Week 4
Topic Modeling

June 7, 2020

Overview

Outline

What's TM about?

- TM is an **unsupervised machine learning** technique for abstracting hidden topics (i.e., themes) fromcollections of documents.
- TM components:
  - Topics are represented as the probability that each of a given *set of words* will occur
  - Documents are represented as a mixture of topics
- Words can be associated with multiple topics

Methodological roots of TM

Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: John Lafferty

Abstract

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA 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. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

Extension of the original model

Dynamic Topic Models

David M. Blei BLEI@CS.PRINCETON.EDU
Computer Science Department, Princeton University, Princeton, NJ 08544, USA

John D. Lafferty LAFFERTY@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213, USA

Abstract

A family of probabilistic time series models is developed to analyze the time evolution of topics in large document collections. The approach is to use state space models on the natural parameters of the multinomial distributions that represent the topics. Variational approximations based on Kalman filters and nonparametric wavelet regression are developed to carry out approximate posterior inference over the latent topics. In addition to giving quantitative, predictive models of a sequential corpus, dynamic topic models provide a qualitative window into the contents of a large document collection. The models are demonstrated by analyzing the OCR'ed archives of the journal *Science* from 1880 through 2000.

ment are assumed to be independently drawn from a mixture of multinomials. The mixing proportions are randomly drawn for each document; the mixture components, or topics, are shared by all documents. Thus, each document reflects the components with different proportions. These models are a powerful method of dimensionality reduction for large collections of unstructured documents. Moreover, posterior inference at the document level is useful for information retrieval, classification, and topic-directed browsing.

Treating words exchangeably is a simplification that it is consistent with the goal of identifying the semantic themes within each document. For many collections of interest, however, the implicit assumption of exchangeable *documents* is inappropriate. Document collections such as scholarly journals, email, news articles, and search query

Outline

TM process

© *Academy of Management Annals*
2019, Vol. 13, No. 2, 586–632.
https://doi.org/10.5465/annals.2017.0099

TOPIC MODELING IN MANAGEMENT RESEARCH: RENDERING NEW THEORY FROM TEXTUAL DATA

TIMOTHY R. HANNIGAN
University of Alberta

RICHARD F. J. HAANS
Erasmus University

KEYVAN VAKILI
London Business School

HOVIG TCHALIAN
Claremont Graduate University

VERN L. GLASER
MILO SHAOQING WANG
University of Alberta

SARAH KAPLAN
University of Toronto

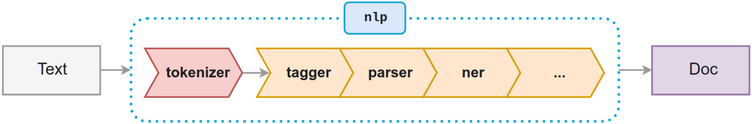
P. DEVEREAUX JENNINGS[1]
University of Alberta

TM process

Source: Adapted from Hannigan et al. 2019 (AMA)

Rendering of corpora

Before TM, the text MUST traverse an NLP pipeline.



Source: spaCy project website.

Pay attention:

- The outcome of the NLP pipeline is a function of the statistical model of the natural language adopted by the researcher.
- Picking-up the right model requires substantial institutional knowledge
- Standard models of the language may not fit the data at hand — this can jeopardize the validity of topic modelling estimates.

Rendering constructs

In this step, analysts make sense of the mixture of words that have been discovered.
Possible activities are:

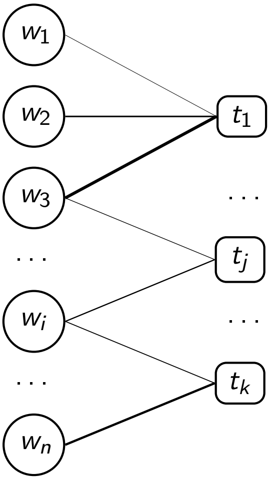
- attaching meaningful labels to mixtures of words (e.g., 'ML', 'DL', 'analytics', 'HR', 'people', 'management' words that co-occur in a same topic seems related to a 'people analytics' topic)
- appreciating how topics map onto documents with different attributes (e.g., 'old documents', 'new documents', documents concerning domestic vs global firms, documents concerning large vs small companies)
- appreciating how topics are connected — that is, exploring the topology of topics

Outline

TM design — rendering of topics

Rendering of topics equates to creating a graph connecting the words observed in a corpus of text with some latent variables.

Since the number of words is given, the critical choice concerns the number of latent variables to retain, i.e., the topics that are supposed to generate the observed documents.



TM design — rendering of artifacts

The evaluation of topic models is a slippery terrain:

- plethora of approaches and metrics
- fluid conversation involving staticians, data scientists, and business/financial analysts

TM design — reading tea leaves?

Reading Tea Leaves: How Humans Interpret Topic Models

Jonathan Chang *
Facebook
1601 S California Ave.
Palo Alto, CA 94304
jonchang@facebook.com

Jordan Boyd-Graber *
Institute for Advanced Computer Studies
University of Maryland
jbg@umiacs.umd.edu

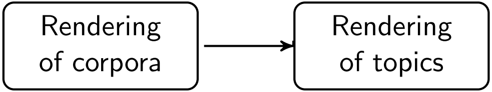
Sean Gerrish, Chong Wang, David M. Blei
Department of Computer Science
Princeton University
{sgerrish,chongw,blei}@cs.princeton.edu

Abstract

Probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus. Practitioners typically assume that the latent space is semantically meaningful. It is used to check models, summarize the corpus, and guide exploration of its contents. However, whether the latent space is interpretable is in need of quantitative evaluation. In this paper, we present new quantitative methods for measuring semantic meaning in inferred topics. We back these measures with large-scale user studies, showing that they capture aspects of the model that are undetected by previous measures of model quality based on held-out likelihood. Surprisingly, topic models which perform better on held-out likelihood may infer less semantically meaningful topics.

Hannigan et al's represention

Rendering of artifacts can be absent



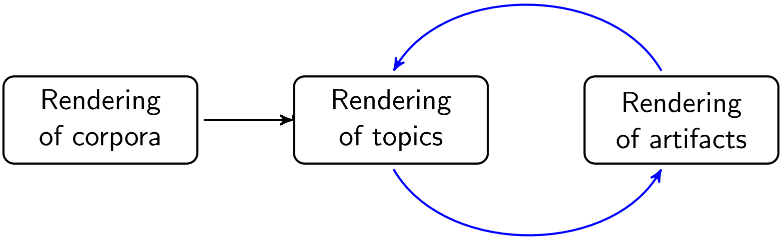
Such a scenario materializes when the goal of TM is producing a set of features to use in a statistical/ML model.

Rendering of artifacts can drive rendering of topics

Rendering of corpora → Rendering of artifacts → Rendering of topics

Such a scenario materializes when analysts want to test the existence of known a prior constructs (we will see this scenario in the webinar.)

Iterative approach



Oftentimes, analysts use an iterative approach — they retain a model with n_i topics, then, they try make sense out of topics; every interpretation informs the next topic model wherein n_j topics are retained (see DiMaggio (2015 — Big Data & Society).

Need for TM guidelines

The next sections of the slide-show survey some TM applications in the econ/organizational fields and try to make order/clarity management scholars evaluate topic models, we aim at:
The goal is twofold:

- clarifying the expectations of authors and reviewers on what constitutes a valid topic model
- promoting the reproducibility of TM estimates

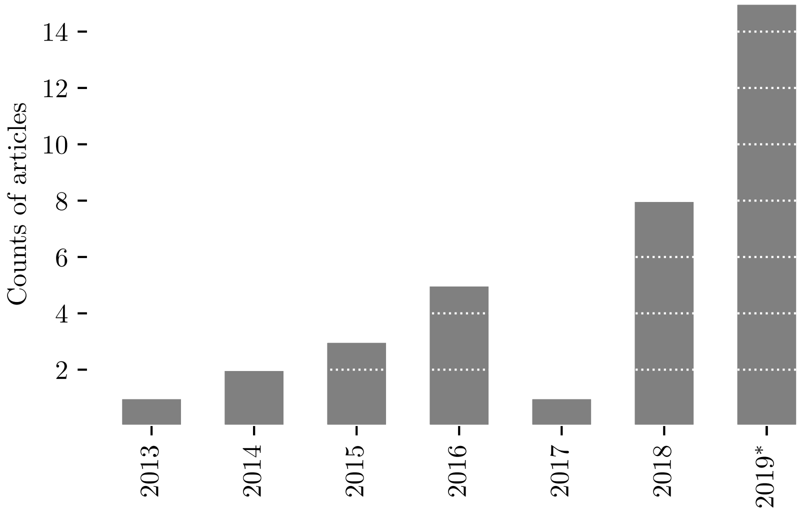
Outline

Sampling

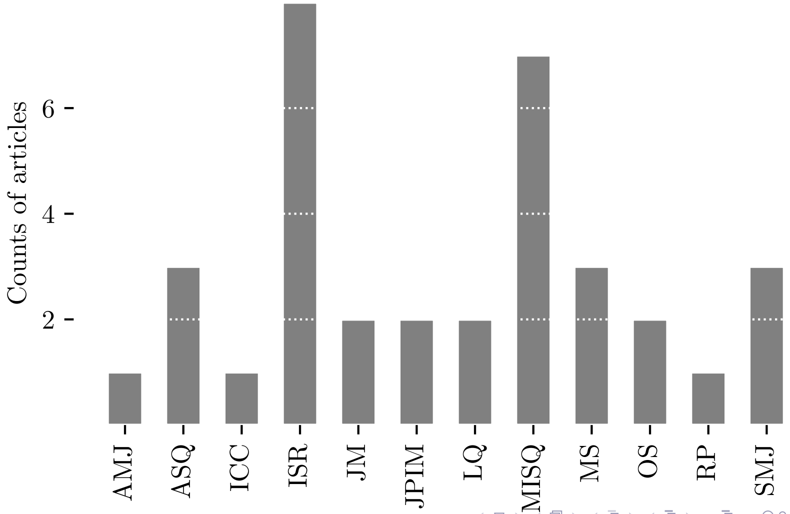
Keywords: ‘topic model*,’ ‘natural language processing,’ ‘nlp,’ ‘latent dirichlet,’ ‘LDA’

Journals: Academy of Management Journal, Administrative Science Quarterly, Entrepreneurship Theory and Practice, Industrial and Corporate Change, Information Systems Research, Journal of Business Venturing, Journal of Management, Journal of Management Studies, Journal of Product Innovation Management, Leadership Quarterly, Management Science, MIS Quarterly, Organization Science, Organization Studies, Research Policy, Strategic Entrepreneurship Journal, Strategic Management Journal, Strategic Organization.

Sample — articles across journals



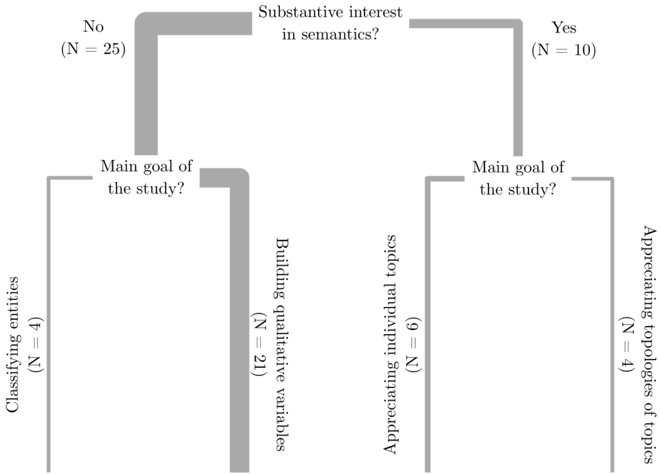
Sample — articles across journals



Coding scheme

Domain	Variable	Synopsis
Scope of the study	Substantial semantic interest	[0 = No; 1 = Yes]
Research design	Empirical goal	[0 = Classification; 1 = qualitative variables; 2 = individual topics; 3 = topology]
Evaluation	Heuristic	[0 = No; 1 = Yes]
	Statistical	[0 = No; 1 = Yes]
		Arun et al. 2010 [0 = No; 1 = Yes]
		Cao et al. 2009 [0 = No; 1 = Yes]
		Deveud et al. 2014 [0 = No; 1 = Yes]
		Dispersion of Residuals [0 = No; 1 = Yes]
		Document-completion Held-out likelihood [0 = No; 1 = Yes]
		Frequency and Exclusivity - FREX [0 = No; 1 = Yes]
		Griffiths and Steyvers 2004 [0 = No; 1 = Yes]
		Perplexity [0 = No; 1 = Yes]
		Semantic Coherence [0 = No; 1 = Yes]
		Silhoutte Coefficient [0 = No; 1 = Yes]
	Eyeballing	[0 = No; 1 = Yes]
		Keywords inspection [0 = No; 1 = Yes]
		Visual inspection [0 = No; 1 = Yes]
	Semantic	[0 = No; 1 = Yes]
		Word intrusion [0 = No; 1 = Yes]
		Topic intrusion [0 = No; 1 = Yes]
		Polysemy inspection [0 = No; 1 = Yes]
		Topic to document inspection [0 = No; 1 = Yes]
		Human coder agreement [0 = No; 1 = Yes]
	External	[0 = No; 1 = Yes]

Results — scope and goals



Results — scope and goals (cont'd)

Abbasi et al. (2018)
Larsen & Bong (2016)
Ruckman & Mcarthy (2017)
Wang et al. (2018)

Jung & Lee (2016)
Geva et al. (2019)
Choudhury et al. (2019)
Haans (2019)
Kaplan & Vaikili (2015)
Ghose et al. (2019)
Huang et al. (2018)
Giorgi & Weber (2015)
Corritore et al. (2019)
Hasan et al. (2015)
Wu (2013)
Adamopoulos et al. (2018)
Yang et al. (2019)
Gong et al. (2018)
Shi et al. (2016)
Yue et al. (2019)
Antons et al. (2019)
Hwang et al. (2019)
Singh et al. (2014)
Khernamnuai et al. (2018)
Huang et al. (2019)

Banks et al. (2019)
Bao and Datta (2014)
Doldor et al. (2019)
Lappas et al. (2016)
Nielsen & Borjëson (2019)
Sieweke & Santoni (2019)

Antons et al. (2016)
Croidieu & Kim (2018)
Giorgi et al. (2019)
Hopp et al. (2018)

No substantive interest in semantics - classification

Industrial and Corporate Change Advance Access published November 29, 2016

Industrial and Corporate Change, 2016, 1–22
doi: 10.1093/icc/dtw046
Original article

OXFORD

Why do some patents get licensed while others do not?

Karen Ruckman[1,*] and Ian McCarthy[2]

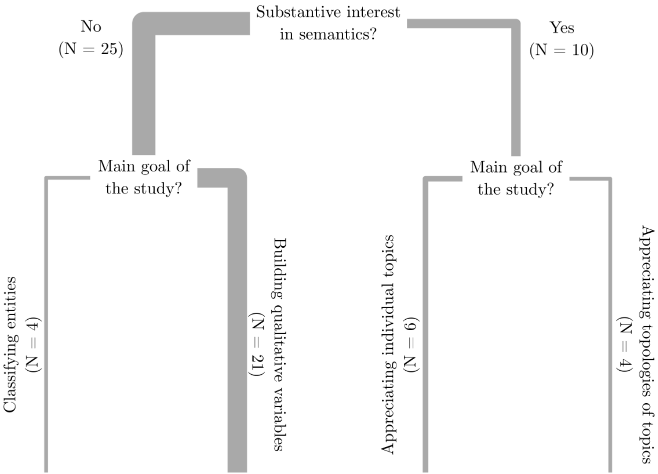
[1]Beedie School of Business, Simon Fraser University, 8888 University Avenue, Burnaby, British Columbia v5a 1s6, Canada. email: ruckman@sfu.ca and [2]Beedie School of Business, Simon Fraser University, 8888 University Avenue, Burnaby, British Columbia v5a 1s6, Canada. email: imccarthy@sfu.ca

*Main author for correspondence.

Abstract

To understand why some patents get licensed and others do not, we estimate a portfolio of firm- and patent-level determinants for why a particular licensor's patent was licensed over all technologically similar patents held by other licensors. Using data for licensed biopharmaceutical patents, we build a set of alternate patents that could have been licensed-in using topic modeling techniques. This provides a more sophisticated way of controlling for patent characteristics and analyzing the attractiveness of a licensor and the characteristics of the patent itself. We find that patents owned by licensors with technological prestige, experience at licensing, and combined technological depth and breadth

Results — scope and goals



No substantive interest in semantics - qual. vars.

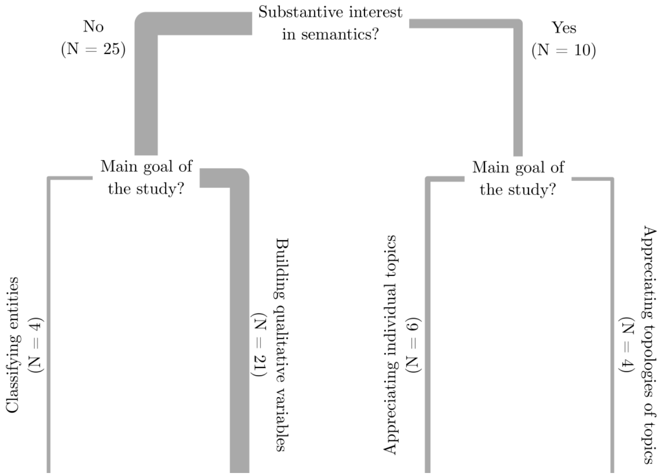
Strategic Management Journal
Strat. Mgmt. J., **36**: 1435–1457 (2015)
Published online EarlyView 2 July 2014 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.2294
Received 28 July 2013; Final revision received 13 May 2014

THE DOUBLE-EDGED SWORD OF RECOMBINATION IN BREAKTHROUGH INNOVATION

SARAH KAPLAN[1]* and KEYVAN VAKILI[2]
[1] *Strategic Management, Rotman School, University of Toronto, Toronto, Ontario, Canada*
[2] *Strategy and Entrepreneurship, London Business School, London, U.K.*

We explore the double-edged sword of recombination in generating breakthrough innovation: recombination of distant or diverse knowledge is needed because knowledge in a narrow domain might trigger myopia, but recombination can be counterproductive when local search is needed to identify anomalies. We take into account how creativity shapes both the cognitive novelty of the idea and the subsequent realization of economic value. We develop a text-based measure of novel ideas in patents using topic modeling to identify those patents that originate new topics in a body of knowledge. We find that, counter to theories of recombination, patents that originate new topics are more likely to be associated with local search, while economic value is the product of broader recombinations as well as novelty.

Results — scope and goals



Substantive interest in semantics - individual topics

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

ELSEVIER

Review

Natural experiments in leadership research: An introduction, review, and guidelines

Jost Sieweke[a,*], Simone Santoni[b]

[a] Vrije Universiteit Amsterdam, Netherlands
[b] Cass Business School, United Kingdom of Great Britain and Northern Ireland

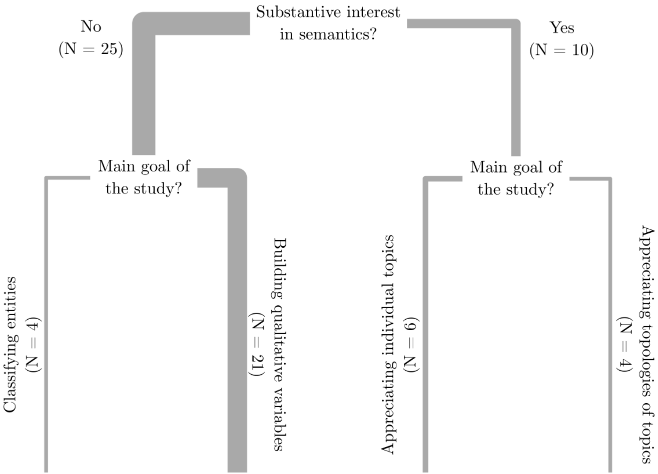
ARTICLE INFO

Keywords:
Causal inference
Leadership
Instrumental variable design
Natural experiment
Regression discontinuity design
Topic modeling

ABSTRACT

Endogeneity is a serious challenge for leadership research. To overcome the problem, researchers increasingly rely upon experimental designs, such as laboratory and field experiments. In this paper, we argue that natural experiments — in the form of standard natural experiments, instrumental variable, and regression discontinuity designs — offer additional opportunities to infer causal relationships. We conduct a systematic, cross-disciplinary review of 87 studies that leverage natural experimental designs to inquire into a leadership topic. We introduce the standard natural experiment, instrumental variable, and regression discontinuity design and use topic modeling to analyze which leadership topics have been investigated using natural experimental designs. Based on the review, we provide guidelines that we hope will assist scholars in discovering natural exogenous variations, selecting the most suitable form of natural experiment and by mobilizing appropriate statistical techniques and robustness checks. The paper is addressed to leadership and management scholars who aim to use natural experiments to infer causal relationships.

Results — scope and goals



Substantive interest in semantics - topology

Administrative Science Quarterly
2018, Vol. 63(1)1–42

DOI: 10.1177/0001839216686531
journals.sagepub.com/home/asq
SAGE

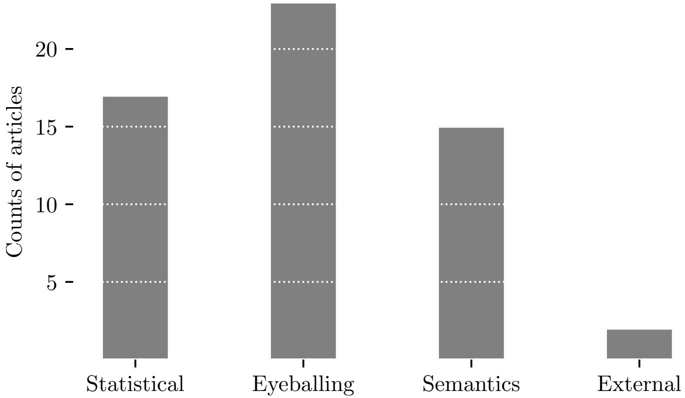
Labor of Love: Amateurs and Lay-expertise Legitimation in the Early U.S. Radio Field

Grégoire Croidieu[1] and Phillip H. Kim[2]

Abstract

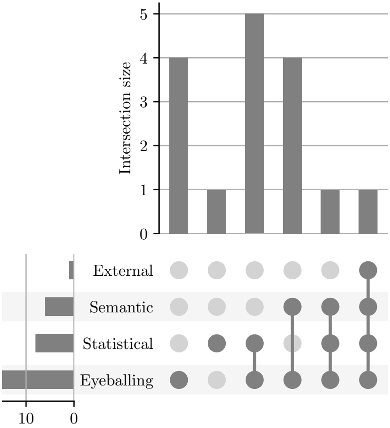
Many actors claim to be experts of specialized knowledge, but for this expertise to be perceived as legitimate, other actors in the field must recognize them as authorities. Using an automated topic-model analysis of historical texts associated with the U.S. amateur radio operator movement between 1899 and 1927, we propose a process model for lay-expertise legitimation as an alternative to professionalization. While the professionalization account depends on specialized work, credentialing, and restrictive jurisdictional control by powerful field actors, our model emphasizes four mechanisms leading to lay-expert recognition: building an advanced collective competence, operating in an unrestricted public space, providing transformational social contributions, and expanding an original collective role identity. Our analysis shows how field

Results — evaluation approaches



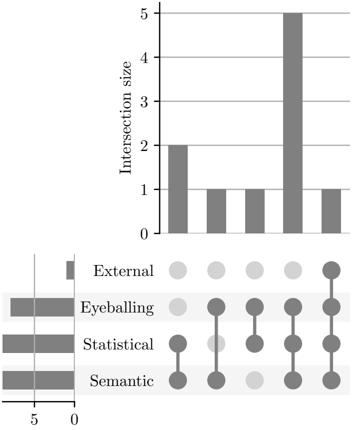
Association between scope and evaluation approaches

Studies without substantive interest in semantics



Association between scope and evaluation approaches

Studies with substantive interest in semantics



Outline

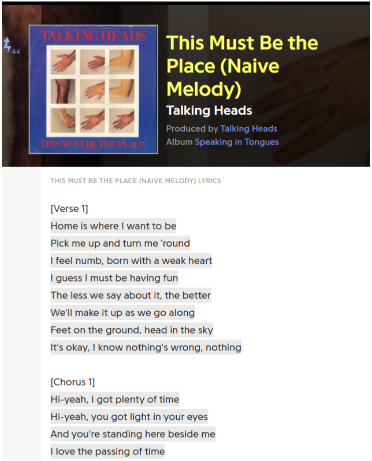
Some TM guidelines. . . in action

Let's suppose to have a dataset containing song lyrics.

TM could be used to analyze the dataset from different angles.

Let's consider some concrete examples and see, case-by-case, how to plausibly assess the validity of the TM at hand.

Mainly, I suggest 'how to best render topics' depends on the goal(s) analysts want to pursue using TM.



Vig. 1: No substantive semantic interest, classification

Study: organization and functioning of 'open' categorization system

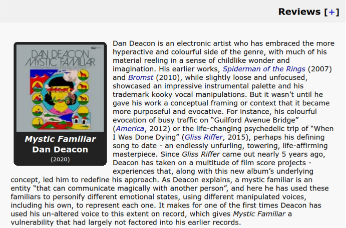
Setting: Rate Your Music

Issue: it is difficult to separate the features that make a category creating song from the effects of social interactions developing within the online community

Role of TM: clustering songs

TM suggestion:

- sample multiple topic models with high numbers of topics
- retain the topic model with the best statistical fit



Vig. 2: No substantive interest, qualitative variables

Study: popularity of pop cultural products [e.g., Askin & Mauskapf, 2017 — ASR]

Setting: Billboard

Issue: classical OVB problem

Role of TM: ceteris paribus comparison in regression settings.

TM suggestion:

- sample multiple topic models with high numbers of topics
- retain the topic model with the best statistical fit
- metrics that highlight topic distinctiveness have priority

ASA
AMERICAN SOCIOLOGICAL ASSOCIATION

American Sociological Review
2017, Vol. 82(5) 910–944

DOI: 10.1177/0003122417728662
journals.sagepub.com/home/asr
SAGE

What Makes Popular Culture Popular? Product Features and Optimal Differentiation in Music

Noah Askin[a] and Michael Mauskapf[b]

Abstract

In this article, we propose a new explanation for why certain cultural products outperform their peers to achieve widespread success. We argue that products' position in feature space significantly predicts their popular success. Using tools from computer science, we construct a novel dataset allowing us to examine whether the musical features of nearly 27,000 songs from *Billboard*'s Hot 100 charts predict their levels of success in this cultural market. We find that, in addition to artist familiarity, genre affiliation, and institutional support, a song's perceived proximity to its peers influences its position on the charts. Contrary to the claim that all popular music sounds the same, we find that songs sounding too much like previous and contemporaneous productions—those that are highly typical—are less likely to succeed. Songs exhibiting some degree of optimal differentiation are more likely to rise to the top of the charts. These findings offer a new perspective on success in cultural markets by specifying how content organizes product competition and audience consumption behavior.

Vig. 3: Substantive interest in semantics, individual topics

Study: category emergence

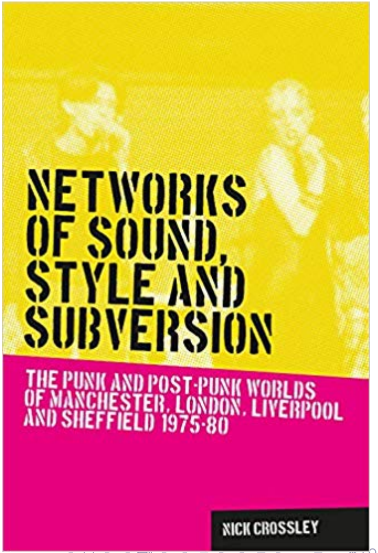
Setting: 1970's punk scene

Issue: detecting the features behind a new style/philosophy

Role of TM: capturing the linguistic manifestations of latent phenomena (e.g., the DIY ethos)

TM suggestion:

- sample models with reasonably low numbers of topics
- semantic and external evaluations should have priority



Vig. 4: Sub. interest in semantics, focus on topologies

Study: technology and systems of cultural production
Setting: early 1980's, advent of synths in the recording music sector
Issue: appreciating the effect of technological constraints on creativity/novelty emergence
Role of TM: comparing and contrasting the distribution of meanings in the cultural production system
TM suggestion:

- sample models with reasonably low numbers of topics
- semantic and external evaluations should have priority
- human (expert) judgements is