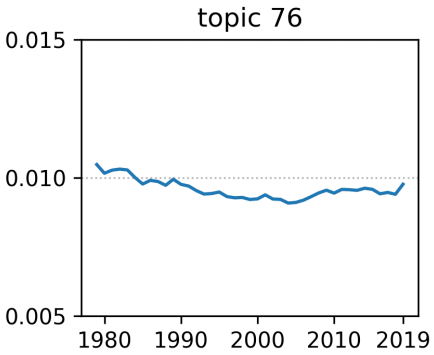


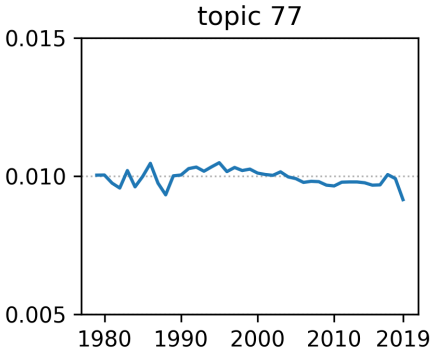


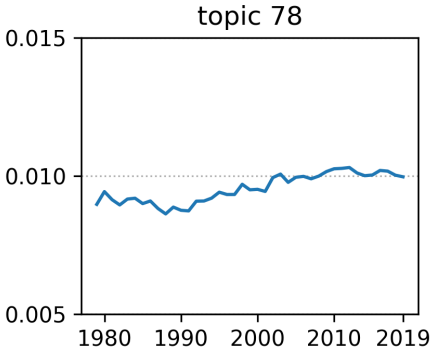


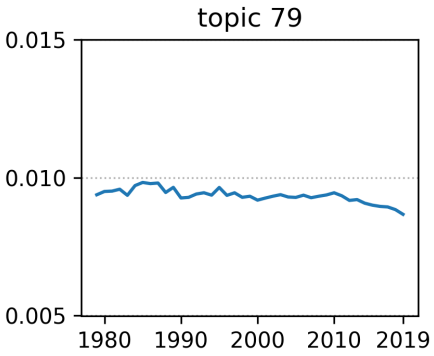


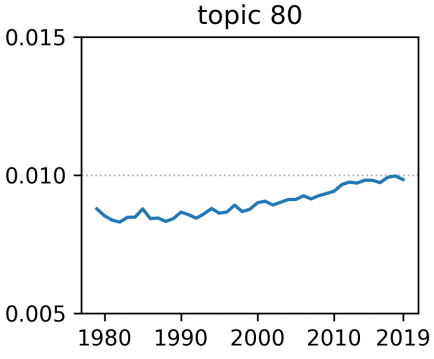
76		
interaction	13.4133	experiment design
action	11.6217	methodology - analysis or design
mechanism	7.5431	theory evidence
hypothesis	6.5583	action
evidence	3.9413	
observation	3.3833	
video	2.3055	
finding	2.1853	
generation	2.1326	
result	1.9021	
77		
family	28.92	late-life research
caregiver	11.506	research on caregivers
home	7.3075	home care
care	6.737	family
death	3.6547	
burden	3.6404	
relative	2.9931	
family_member	2.1953	
end	1.9748	
life	1.8498	
78		
community	13.9321	race studies in USA
income	3.1756	minority
culture	2.7576	afro american minority issues
discrimination	2.721	district
population	2.6372	
minority	2.5456	
race	2.4687	
member	2.039	
african_american	1.6937	
united_states	1.4331	
79		
distress	15.0421	psychological distress research
illness	13.8695	-
study	12.042	disorders
impact	3.8595	act
sample	2.8974	
mdd	2.7997	
presence	2.6101	
episode	2.3635	
questionnaire	2.14	
morbidity	2.1382	
80		
decision	10.0881	decision making studies
choice	8.2397	research on decision/choice
preference	7.4649	decision making
cost	5.2194	choice
reward	4.999	
benefit	4.7054	
decision_making	4.4374	
option	2.7817	
probability	2.5263	
uncertainty	2.1636	



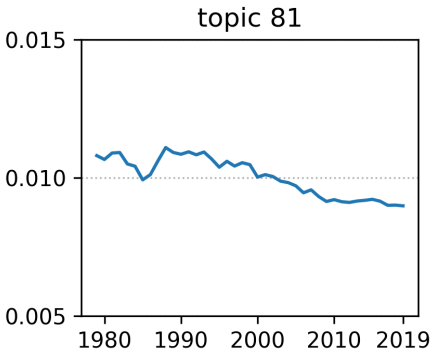


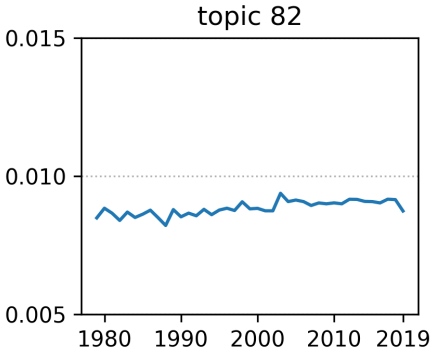


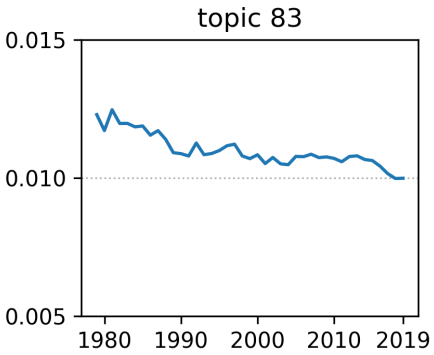


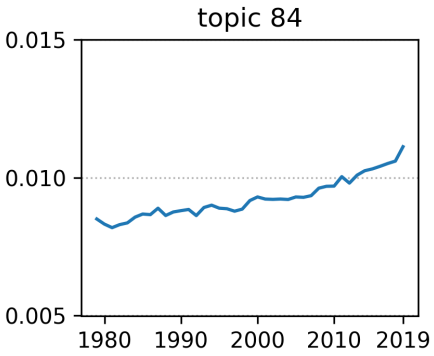


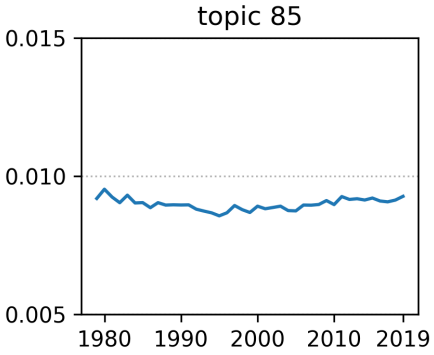
81		
day	21.6492	time descriptions
week	13.9603	methodology - time frame
period	9.4586	time intervals
session	6.3005	work time
time	5.0849	
hour	4.1448	
min	2.8615	
increase	2.6043	
reduction	2.3358	
minute	2.0374	
82		
study	9.3957	-
concern	7.827	-
reason	5.8954	-
acceptance	3.509	interest
interest	3.4836	
asd	3.4013	
center	2.5465	
donor	2.0893	
rejection	1.2249	
donation	1.1507	
83		
test	20.8804	schizophrenia studies
impairment	9.8486	Testing
deficit	9.0543	deficits effects on performance
schizophrenia	8.2777	activity
cognition	4.9057	
domain	4.5503	
performance	4.2611	
dysfunction	3.5935	
function	2.3455	
battery	2.1981	
84		
model	42.8898	regression models
datum	4.0744	methodology - analysis and prediction
analysis	3.6747	statistical modeling
prediction	3.3166	concept
transition	2.5985	
trajectory	1.9394	
path	1.798	
regression	1.7858	
data	1.4459	
variable	1.2957	
85		
condition	34.7121	-
study	10.1856	-
motivation	6.6174	particular study
result	5.5205	activity
music	2.935	
participant	2.3718	
impact	1.9665	
finding	1.7286	
pharmacist	1.5282	
skin	1.3176	



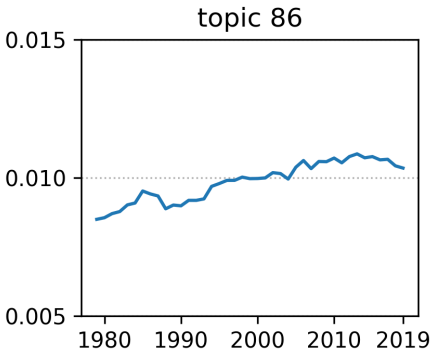


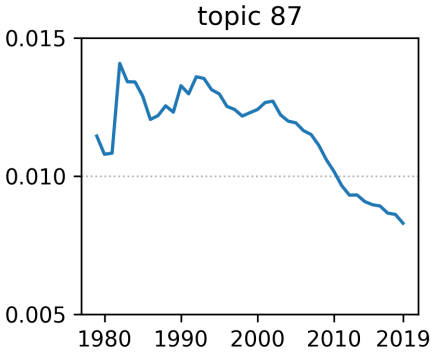


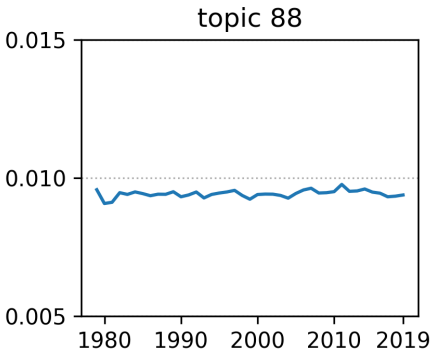


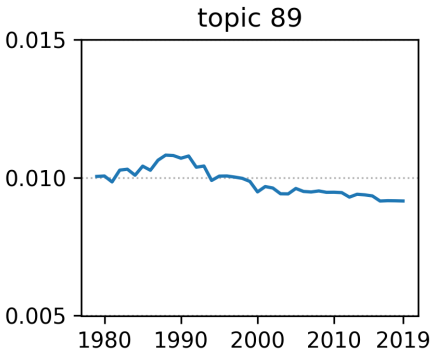


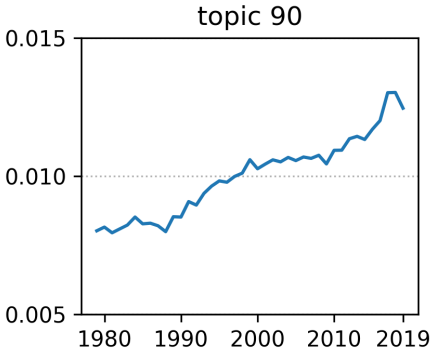
86		
symptom	48.9074	PTSD research
ptsd	8.7869	trauma symptomatology
severity	7.5696	posttraumatic stress disorder
trauma	6.412	cognition
veteran	2.2844	
exposure	1.2434	
symptomatology	1.1831	
posttraumatic_stress_disorder	1.1664	
combat	0.7688	
war	0.7567	
87		
treatment	10.5959	pharmacotherapy efficacy research
placebo	4.6078	medication treatment
week	3.7593	treatment evaluation
trial	3.1784	treatment
efficacy	2.7457	
drug	2.7192	
improvement	2.6192	
response	2.0225	
antidepressant	2.0191	
medication	1.9682	
88		
measure	27.9951	fear and avoidance studies
study	8.2836	measure of distress
report	7.4898	study bias
fear	7.4238	document
bias	6.1329	
result	4.339	
finding	3.8358	
threat	3.8161	
avoidance	3.6756	
sample	2.6698	
89		
stress	27.9171	stress response studies
response	10.1458	stress research
stressor	4.6894	stress symptoms
level	2.9846	physical phenomenon
cortisol	2.6097	
reactivity	2.5257	
increase	1.7813	
exposure	1.6481	
resistance	1.2263	
restraint	1.2253	
90		
care	15.88	health care
service	15.178	health care service
health_care	6.4669	health care service
provider	5.3767	care
barrier	4.3551	
healthcare	3.371	
access	2.9479	
professional	2.8908	
carer	1.8158	
community	1.7716	



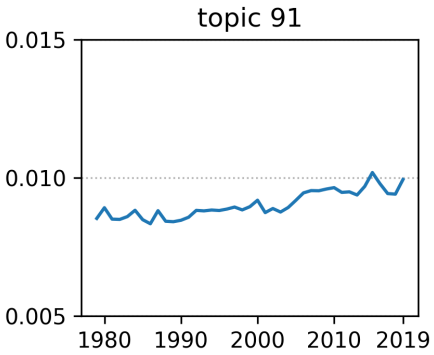


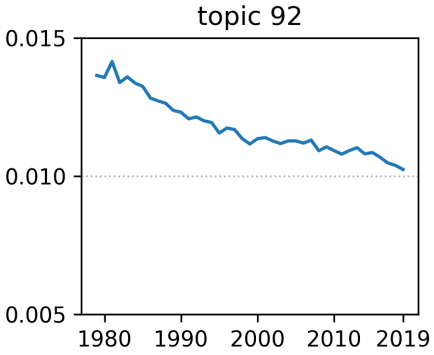


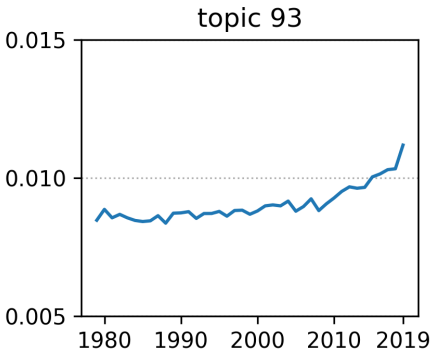


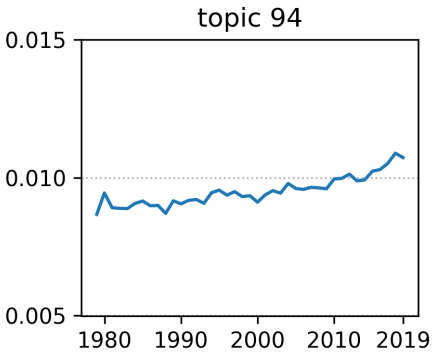


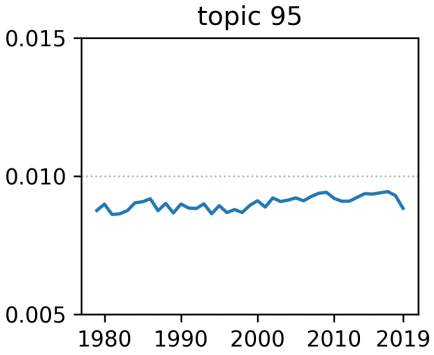
91		
study	4.8279	genetics research
testing	3.9889	-
association	3.8539	study analysis
variation	3.5416	examination
gene	3.1646	
sample	2.4012	
result	2.2873	
analysis	1.9674	
finding	1.8215	
influence	1.8204	
92		
group	79.7895	ANOVA as a method
difference	7.1285	group differences
control	6.992	group difference
comparison	2.8379	group
statistically_significant	0.5118	
experimental	0.2809	
membership	0.2245	
compare	0.1347	
anova	0.1041	
covariance	0.0956	
93		
system	13.0793	system safety
safety	4.3376	technological (system) testing
environment	4.1423	technology
technology	3.211	plan
driver	2.5147	
computer	2.0162	
device	1.9534	
design	1.7467	
user	1.6666	
application	1.6397	
94		
work	19.3335	work environment research
worker	4.9535	research about organisations
job	3.5965	work evaluation
employment	2.6624	employment
employee	2.5562	
organization	2.5492	
resilience	2.511	
burnout	2.3002	
environment	2.2663	
workplace	1.9914	
95		
management	11.6959	chronic disease management
medication	8.0084	medication management
adherence	7.7925	-
epilepsy	4.1434	act
diabetes	3.6504	
patient	3.0349	
study	2.9559	
seizure	2.548	
hypertension	2.2263	
disease	2.0542	



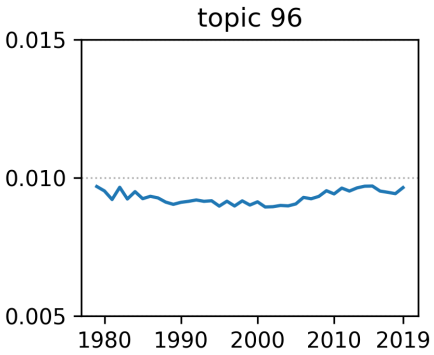


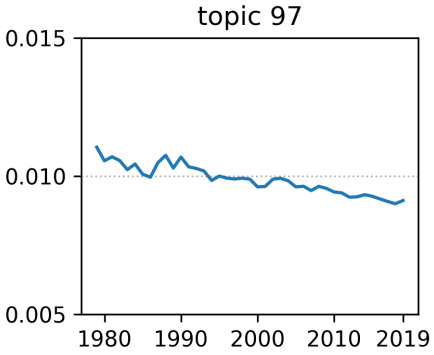


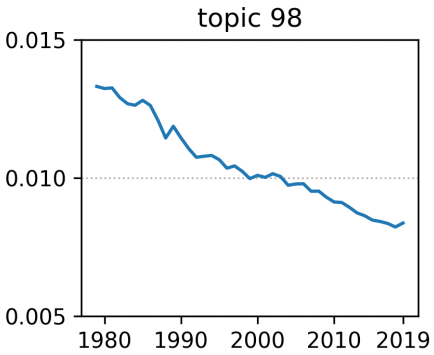


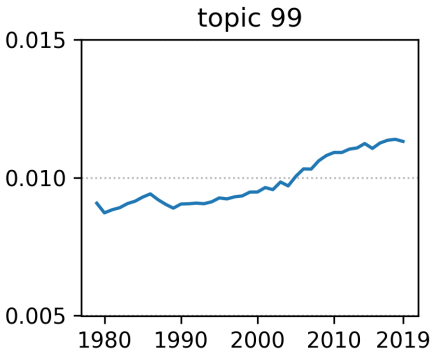


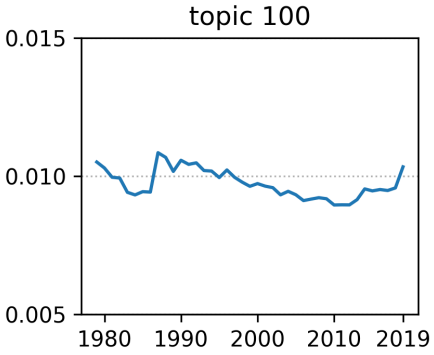
96		
control	40.4723	using control group in study design
individual	25.8775	control group studies
goal	7.7109	control study
study	6.1025	thinking
regulation	3.0454	
result	2.1888	
finding	2.1413	
group	1.2928	
flexibility	0.9526	
participant	0.8468	
97		
difference	30.6556	gender differences in research
gender	13.8128	gender differences
sex	12.5674	gender difference
male	7.9252	act
female	5.7428	
study	2.9762	
age	2.8217	
sample	1.6924	
males_females	1.6045	
ibs	1.114	
98		
event	15.5202	cognitive research
state	12.859	-
situation	7.177	-
author	4.9968	physical phenomenon
reaction	4.9875	
experience	3.3741	
feeling	3.2349	
thought	2.4641	
consequence	2.0956	
attribution	1.506	
99		
association	21.1292	self esteem study design
variable	12.4974	prediction and association
predictor	8.7085	statistical analysis
analysis	6.326	process
characteristic	3.2078	
self_esteem	3.0899	
factor	2.9061	
logistic_regression	2.4324	
cross_sectional	2.3833	
status	2.0121	
100		
frequency	11.8548	EEG studies
power	3.4122	brainwave research
variability	2.9908	EEG study
threshold	2.8262	relation
parameter	2.8224	
eeg	2.5794	
distribution	2.3962	
signal	1.9547	
noise	1.9101	
range	1.6642	











B Source code

The source code for the steps described in this thesis can be accessed from this address:
`https://github.com/ottmartens/psy-topic-models`

C Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Ott-Kaarel Martens**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Evolution of Topics in the Psychology Domain,

   supervised by Eduard Barbu.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Ott-Kaarel Martens

08.05.2020