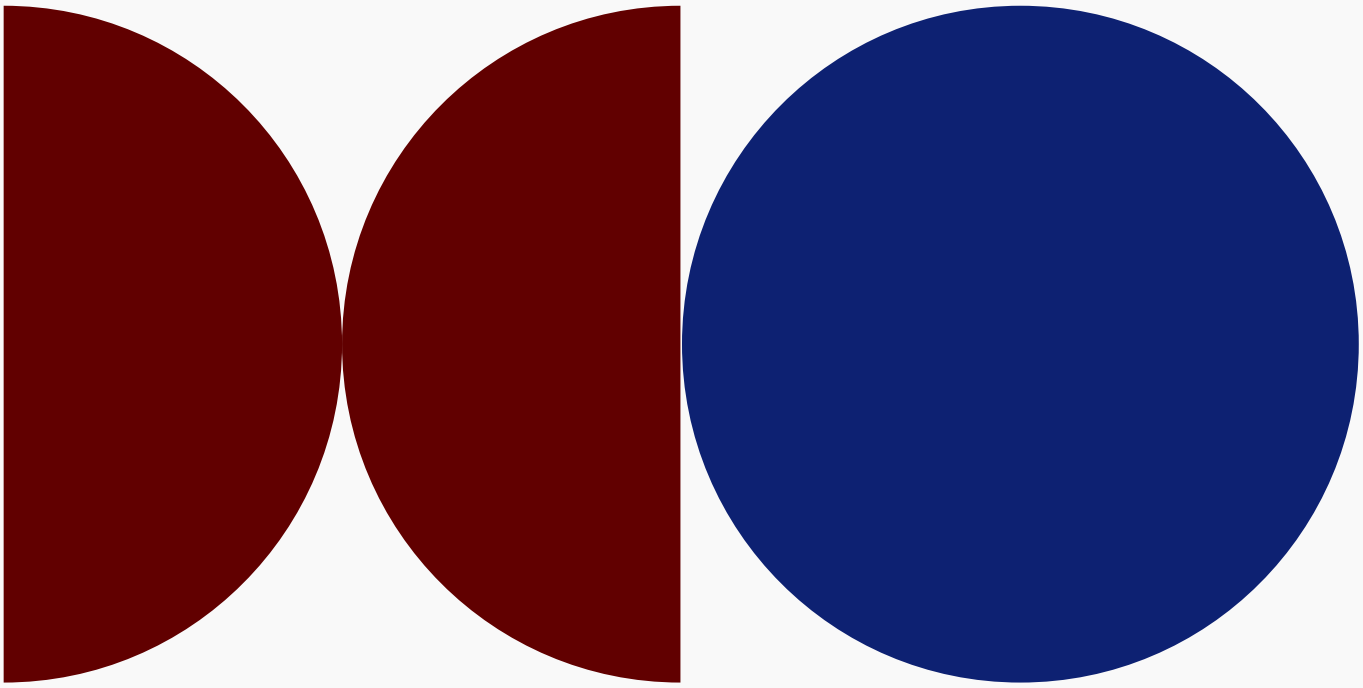
Vetenskapsrådet

Bilaga

Bilaga till forskningsöversikt 2023

Konstnärlig forskning

Sammanfattning av resultat från ämnesmodellering



2023

Bilaga

Forskningsöversikt 2023: Konstnärlig forskning. Sammanfattning av resultat från automatisk ämnesmodellering.

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

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

Innehållsförteckning

Sammanfattning av resultat från automatisk ämnesmodellering utifrån inkomna ansökningar om fria projektbidrag inom konstnärlig forskning 4

Inledning 4
Metodbeskrivning 4
Vidare analys av resultatet 8
Tekniska detaljer 11

Sammanfattning av resultat från automatisk ämnesmodellering utifrån inkomna ansökningar om fria projektbidrag inom konstnärlig forskning

Inledning

I arbetet med forskningsöversikten utgick Kommittén för konstnärlig forskning inledningsvis från det så kallade datapaketet, det statistiska underlag som årligen sammanställs utifrån inkomna ansökningar. Här ges värdefulla indikationer gällande söktryck, könsfördelning, bidragsbelopp med mera. Som enda finansiär av konstnärlig forskning i Sverige förfogar Vetenskapsrådet genom ansökningarna över ett material som potentiellt sett skulle kunna säga mycket om den konstnärliga forskningens förutsättningar och förändring. Kommittén menade att det är ett material, förvisso huvudsakligen skriftligt, som är sammanställt av de aktiva forskarna själva, och som sådant en möjlig utgångspunkt för en fördjupad kvantitativ-kvalitativ analys.

Därför genomfördes en automatisk ämnesmodellering (eng. ”topic modelling”) eller en så kallad textmining-analys[1]. Analysen utvecklades inom ramen för ett pilotprojekt med syftet att undersöka möjligheterna att utveckla ett egenutvecklat skript för automatiserad ämnesmodellering samt att utbilda Vetenskapsrådets personal i att använda detta. Genom ett fokus på ämnesmodellering skulle pilotprojektet också kunna leda till att ta fram kompletterande underlag i arbetet med forskningsöversikter och forskningsagendor. I denna bilaga beskrivs kortfattat processen och i ett appendix den underliggande metoden.

Metodbeskrivning

Metoder och algoritmer för automatiserad textanalys, så kallad text-mining, har utvecklats för flera olika syften och tillämpningsområden. Generellt bygger dessa metoder på olika sätt att omvandla textsamlingar till beräkningsbara, numeriska representationer från vilka man kan vidareutveckla analytiska stöd som exempelvis textklassificering, sökning och utforskning. Det finns en uppsjö av olika metoder, och utvecklingen av nya metoder pågår konstant. Till stor del bygger metoderna på maskininlärning. En huvuduppdelning mellan typer av metoder kan beskrivas som sådana som är "övervakade" och "oövervakade". Med övervakade metoder menas sådana som kräver att man lär algoritmen med exempeldata på det man vill att algoritmen ska kunna göra automatiskt; med oövervakade metoder menas sådana som identifierar mönster eller beskrivningar av data utan att den ges några exempel.

[1] Analysen genomfördes av forskningssekreterare Justiina Dahl (Avdelningen för forskningsinfrastruktur) och Sumithra Velupillai (Enheten för data som strategisk resurs).

Ämnesmodellering (eng: topic modelling) är ett exempel på oövervakade metoder. Tanken med ämnesmodellering är att automatiskt identifiera tematiska eller ämnesmässiga grupperingar i textsamlingar genom olika beräkningsprocesser. De kan fungera som ett stöd i att utforska innehållet i stora textsamlingar, till exempel abstract i alla projektansökningar hos en forskningsfinansiär. Algoritmerna genererar olika ämnen/teman från textsamlingarna genom att beräkna likhet mellan ordförekomster mellan dokument i textsamlingarna. Kortfattat handlar det om att läsa in en textsamling och utifrån en språkmodell dela upp texten i meningar och ord, för att sedan skapa en matris där fördelningen av ord i hela textsamlingen beräknas och representeras (så kallad bag-of-words). Varje ord i hela textsamlingen räknas och ges ett frekvensvärde för varje dokument.

När bag-of-words-matrisen är skapad tillämpas sedan ämnesmodelleringen. I pilotprojektet användes en algoritm som heter NMF (Non-Negative Matrix Factorization). NMF är en metod som bygger på linjär algebra där en multidimensionell matris bryts ner till mindre matriser, men som bibehåller essensen i ursprungsmatrisen. Därigenom fångas latenta strukturer, det vill säga de ämnen/teman som representeras i textsamlingen. När man tillämpar NMF för ämnesmodellering från en textsamling skapas en representation som visar vilka ämnen som finns i varje dokument och vilka ord som bäst representerar de olika ämnena som genererats från textsamlingen. Man måste ange hur många ämnen man vill generera från en textsamling. Eftersom detta är en oövervakad metod finns det inget rätt svar på vilket antal som är lämpligast, utan detta måste utforskas iterativt.

Det är alltså fråga om en kombination av kvantitativa och kvalitativa moment, som i det aktuella fallet tog sin utgångspunkt en preliminär analys av titlar, abstracts och nyckelord på engelska från 441 av totalt 476 ansökningar från perioden 2014–2021. En förbearbetning av data genomfördes där så kallade ”stoppord” (funktionsord såsom ”and”, ”is”, ”of” etcetera) filtrerades bort från alla abstracts. Därefter omvandlades samlingen av abstracts till en representation på vilken ämnesmodelleringsalgoritmen applicerades där det antal kluster (”topics”) som modellen skulle generera angavs.

En första modellering genomfördes utifrån 30 respektive 40 kluster, varpå beslutades att analysera vidare utifrån det lägre antalet. Ett kluster 0 innehållande väldigt generella ord filtrerades sedan bort, varefter materialet genomgick ytterligare en modellering utifrån 20 kluster eller ämnen som redovisas i tabell 1 nedan.

Tabell 1. Resultat av ämnesmodellering utifrån 20 kluster redovisade utifrån de 10 mest förekommande orden.

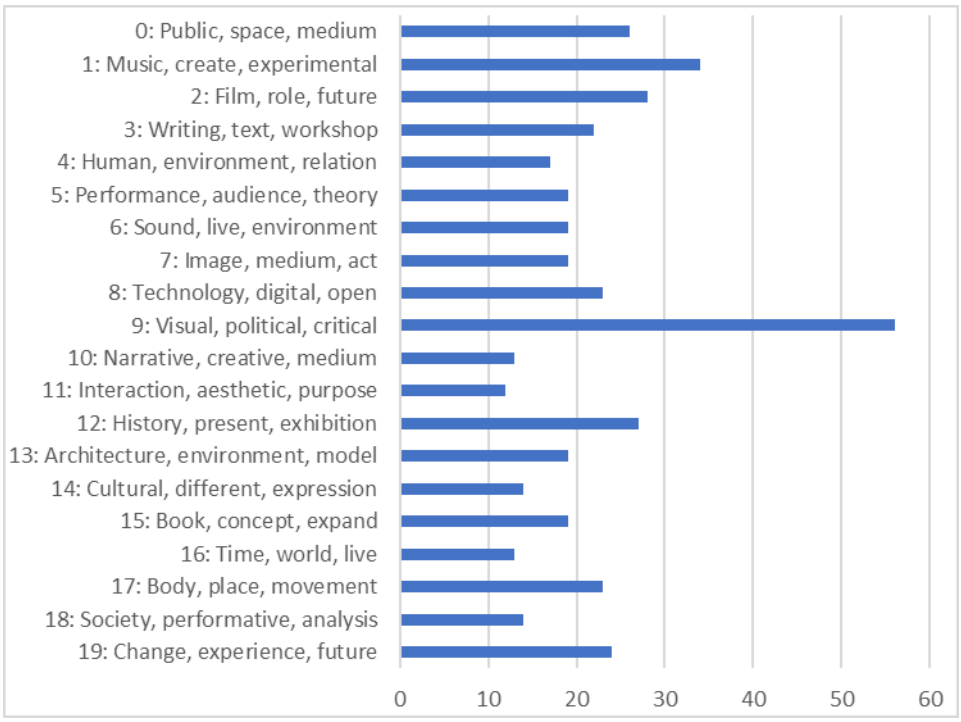
Kluster	**Antal ansökningar**
0: public, space, medium, place, production, site, exhibition, specific, engage, result	26
1: music, musical, different, create, possibility, exploration, creation, experimental, creative, contribute	34
2: film, role, create, future, aesthetic, people, interview, produce, 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
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

Kluster	Antal ansökningar
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

*kvarglömt "stoppord"

Resultaten från en ämnesmodell ger alltså ett nytt slags helhetsbild av hela textsamlingen och de ämnen som finns representerade där. Matrisen innehåller också vikter för varje dokument som ingår i varje ämne och för de ord som bäst representerar varje ämne. Med dessa kan man alltså analysera textsamlingen på olika sätt och utifrån olika antal "gravitationspunkter".

Något som använts i pilotprojektet är skapandet av en graf utifrån de dokument som fått högst vikt i ett ämne, och visa de tio högst rankade orden för varje ämne (Figur 1). Det visar fördelningen av antal dokument utifrån ett förbestämt antal ämnen, och ger genom de vanligast förekommande orden en bild av vad varje ämne handlar om. Resultatet kan sedan genom input från t.ex. en ämneskunnig kokas ner till ett övergripande tema istället för de tio orden, eller på annat sätt analyseras och beskrivas.



Figur 1. Antal ansökningar per tema efter ämnesmodellering utifrån engelska abstracts i ansökningar 2014–2021. Antalet ämnen har satts till 20, och varje ämne visas på y-axeln (representerat med de tre viktigaste av de 10 ord som fått högst vikt i det ämnet). På x-axeln visas antalet ansökningar som har högst värde i respektive ämne (endast ett värde per ansökan). Ett dokument i textsamlingen kan tillhöra flera ämnen, vilket inte fångas i den ovanstående grafen Alla ord som ingår i varje tema finns i tabell 1 ovan.

Vidare analys av resultatet

Resultatet av ämnesmodelleringen presenterades och diskuterades under de två hearingarna som anordnades i mars 2022, som ett led i arbetet med forskningsöversikten. De kommentarer som gavs gällde framför allt svårigheten att göra ett så varierat och föränderligt område som konstnärlig forskning rättvisa. Dessutom påtalades begränsningarna i att utgå ifrån ett skriftligt material. Risken finns, menade vissa forskare, att de allra mest centrala, närmast förgivettagna föreställningarna inte framträder då de är av visuell, materiell eller annan natur. Andra såg tvärtom i klustren ett tematiskt och erfarenhetsbaserat djup som ibland förbisetts i ett starkt verks- eller resultatorienterat utforskande.

En möjlig tolkning av resultatet av ämnesmodelleringen är att form, medium och innehåll är aspekter som aktivt kopplas samman inom konstnärlig forskning på ett sätt som inte följer uppdelningen i konstnärliga genrer. Det är en tolkning som finner stöd genom en jämförelse med de SCB-koder som angetts i de ansökningar som klustrats tillsammans. Den jämförande granskningen kunde till viss del verifiera förekomsten av tvärgående teman och forskningsfrågor gemensamma för forskare från olika konstnärliga praktikområden.

Här framgår exempelvis att även om bildkonst är det område som utifrån SCB-kodningen har den högsta beviljandegraden och därmed kan betraktas som

dominerande, så är det tematiskt och metodmässigt ett mycket varierat fält som i hög grad inbegriper ämnen med andra SCB-koder, exempelvis filmisk gestaltning, potentiellt med inriktning mot dokumentärt berättande. I andra fall kan noteras att en tematik som den som innefattade begrepp som ”history” och ”experience” behandlades utifrån flera av de SCB-kodade ämnena. När det gäller forskning med inriktning mot rumslighetens problematik tydde analysen också på att det är ett tema som – genom begrepp som "public", "space" och "place"– utvecklas på tvärs av det konstnärliga forskningsfältet. Vissa undantag framträdde också, framför allt av klustret runt begreppet ”music” där de 34 ansökningarna uteslutande hade SCB-kodats som just musik.

Tabell 2. Några exempel på ämnen i förhållande till de SCB-koder som uppgetts i ansökan. Beskrivande ord till tema och relation till första angivna forskningsämne/SCB-kod i ansökan.

Tabell 2 a – Tema 2: film, role, create, future, aesthetic, people, interview, produce, story, result

Forskningsämne	Antal ansökningar
Bildkonst	16
Filmvetenskap	7
Arkitektur	1
Historia	1
Musik	1
Scenkonst	1
Övrig annan humaniora	1
Totalt antal ansökningar	**28**

Tabell 2 b – Tema 12: history, historical, present, exhibition, experience, place, representation, contemporary, challenge, Swedish

Forskningsämne	Antal ansökningar
Bildkonst	21
Design	3
Scenkonst	1
Socialantropologi	1
Övrig annan humaniora	1
Totalt antal ansökningar	**27**

Tabell 2 c – Tema 0: public, space, medium, place, production, site, exhibition, specific, engage, result

Forskningsämne	Antal ansökningar
Bildkonst	7
Arkitektur	6
Design	4
Scenkonst	4
Filmvetenskap	3
Mikrobiologi	1
Övrig annan teknik	1
Totalt antal ansökningar	**26**

Observera att det finns andra vanliga algoritmer för ämnesmodellering som bygger på andra beräkningsmetoder. En av de mest använda är LDA (Latent Dirichlet Allocation), som är probabilistisk (bygger på sannolikhetsberäkningar). Anledningen till att NMF användes i pilotprojektet är att den kan fungera bättre på små textsamlingar. LDA och andra probabilistiska metoder fungerar generellt bättre på större textsamlingar.

Det finns också andra sätt att presentera och visualisera resultat från ämnesmodeller, även med interaktiva gränssnitt där man som användare kan klicka sig fram och utforska innehållet i ämnesmodellen. Ett exempel på en sådan lösning är LDAViz, men den bygger på att man skapat en ämnesmodell med LDA-algoritmen.

Tekniska detaljer

Skriptet som tagits fram är skrivet i programmeringsspråket python (version 3.8.8) med öppen källkod och i Jupyter Notebook. Excel-filen som innehåller den samling abstracts man vill skapa ämnesmodeller från, läses in med paketet pandas (version 1.2.4) för innehållsbearbetning. Texterna har språkklassificerats, delats upp i ord och omvandlats till grundform (så kallade lemman) med paketet SpaCy (version 3.2.1) och den engelska SpaCy-språkmodellen ”en_core_web_sm”. Fördefinierade engelska stoppord från paketet nltk (version 3.6.1) används i kombination med en extern Excel-fil för att filtrera vanlig förekommande icke-innehållsbärande ord. Ämnesmodellerna genereras med paketet sklearn (version 0.24.1): CountVectorizer, TfidfTransformer (för datarepresentationen) och NMF (för ämnesmodellen). En resultatfil i Excel skapas också i skriptet som kan användas för att analysera resultaten för varje dokument mer i detalj.

Vetenskapsrådet