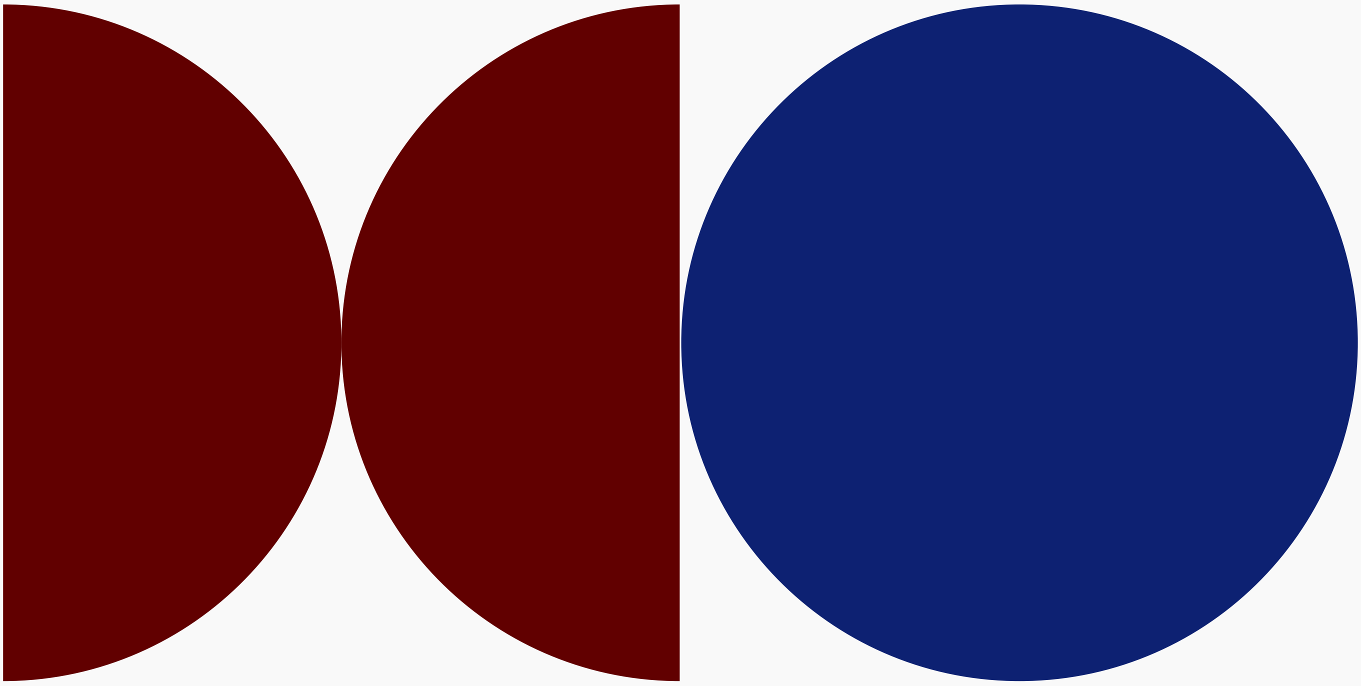
Swedish
Research
Council

Appendix

Appendix to research review 2023

Artistic research

Summary of the results from automated topic modelling



2023

Appendix

Research review 2023: Artistic research.
Summary of the results from automated topic modelling

Dnr 3.5-2022-05069
ISBN 978-91-88943-82-8

Swedish Research Council
Vetenskapsrådet
Box 1035
SE-101 38 Stockholm, Sweden

Contents

Summary of the results from automated topic modelling based on applications received for undirected project grants in artistic research

Introduction

In its work on the research review, the Committee for Artistic Research initially relied on the so-called ‘data package’, the statistical documentation compiled annually from applications received. The statistics provides valuable indications regarding factors such as the number of applications, gender distribution, grant amounts and so on. As the Swedish Research Council is the only funding body for artistic research in Sweden, the applications present a material that could potentially tell a lot about the conditions and changes of artistic research. Considering that this material, although primarily in writing, has been compiled by active researchers, the Committee wanted to explore its potential as a starting point for a more in-depth quantitative-qualitative analysis.

An automated topic modelling, or what is known as a ‘text mining analysis’ was therefore carried out.[1] A pilot project within a larger framework aimed at investigating the opportunities for developing an in-house script for automated topic modelling, the analysis of artistic research applications also offered an opportunity to train personnel at the Swedish Research Council to use this tool for complementary documentation in the work on research reviews and research agendas.

This appendix describes the process in brief, with the underlying method presented in a further appendix.

Method description

Methods and algorithms for automated text analysis, or ‘text mining’, have been developed for several different purposes and applications. In general, these methods are based on different ways of converting text collections into computable, numeric representations, from which analytical support can be further developed, such as text classification, searches and exploration. There is an abundance of different methods, and new methods are constantly being developed. To a large extent, the methods are based on machine learning. A primary division of different types of methods can be made into those that are “monitored” and “unmonitored”. Monitored methods are those that require the algorithm to be ‘taught’ using example data of what you want the algorithm to

[1] The analysis was conducted by senior research officers Justiina Dahl (Department of Research Infrastructures), and Sumithra Velupillai (Strategic Data Resources Unit).

do automatically, while unmonitored methods are those that identify patterns or descriptions of data without being provided with any examples.

Topic modelling is an example of an unmonitored method. The idea behind topic modelling is to automatically identify thematic or topic-related groupings in text collections, using different computation methods. They can act as support in exploring the content of large text collections, such as abstracts in all project applications to a research funding body. The algorithms generate different topics/themes from the text collections by computing the similarity between word occurrences in documents in the text collections. In brief, it is about reading a text collection and dividing up the text into sentences and words, based on a language model, and then creating a matrix where the distribution of words in the entire text collection is computed and represented (known as 'bag-of-words'). Each word in the entire text collection is counted and given a frequency value for each document.

When the bag-of-words matrix has been created, the topic modelling is then applied. In the pilot project, we used an algorithm known as NMF (Non-Negative Matrix Factorisation). NMF is a method based on linear algebra, where a multi-dimensional matrix is broken down into smaller matrices while retaining the essence of the original matrix. In this way, latent structures are captured; that is, the topics/themes that are represented in the text collection. When using NMF for topic modelling of a text collection, a representation is created that shows which topics are included in each document, and what words best represent the different topics generated from the text collection. You have to state how many topics you want to generate from a text collection. As this is an unmonitored method, there is no correct answer to what number is the most suitable; instead, this has to be explored iteratively.

In other words, it is a combination of quantitative and qualitative elements, which in this case were based on a preliminary analysis of titles, abstracts and key words in English from 441 of a total of 476 applications from the period 2014–2021. Pre-processing of data was carried out, where 'stop words' (function words such as 'and', 'is', 'of' and so on) were filtered out from all abstracts. Thereafter, the collection of abstracts was converted into a representation, on which the topic modelling algorithm was applied, where the number of clusters ('topics') the model was to generate was stated.

Initial modelling was carried out based on 30 and 40 clusters respectively, and a decision was then made to continue the analysis based on the lower number. A 0 cluster containing very general words was then filtered out, after which the material went through a further modelling based on 20 clusters or topics, as shown in Table 1 below.

Table 1. Result of topic modelling based on 20 clusters, described based on the 10 most frequently used words.

Clusters	**Number of applications**
0: public, space, medium, place, production, site, exhibition, specific, engage, result	26
1: music, musical, different, create, possibility, exploration, creation, experimental, creative, contribute	34
2: film, role, create, future, aesthetic, people, interview, produce, story, result	28
3: writing, text, write, workshop, experience, main, theoretical, emerge, consist, en*	22
4: human, environment, relation, life, create, challenge, relationship, exhibition, propose, bring	17
5: performance, perform, audience, theory, model, result, combine, group, strategy, act	17
6: sound, live, environment, approach, collaboration, aspect, affect, relationship, space, creation	19
7: image, medium, visual, want, examine, discussion, act, series, engage, representation	19
8: technology, digital, open, system, medium, development, possibility, interdisciplinary, framework, life	23
9: visual, political, critical, aesthetic, strategy, issue, methodology, open, context, dialogue	56
10: narrative, creative, medium, investigate, theory, important, individual, create, story, different	13

Clusters	**Number of applications**
11: interaction, aesthetic, purpose, conduct, investigate, system, expression, experience, perspective, space	12
12: history, historical, present, exhibition, experience, place, representation, contemporary, challenge, Swedish	27
13: architecture, environment, model, build, structure, object, development, exhibition, making, generate	19
14: cultural, different, expression, life, provide, establish, area, culture, creation, idea	14
15: book, concept, exhibition, theory, workshop, expand, text, present, theoretical, publication	19
16: time, world, live, want, researcher, experimental, life, en*, relationship, present	13
17: body, place, movement, experience, collaboration, investigate, workshop, relation, affect, want	23
18: society, performative, analysis, production, contemporary, perform, collaboration, Swedish, create, understanding	14
19: change, experience, future, people, story, society, involve, approach, public, culture	24

* overlooked ‘stop word’

The results from topic modelling therefore produces a new type of overall picture of the entire text collection and the topics that are represented in it. The matrix also includes weights for each document included in each topic, and for the words that best represent each topic. Using these, it is therefore possible to analyse the text collection in different ways and based on different numbers of “gravity points.”

In the pilot project, we created a graph based on the documents given the greatest weight in a topic and showed the ten most highly ranked words for each topic (Figure 1). This shows the distribution of the number of documents based on a pre-set number of topics and provides a picture of what each topic is about, based on the most frequently used words. With input from a subject expert or similar it is then possible to further boil down or interpret the ten-word clusters, or analyse and describe the result in another way.

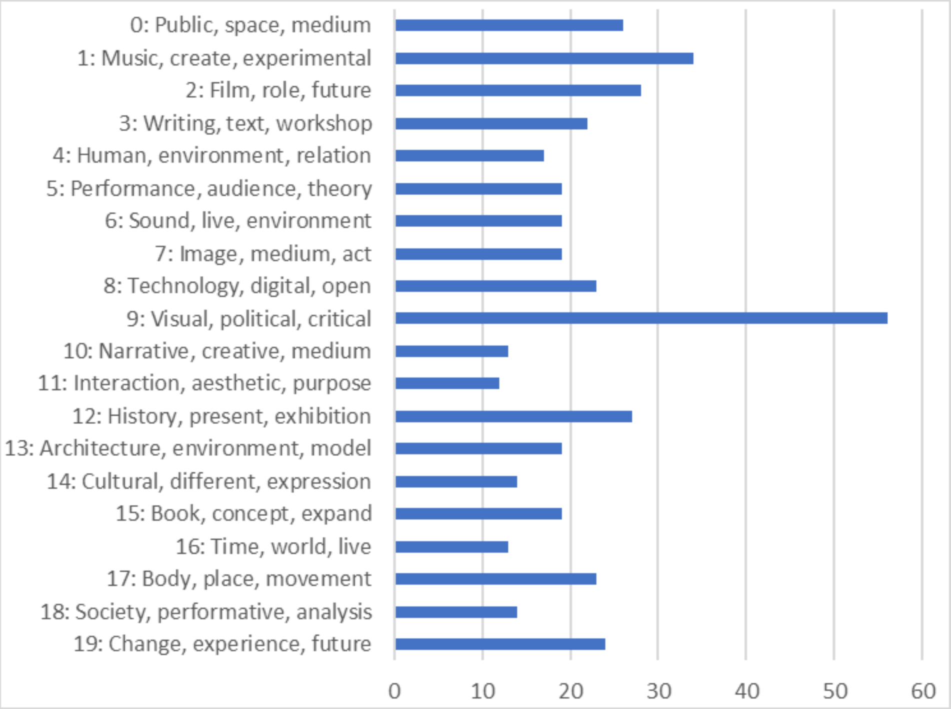


Figure 1. Number of applications per theme following topic modelling based on English abstracts in applications, 2014–2021. The number of topics has been set at 20, and each topic is shown on the Y axis (represented by the 3 most important of the 10 words given the greatest weight for the topic). The X axis shows the number of applications with the highest value in each topic (only one value per application). A document in the text collection can belong to several topics, which is not captured in the figure above. All the words included in each theme are shown in Table 1 above.

Further analysis of the result

The result of the topic modelling was presented and discussed during two hearings in March 2022. The comments made related in particular to the difficulty of doing justice to an area that is as varied and changeable as artistic research. The limitations of basing the review on written material were also pointed out. There is a risk, some researchers considered, that the most central concepts, which are almost taken for granted, do not emerge, as they are of a visual, material or other nature. Others, on the contrary, saw a thematic and experience-based depth in the clusters that has sometimes been disregarded in strongly output-oriented or result-oriented exploration.

A possible interpretation of the result of the topic modelling is that form, medium and content are aspects that are actively connected in artistic research in ways that do not follow the division into artistic genres. This is an interpretation that finds support through comparison with the SCB/Statistics Sweden codes stated in the applications clustered together. The comparative assessment could to some extent verify the existence of cross-sectorial themes and research questions that are shared by researchers from different artistic practice areas. It showed, for example, that even if visual arts is the area that has the highest approval rate based on the SCB/Statistics Sweden coding and therefore can be regarded as the dominant area, it is a thematically and methodologically varied field that to a large extent includes subjects with other SCB/Statistics Sweden codes, such as filmic expression, potentially focused towards documentary story-telling. In other cases, it can be noted that a theme, such as the one including among others the concepts "history" and "experience," was processed based on several of the SCB/Statistics Sweden-coded subjects. When it comes to research focused on the problems of spatiality, the analysis also indicated that this is a theme that, by reference to concepts such as "public," "space" and "place," is developing across the artistic research field. Some exceptions also emerged, in particular the cluster around the concept of "music," where the 34 applications were exclusively coded simply as "music."

Table 2. Some examples of topics in relation to the SCB/Statistics Sweden codes stated in the application. Descriptive words of theme and relationship to first stated research topic/SCB/Statistics Sweden-code in the application.

Table 2 a - Theme 2: film, role, create, future, aesthetic, people, interview, produce, story, result

Research topic	**Number of applications**
Visual arts	16
Film studies	7
Architecture	1
History	1
Music	1
Dramatic art	1
Other humanities	1
Total number of applications	**28**

Table 2 b - Theme 12: history, historical, present, exhibition, experience, place, representation, contemporary, challenge, Swedish

Research topic	**Number of applications**
Visual arts	21
Design	3
Dramatic art	1
Social anthropology	1
Other humanities	1
Total number of applications	**27**

Table 2 c - Theme 0: public, space, medium, place, production, site, exhibition, specific, engage, result

Research topic	**Number of applications**
Visual arts	7
Architecture	6
Design	4
Dramatic art	4
Film studies	3
Microbiology	1
Other technology	1
Total number of applications	**26**

It is important to note that there are other common algorithms for topic modelling based on other computation methods. One of the most used is LDA (Latent Dirichlet Allocation), which is probabilistic (based on probability computations). The reason why NMF was used in the pilot project is that it can work better for small text collections. LDA and other probabilistic methods generally work better for larger text collections.

There are also other ways of presenting and visualising results from topic modelling, including with interactive interfaces, where users can click their way through and explore the content of the topic model. One example of such a solution is LDAViz, but it is based on creating a topic model using the LDA algorithm.

Technical details

The script produced is written using the programming language Python (version 3.8.8), with open-source code and in Jupyter Notebook. The Excel file that contains the collection of abstracts from which topic models are to be created are scanned using the Pandas package (version 1.2.4) for content processing. The texts have been language classified, divided up into words and converted into basic format (known as lemmata), using the package SpaCy (version 3.2.1) and the English SpaCy language model “en_core_web_sm.” Pre-defined English stop words from the package nltk (version 3.6.1) are used in combination with an external Excel file to filter commonly used non-content-carrying words. The topic models are generated using the package Sklearn (version 0.24.1): CountVectorizer, TfidfTransformer (for the data representation) and NMF (for the topic model). A result file in Excel is also created in the script, which can be used to analyse the results for each document in more detail.

Swedish
Research Council