

latent variables

unsupervised learning

DTL SU @ AU

kristoffer l nielbo

kln@cas.au.dk

github.com/kln-courses/tmgu17

tmgu17.slack.com

DAI|IMC|AARHUS UNIVERSITY

Topic modeling

Probabilistic modeling within the framework of unsupervised mixed models. Each document is more or less likely within each clusters



- ① Discover thematic structure
 - ② Annotate documents
 - ③ Use the annotations to visualize, organize, summarize, ...

Topic analysis of multiple documents

Document: structured* or non-random collection of words

What is the structuring agent?

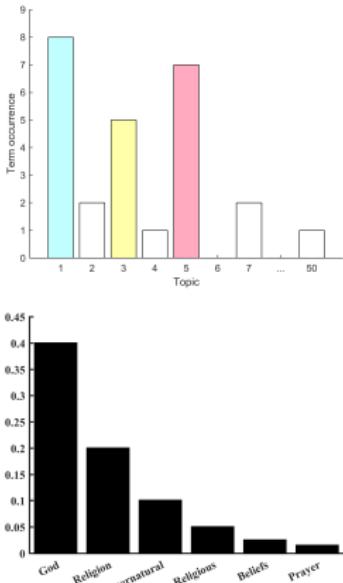
How to create a document

- ① Choose a distribution over topics
 - ② : For each word in the document:
 - Choose a topic from 1
 - Choose a word from the topic's distribution over the vocabulary

Topics create documents: $\text{topics} \Rightarrow \text{doc} \& \text{words}$

If we want to extract the **hidden topics** that generated the documents, we have to reverse the process: $\text{words} \& \text{doc} \Rightarrow \text{topics}$

Document creation



ABSTRACT—We present two studies aimed at resolving experimentally whether **religion** increases prosocial behavior in the anonymous dictator game. Subjects allocated more money to anonymous strangers when **God-concepts** were implicitly activated than when neutral or no concepts were activated. This effect was at least as large as that obtained when concepts associated with secular moral institutions were primed. A trait measure of self-reported religiosity did not seem to be associated with prosocial behavior. We discuss different possible mechanisms that may underlie this effect, focusing on the hypotheses that the religious prime had an **ideomotor effect** on generosity or that it activated a felt presence of supernatural watchers. We then discuss implications for theories positing religion as a facilitator of the emergence of early large-scale societies of **cooperators**.

Many theorists have suggested that the cognitive availability of omniscient and omnipotent **supernatural** agents has had a dramatic impact on the development of large-scale human societies. The imagined presence of such agents, along with emotional ritual and costly commitment to the social group they govern, may have been the major development that allowed genetically unrelated individuals to interact in cooperative ways (e.g., Aron & Norenzayan, 2004; Irons, 1991; Sosis & Ruffle, 2004). The research reported in this article experimentally investigated this link between two broad classes of culturally widespread phenomena of interest to social science—**religious** beliefs and **cooperative** behavior among unrelated strangers.

Although anecdotes documenting religion's prosocial and antisocial effects abound, the empirical literature has produced mixed results regarding religion's role in prosocial behavior.

Sosis and Ruffle (2004) examined levels of generosity in an experimental cooperative pool game in religious and secular kibbutzim in Israel and found higher levels of cooperation in the religious ones, and the highest levels among religious men who engaged in daily communal **prayer**. Batson and his colleagues (Batson et al., 1999; Batson, Schoenrade, & Ventis, 1993) have shown that although religious people report more explicit willingness to care for others than do nonreligious people, controlled laboratory measures of altruistic behavior often fail to corroborate this difference. Furthermore, when studies demonstrate that helpfulness is higher among more devoted people, this finding is typically better explained by egoistic motives such as seeking praise or avoiding guilt, rather than by higher levels of compassion or by a stronger motivation to benefit other people.

However insightful these findings are, research on religion and prosocial behavior has been limited by its overwhelming reliance on **correlational** designs. If religiosity and prosocial behavior are found to be correlated, it is just as likely that having a prosocial disposition causes one to be religious, or that some third variable such as guilt proneness or dispositional empathy causes both cooperative behavior and religiosity, as that religious beliefs somehow cause prosocial behavior. Only rarely have studies induced **supernatural beliefs** to examine them as a causal factor. Bering (2003, 2006) inhibited 3-year-old children's tendencies to cheat (i.e., open a "forbidden box") by telling them that an invisible agent ("Princess Alice") was in the room with them. In a different study, college students who were casually told that the ghost of a dead graduate student had been spotted in their private testing room were less willing to cheat on a computerized spatial-reasoning task than were those told nothing (Bering, McLeod, & Shackelford, 2005). These studies suggest that explicit thoughts of supernatural agents curb cheating behavior.

In the research reported here, we examined the effect of **God** concepts specifically on selfish and prosocial behavior. Our research design was novel in two ways. First, we introduced an



Latent Dirichlet Allocation

Simple topic model that uses expectation-maximization (VEM) or Gibbs sampling (GIBBS) to fit the model

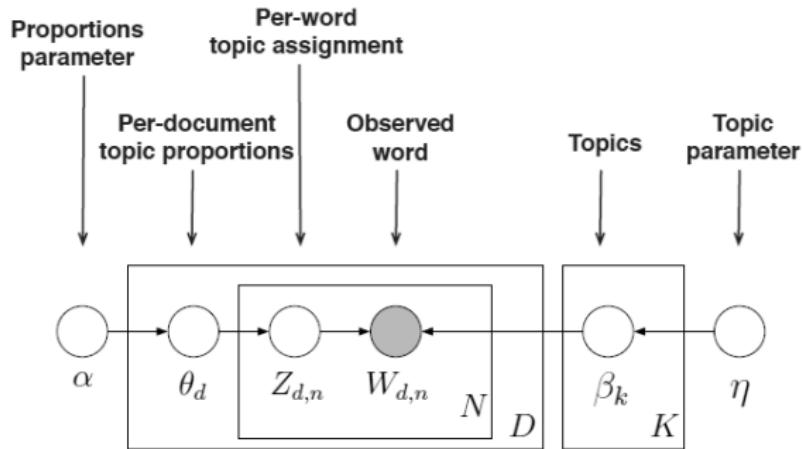
Competing sparsity (goals of LDA)

- ① For each *document*, allocate its words to as *few topics* as possible
- ② For each *topic*, assign high probability to as *few terms* as possible

These goals are at odds, because putting a topic in a single document complicates 2
hard and putting few words in each topic complicates 1

All of the words in a document have a probability under each topic (1), but to cover a
document's words, we must assign many topics to it (2)

→ trade off between these goals results in tightly co-occurring words within each topic



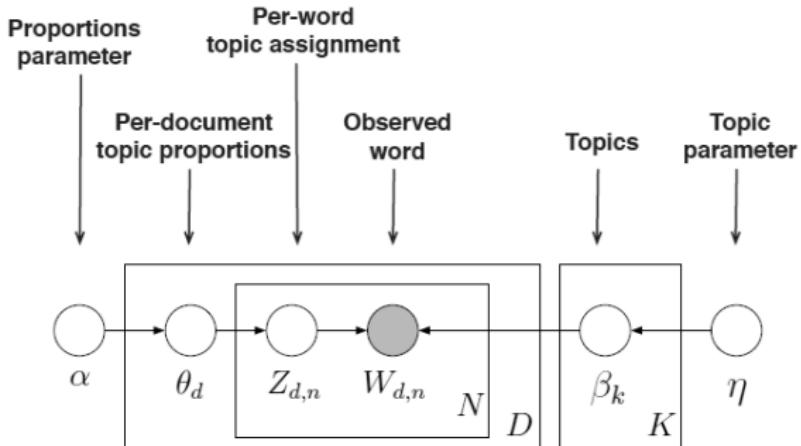
The joint distribution defines a posterior probability: $p(\theta, z, \beta | w)$

From a collection of documents we can infer

- Per-word topic assignment $Z_{d,n}$
- Per-document topic proportions θ_d
- Per-corpus topic distributions β_k

Use posterior expectations to perform given task





k determines the model's granularity and can be estimated using model perplexity

Low α value places more weight on documents composed of only a few dominant topics (high value will return more relatively dominant topics) $\rightarrow \frac{50}{k}$

A low η/β value places more weight on having each topic composed of only a few dominant words $\rightarrow 0.01$

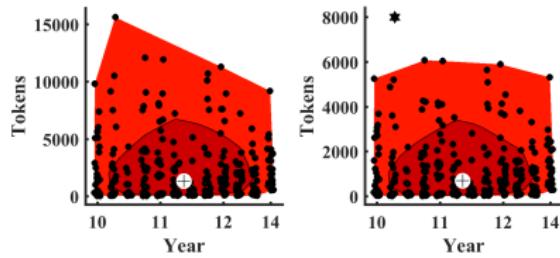
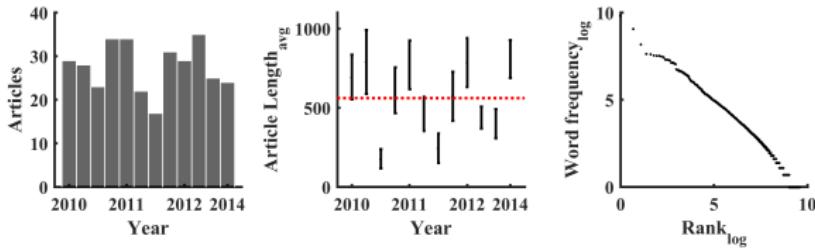
Jihadist macroscope



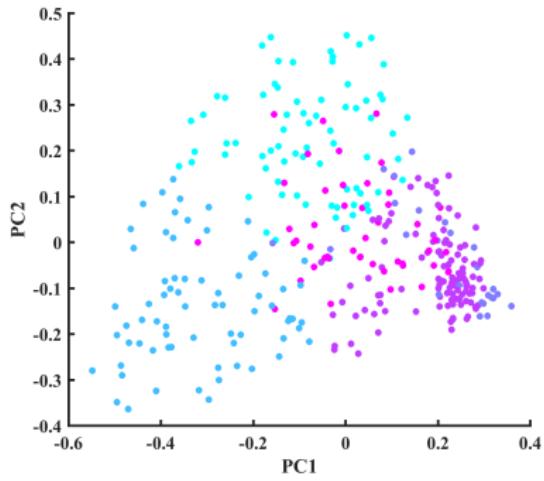
Religious explanation vs. kinship explanation (+ socio-economic)

Project that relies heavily on humanities domain expertise in selection, pre-processing, modeling, evaluation and interpretation

Corpus



Partitional clustering



KEYWORDS

brother
fight
mujahidin
ummah
women

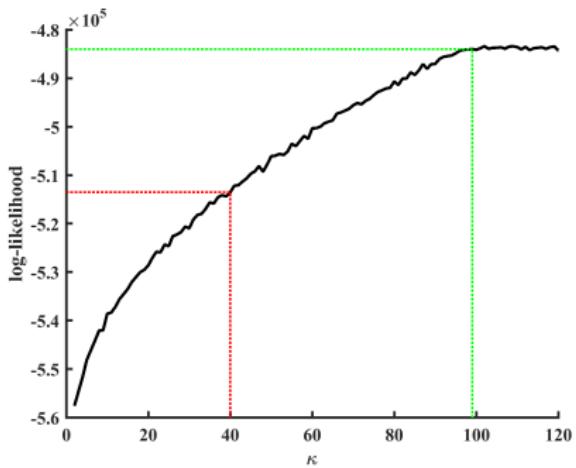
abu
brother
children
love
people

Al Qaeda
AQAP
Inspire
issue
magazine

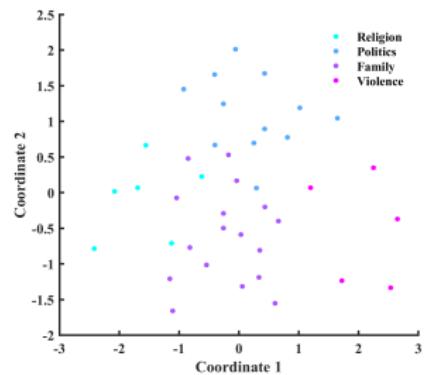
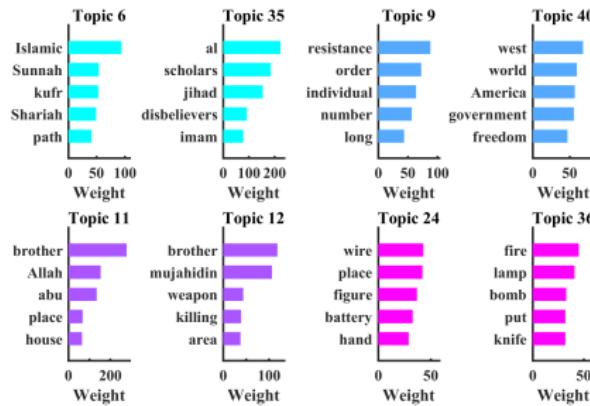
America
attack
bomb
people
world

Allah*
Islam
message
Muslim
time

Determining k number of topics



LDA model with 40 topics



Further explorations

