Unsupervised Learning Methods

Computational Text Analysis

Theresa Gessler

University of Zurich | http://theresagessler.eu/ | @th_ges

2022-05-09

Program

- unsupervised models
  - general idea
  - types of models
- topic models
  - what are they
  - why do we use them
- the structural topic model
  - basic implementation in R using `stm`
  - special features of structural topic models

Text Analysis Methods

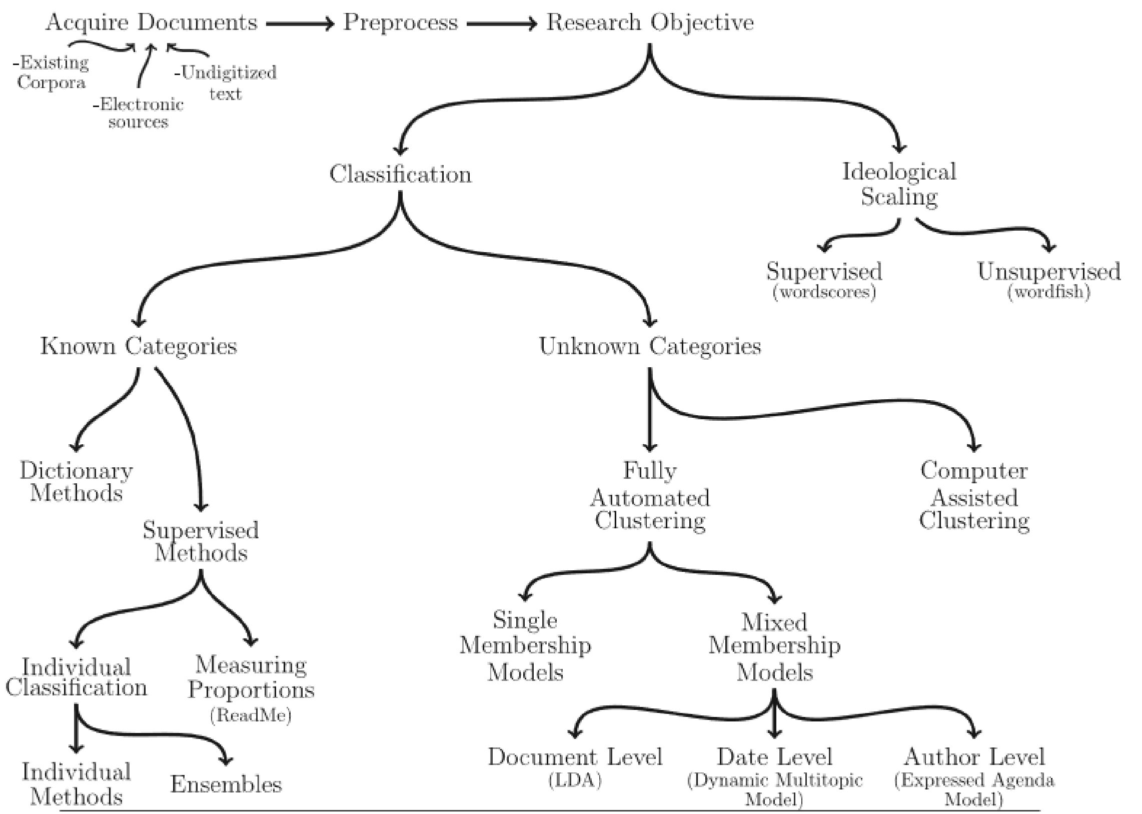


Fig. 1 An overview of text as data methods.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21, 267-297.

Unsupervised Models

- search for patterns in feature space without pre-existing labels
    - typically: cluster formation

→ purpose: **discovery** of relevant categories

Ingredients (Grimmer & Stewart 2013):

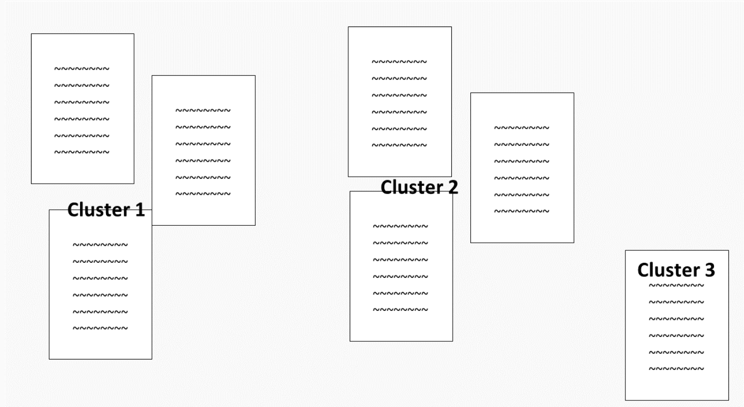
- definition of document similarity
- function that operationalizes ideal clustering
- optimization algorithm
    - often iterative

Unsupervised models

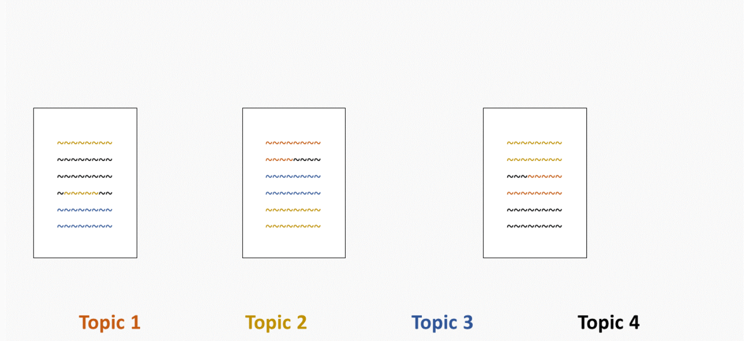
Typology (an attempt)

- **clustering models** (single membership)
  - k means clustering
  - hierarchical clustering, text similarity → quanteda: `textstat_similarity()`
- **topic models** (mostly multi-membership)
  - often: inclusion of data structure, e.g. expressed agenda model or dynamic topic model
- **semi-supervised and computer-assisted methods**
  - 'seeded' topic models and other assisted clustering methods (Grimmer and King 2011)

Clustering



Topic models



Topic models

Topic models

Topic models

- algorithm to find most important 'topics' in an unstructured corpus through analyzing co-occurence of words

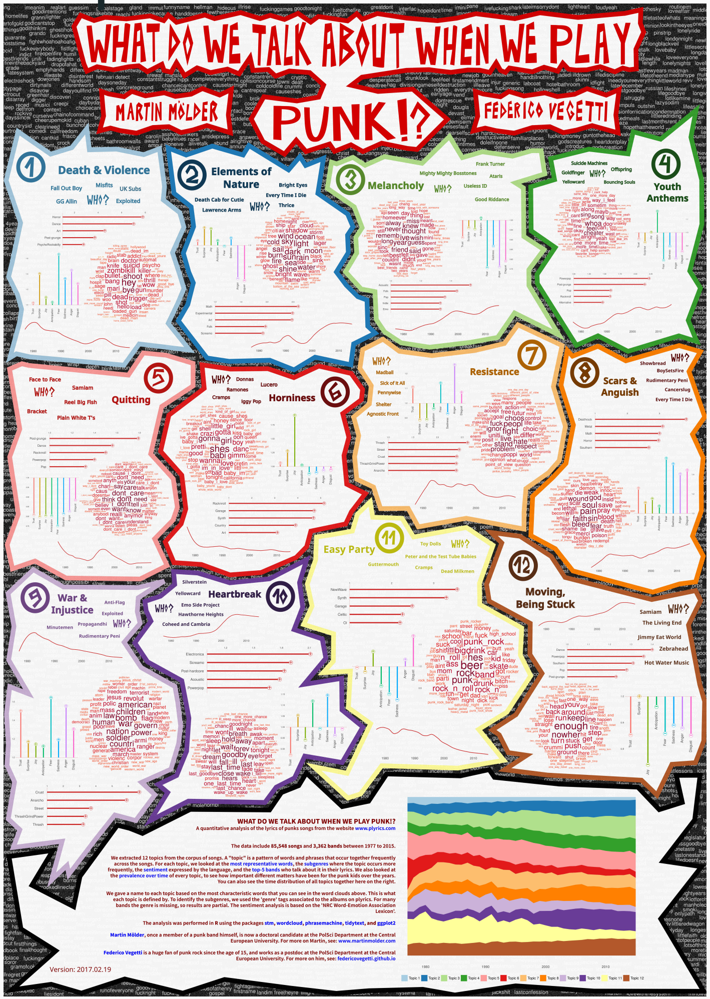
Typical use

- to get a first impression of data
    - to get an overview when you face large amounts of data
- to study the trajectory of large topics in diverse corpora
    - in scientific journals, twitter data, archival texts, music lyrics, ...
- to study different frames of the same debate (DiMaggio, Manish & Blei 2013)
    - different aspects highlighted, different words used

Topic models

A non-scientific example of a topic model

Poster by Martin Mölder & Federico Vegetti



WHAT DO WE TALK ABOUT WHEN WE PLAY PUNK!?

A quantitative analysis of the lyrics of punks songs from the website www.plyrics.com

The data include 85,548 songs and 3,362 bands between 1977 to 2015.

We extracted 12 topics from the corpus of songs. A "topic" is a pattern of words and phrases that occur together frequently across the songs. For each topic, we looked at the most representative words, the subgenres where the topic occurs more frequently, the sentiment expressed by the language, and the top-5 bands who talk about it in their lyrics. We also looked at the prevalence over time of every topic, to see how important different matters have been for the punk kids over the years. You can also see the time distribution of all topics together here on the right.

We gave a name to each topic based on the most characteristic words that you can see in the word clouds above. This is what each topic is defined by. To identify the subgenres, we used the 'genre' tags associated to the albums on plyrics. For many bands the genre is missing, so results are partial. The sentiment analysis is based on the 'NRC Word-Emotion Association Lexicon'.

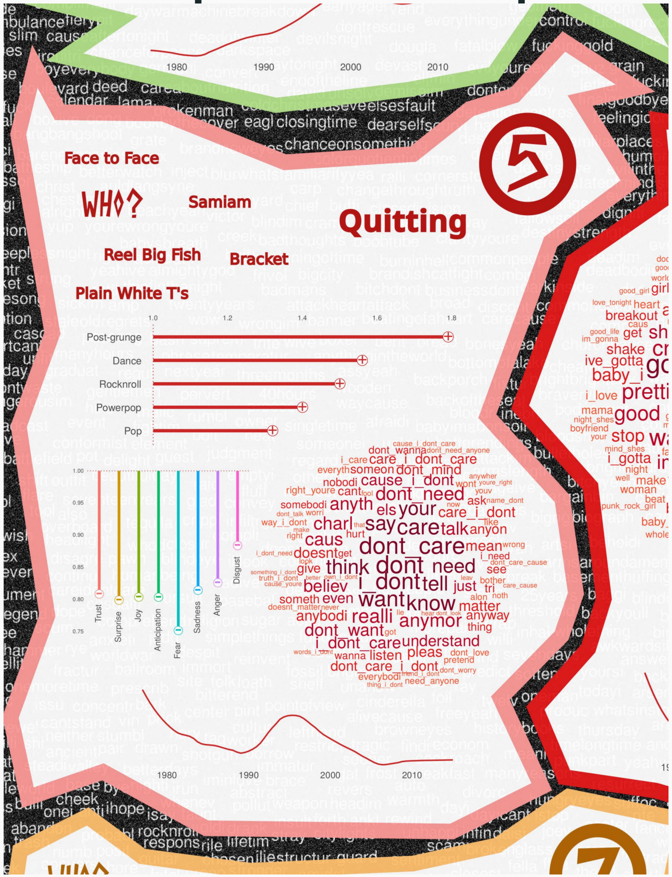
The analysis was performed in R using the packages stm, wordcloud, phrasemachine, tidytext, and ggplot2

Martin Mölder, once a member of a punk band himself, is now a doctoral candidate at the PolSci Department at the Central European University. For more on Martin, see: www.martinmolder.com

Federico Vegetti is a huge fan of punk rock since the age of 15, and works as a postdoc at the PolSci Department at the Central European University. For more on him, see: federicovegetti.github.io

Topic models

A non-scientific example of a topic model



Topic models

Assumptions

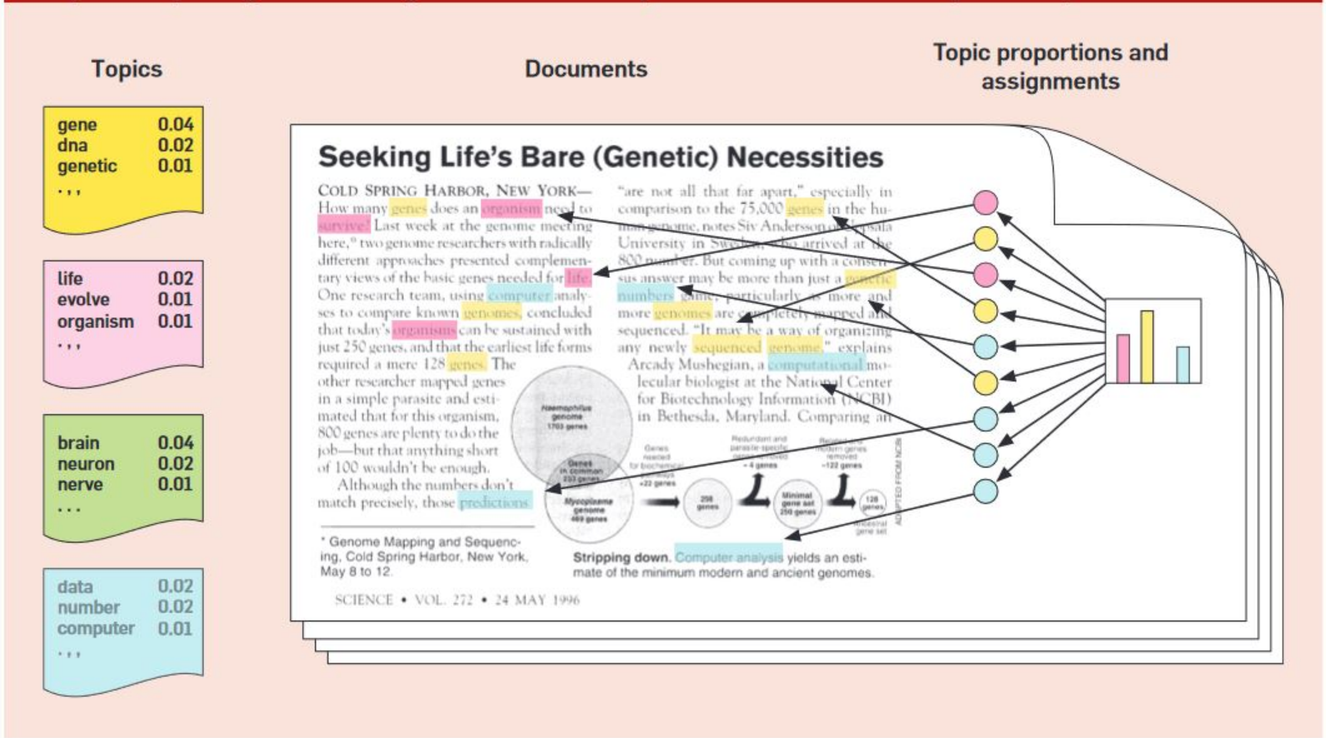
- each document consists of a mixture of topics
    - *drawn from a distribution of topics*
- documents with similar topics will use similar groups of words
    - every topic is connected more closely to some words, compared to others
- documents are created by
    - first choosing shares of topics
    - then choosing the words

→ unrealistic but useful assumptions about language

Topic models

Latent Dirichlet Allocation

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
Redundant and parasite-specific genes removed −4 genes
Related and modern genes removed −122 genes
256 genes
Minimal gene set 250 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

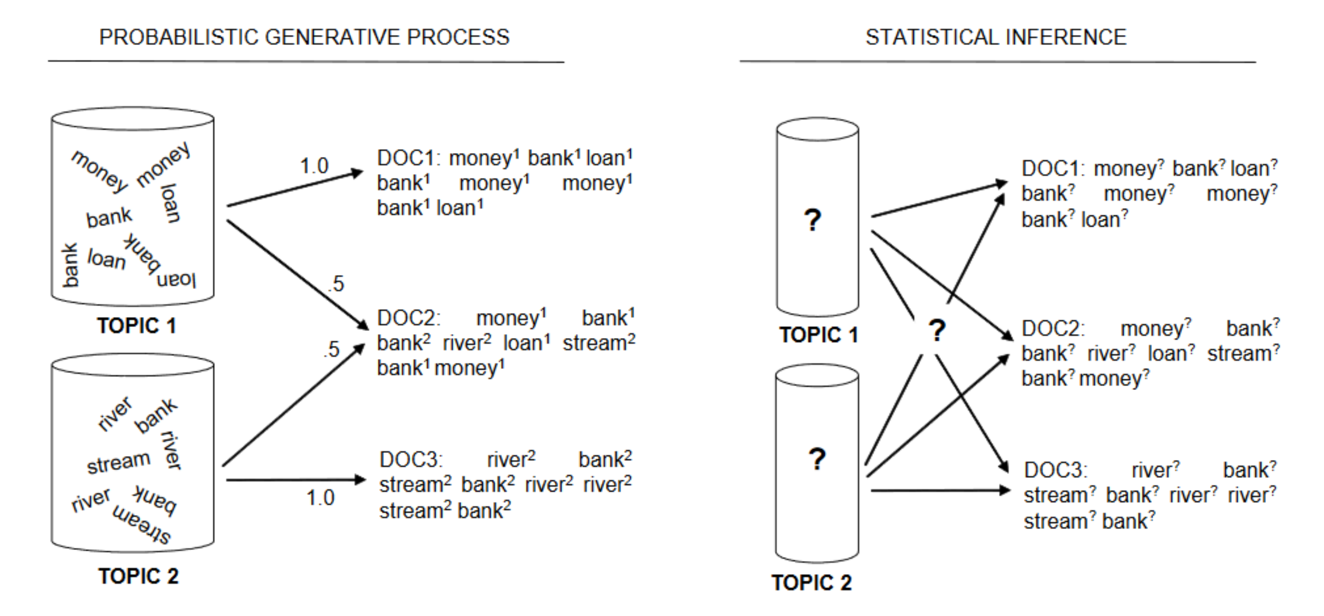
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

David Blei (2012): Probabilistic topic models

Topic models

Latent Dirichlet Allocation


PROBABILISTIC GENERATIVE PROCESS
STATISTICAL INFERENCE
money money loan bank bank loan bank loan
TOPIC 1
river bank river stream river bank stream
TOPIC 2
1.0
.5
.5
1.0
DOC1: money¹ bank¹ loan¹ bank¹ money¹ money¹ bank¹ loan¹
DOC2: money¹ bank¹ bank² river² loan¹ stream² bank¹ money¹
DOC3: river² bank² stream² bank² river² river² stream² bank²
?
TOPIC 1
?
TOPIC 2
?
DOC1: money? bank? loan? bank? money? money? bank? loan?
DOC2: money? bank? bank? river? loan? stream? bank? money?
DOC3: river? bank? stream? bank? river? river? stream? bank?

Steyvers & Griffiths (2007): Probabilistic Topic Models

Topic models

Latent Dirichlet Allocation

- topic models are estimated iteratively
    - random assignment of topics to tokens at first
    - changing topic weights with each iteration to maximize 2 goals
        - tokens of same type should belong to one topic
        - words in same document should belong to one topic

More theory

- original article on LDA
- a lecture on topic models by David Blei
- shorter video description of LDA
- an understandable introduction

The structural topic model

The structural topic model

- fast implementation of topic models in `stm` R-package
- many statistical tools for estimation of effects, rather than description
  - linked to attempts to make causal inference with text
- possibility to include document information into model
  - as explanation of topic *prevalence*
  - as factor that varies *content* of topic

Further reading: Roberts et.al. (2013), Roberts et.al.(2014), Roberts et.al. (2016)

The structural topic model

Input

3 components:

- documents
- vocabulary
- meta data

Preparation

- within `stm` package
  - `textProcessor()` and `prepDocuments()`
  - converting from quanteda with `convert(x, to="stm")`
- direct use of dfm (but: inconsistent number of documents for empty documents)

The structural topic model

Example: Gadarian Dataset

- 341 open-ended responses:
  - treatment: what makes you anxious about immigration
  - control: write about immigration
  - party ID

Literature

Gadarian, Shana Kushner, and Bethany Albertson. "Anxiety, immigration, and the search for information." Political Psychology 35.2 (2014): 133-164.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. "Structural Topic Models for Open-Ended Survey Responses." American Journal of Political Science 58, no 4 (2014): 1064-1082.

The structural topic model

Preparing the data

- how much pre-processing is needed and useful is an open question
  - some studies have discouraged the use of extensive pre-processing for topic models
  - but, pre-processing also speeds up estimation

```
# data prep
gadarian_corpus ← corpus(gadarian,text_field="open.ended.response")
gadarian_toks ← tokens(gadarian_corpus,
                        remove_punct=T,
                        remove_numbers=T) %>%
  tokens_remove(stopwords("en"))
gadarian_dfm ← dfm(gadarian_toks)
```

The structural topic model

Estimating a topic model

Simplest form:

```
gad_stm_1 ← stm(gadarian_dfm,K=10)
# from stm object:
#gad_stm_1 ← stm(gadarian$documents, gadarian$vocab, K=20)
```

Parameters

- texts and vocabulary
  - as dfm
  - or explicitly for `prepDocuments()` output
- number of topics (K)
  - or with K=0: automatic selection - *but this is not necessarily an ideal number*

The structural topic model

Estimating a topic model

- topic models are estimated iteratively
  - random assignment of topics to tokens at first
  - changing topic weights with each iteration to maximize 2 goals
    - tokens of same type should belong to one topic
    - words in same document should belong to one topic

```r
gad_stm_1 <- stm(gadarian_dfm,K=10,seed=2020)
```

```
## Beginning Spectral Initialization
##      Calculating the gram matrix...
##      Finding anchor words...
##      ..........
##      Recovering initialization...
##      .............
## Initialization complete.
##
......................................................................................
...........................
## Completed E-Step (0 seconds).
```

The structural topic model

Interpreting results

```
labelTopics(gad_stm_1)
```

```
## Topic 1 Top Words:
##       Highest Prob: get, new, come, country, laws, right, nation
##       FREX: new, amount, right, learning, laws, nation, know
##       Lift: congress, lobbyists, required, -english, -jobs, -more, -us
##       Score: new, amount, right, laws, support, know, congress
## Topic 2 Top Words:
##       Highest Prob: life, better, done, us, worry, citizens, immigration
##       FREX: another, im, done, life, better, major, wall
##       Lift: abroad, accross, alilens, anymore.and, attack, bank, became
##       Score: life, im, better, assimilate, another, family, wall
## Topic 3 Top Words:
##       Highest Prob: americans, security, illegals, jobs, social, healthcare, cost
##       FREX: healthcare, security, americans, social, mexicans, illegals, cost
##       Lift: act, activity, allegiance, along, amounts, annoying, began
##       Score: security, healthcare, americans, social, cost, mexicans, jobs
## Topic 4 Top Words:
##       Highest Prob: immigrants, english, many, economy, tax, paid, taking
##       FREX: immigrants, economy, tax, just, paid, english, taking
```

The structural topic model

Interpreting results

`labelTopics()` gives different outcomes:

- highest probability words
- FREX words: frequent and exclusive
- Lift & Score: commonly used indicators from other packages (`lda` / `maptpx`)

The structural topic model

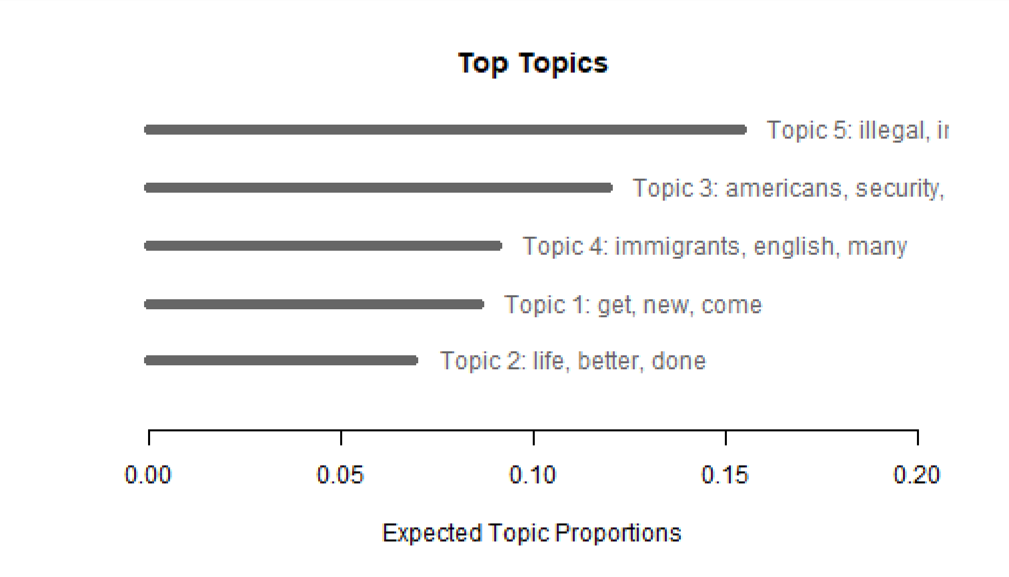
Interpreting results

```
labelTopics(gad_stm_1)
```

```
## Topic 1 Top Words:
##       Highest Prob: get, new, come, country, laws, right, nation
##       FREX: new, amount, right, learning, laws, nation, know
##       Lift: congress, lobbyists, required, -english, -jobs, -more, -us
##       Score: new, amount, right, laws, support, know, congress
## Topic 2 Top Words:
##       Highest Prob: life, better, done, us, worry, citizens, immigration
##       FREX: another, im, done, life, better, major, wall
##       Lift: abroad, accross, alilens, anymore.and, attack, bank, became
##       Score: life, im, better, assimilate, another, family, wall
## Topic 3 Top Words:
##       Highest Prob: americans, security, illegals, jobs, social, healthcare, cost
##       FREX: healthcare, security, americans, social, mexicans, illegals, cost
##       Lift: act, activity, allegiance, along, amounts, annoying, began
##       Score: security, healthcare, americans, social, cost, mexicans, jobs
## Topic 4 Top Words:
##       Highest Prob: immigrants, english, many, economy, tax, paid, taking
##       FREX: immigrants, economy, tax, just, paid, english, taking
```

Interpreting results

```
# formatting the plot (optional)
par(bty="n",col="grey40",lwd=5)
# plotting
plot(gad_stm_1,topics=1:5,type="summary",xlim=c(0,0.2))
```



The structural topic model

Looking at original documents

e.g. topic 3

```
labelTopics(gad_stm_1,topics=3)
```

```
## Topic 3 Top Words:
##       Highest Prob: americans, security, illegals, jobs, social, healthcare, cost
##       FREX: healthcare, security, americans, social, mexicans, illegals, cost
##       Lift: act, activity, allegiance, along, amounts, annoying, began
##       Score: security, healthcare, americans, social, cost, mexicans, jobs
```

The structural topic model

Looking at original documents

```
findThoughts(gad_stm_1, gadarian$open.ended.response, topics=3, n=3)
```

```
##
##  Topic 3:
##       robbing americans of their social security; rising costs of social programs
to cover illegals; rising costs of incarceration due to illegals; not being able to
understand what they are saying (spanish and other languages); the fact that illegal
immigrants are here with their children who go through our school systems then cannot
get financial aid for college because their parents never filed for naturalization
leaving the kids to work dead end jobs and never get naturalized; the fact that they
are doing the work that americans don't want to do anyway, and then the americans
complain about it; border walls; increasing amounts of drugs in america because it is
so easy to get into our country
##      i want illegals gone.  they cost my husband his business.  they come here
undercut the americans, and then raise their prices after they've driven hard working
americans out of their jobs.
##      migrant workers, mexicans, green cards, passports, border security, people
willing to do jobs most americans don't want, multi-lingual culture, ignorance, fear,
hypocrisy of americans
```

The structural topic model

Looking at original documents

e.g. topic 5

```
labelTopics(gad_stm_1,topics=5)
```

```
## Topic 5 Top Words:
##       Highest Prob: illegal, immigration, welfare, crime, legal, jobs, services
##       FREX: illegal, welfare, hospitals, fences, services, paying, crime
##       Lift: =, average, barriers, bigger, boarders, bums, car
##       Score: illegal, hospitals, crime, fences, schools, services, loss
```

The structural topic model

Looking at original documents

```
findThoughts(gad_stm_1, gadarian$open.ended.response, topics=5, n=5)
```

```
##
##  Topic 5:
##       illegal, draining our goverment resources and making no contibution to our
society.  as long as they will do the menial labor, our lazy welfare bums will
continue to rob, steal, and murder.  our jails are overcrowded!
## if welfare recipients, who are physically able, wont work then no welfare.
##      legal immigration is ok and needed in u.s.
## illegal immigration is very wrong for u.s.
##  - causes job loss for citizens
##  - drain on "medical system"
##  - crime rises in communities where
##    illegals stay
##      opportunity for the immagrants
## higher burden on the taxed citizens
## founding fathers
## we're paying for their health costs
## illegal immagration is illegal.
```

Inclusion of covariates

Inclusion of covariates

The case for structural topic models

"In practice, social scientists often know more about a document than its word counts" (Roberts et al 2016, 989)

→ **relation between covariates and latent topics** as focus of interest

→ **covariates shape priors** and are used for **partial pooling**

Inclusion of covariates

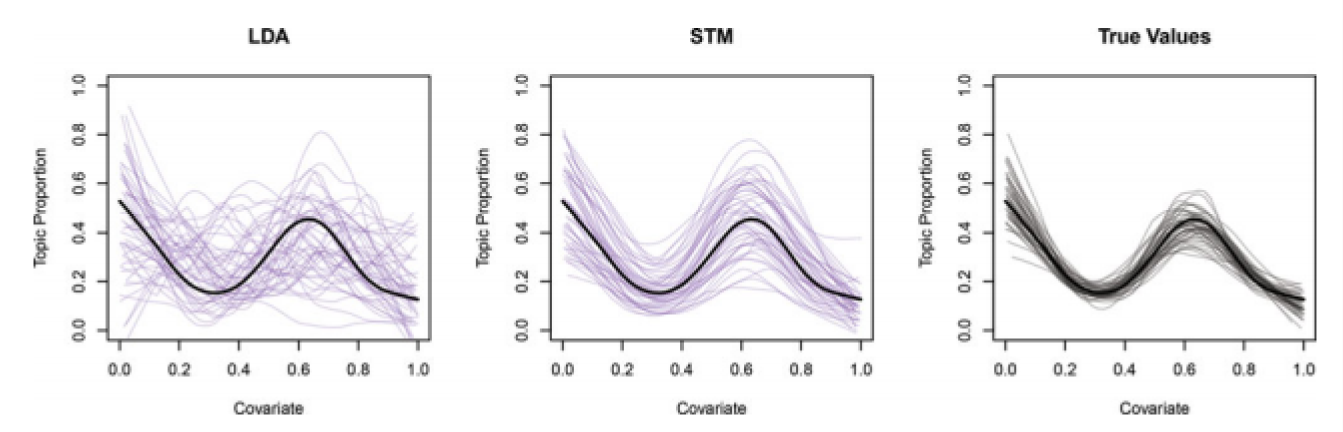
The case for structural topic models

"We leverage **generalized linear models (GLMs henceforth)** to introduce covariate information into the model. Prior distributions with globally shared mean parameters in the latent Dirichlet allocation model are replaced with means parameterized by a linear function of observed covariates. Specifically, for topic prevalence, the Dirichlet distribution that controls the proportion of words in a document attributable to the different topics is replaced with a logistic Normal distribution with a mean vector parameterized as a function of the covariates (Aitchison and Shen 1980). For topical content, we define the distribution over the terms associated with the different topics as an exponential family model, similar to a multinomial logistic regression, parameterized as a function of the marginal frequency of occurrence deviations for each term, and of deviations from it that are specific to topics, covariates, and their interactions. We shall often refer to the resulting model as the structural topic model (STM), because the **inclusion of covariates is informative about structure in the document collection and its design**. From an inferential perspective, including covariate information allows for **partial pooling of parameters along the structure defined by the covariates.**" (Roberts et al 2016, 989)

Inclusion of covariates

Prevalence and content

- stm can introduce **metadata** into the model that affects
  - **prevalence**: how prevalent a topic is for the group of documents
  - **content**: what is the content of a topic in the group of documents



Source: Roberts et al 2016

- standard formula notation (with ~), including log-transformations and smooting
  - e.g. `stm(gadarian_dfm, K=10, content=~date, data=docvars(gadarian_dfm))`

Inclusion of covariates

Prevalence

```
gad_stm_2 <- stm(gadarian_dfm,K=10,
                 prevalence=~treatment,
                 data=docvars(gadarian_dfm),verbose=F)
```

```
labelTopics(gad_stm_2)
```

```
## Topic 1 Top Words:
##       Highest Prob: get, new, come, country, laws, right, immigration
##       FREX: new, amount, right, learning, laws, nation, know
##       Lift: -english, -jobs, -more, -us, aboration, accidents, actively
##       Score: new, amount, right, laws, support, congress, lobbyists
## Topic 2 Top Words:
##       Highest Prob: life, better, done, government, immigration, assimilate,
another
##       FREX: another, im, done, life, major, wall, better
##       Lift: abroad, accross, alilens, ancestors, another, anymore.and, attack
##       Score: im, life, family, assimilate, better, another, wanting
## Topic 3 Top Words:
##       Highest Prob: jobs, americans, security, illegals, social, etc, healthcare
##       FREX: healthcare, violence, security, americans, social, jobs, illegals
```

Inclusion of covariates

Content

```
gad_stm_3 <- stm(gadarian_dfm,K=10,
                 content=~treatment,
                 data=docvars(gadarian_dfm),verbose=F)
```

- caution: *considerably slows down model estimation and results are sometimes poor*

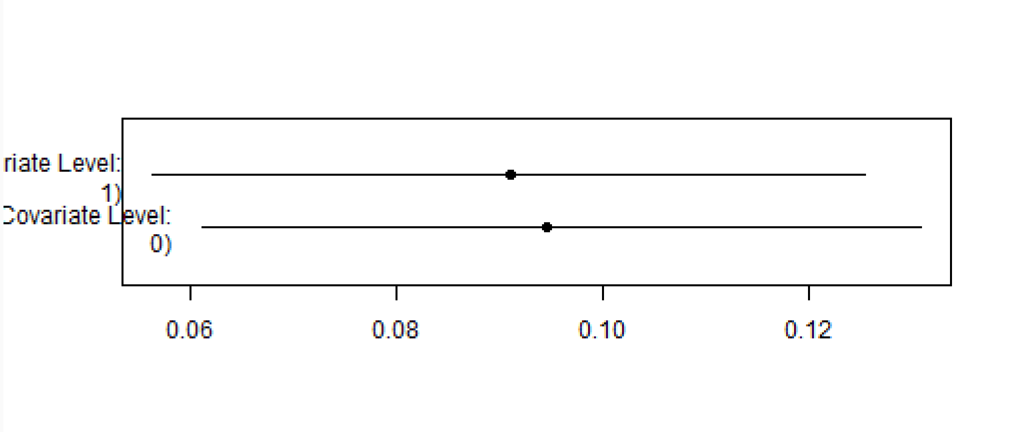
```
labelTopics(gad_stm_3,topics=3)
```

```
## Topic Words:
##  Topic 3: systems, america, americans, social, much, one, country
##
##  Covariate Words:
##  Group 0: themsleves, allour, assistance, finincal, imigerants, subsidize, cheap
##  Group 1: large
##
##  Topic-Covariate Interactions:
##  Topic 3, Group 0: stop, immagrants, =, higher, students, unattainable, mexicans
##  Topic 3, Group 1: usa, entering, history, politicians, crime, desperation,
accountability
##
```

Estimation of effects (prevalence)

```
gad_est ← estimateEffect(1:10~treatment,
                          gad_stm_2,docvars(gadarian_dfm))
```

```
plot(gad_est,topics=1,covariate="treatment")
```

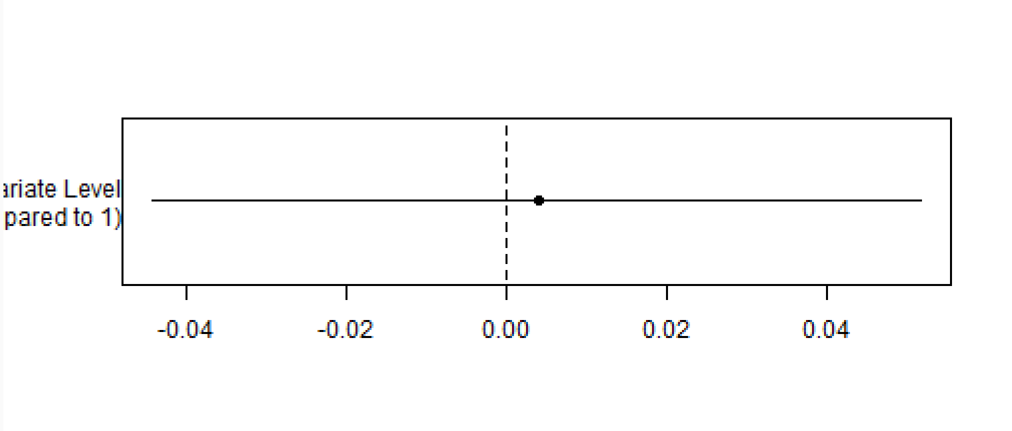


Inclusion of covariates

Estimation of effects (prevalence)

- three types of plots for effects
  - pointestimate
  - continuous (e.g. for time)
  - difference (for comparison of two values)

```
plot(gad_est,method="difference",cov.value1="0",
  cov.value2="1",covariate="treatment",topics=1)
```



Validation

Validation

Which model should we choose?

- technical validation
  - fit and coherence statistics of models (with different random strating points)
    - e.g. semantic coherence: at maximum when words in a topic frequently co-occur (Mimno et al 2011)
    - e.g. held-out-likelihood for new documents (Wallach et al 2009)

- manual validation / judgement
  - reading of typical texts for topics with `findThoughts()`
  - coherence of different topic descriptors

- more thorough manual validation / quantification
  - see e.g.: Chang et al 2009 (NIPS), Quinn et.al. 2010 (AJPS)
  - talk on Chang et al 2009 - video
  - word and topic intrusion test with `oolong` - *which is also a fantastic tool for validating dictionary methods!*

Validation

Technical: Different numbers of topics (k)

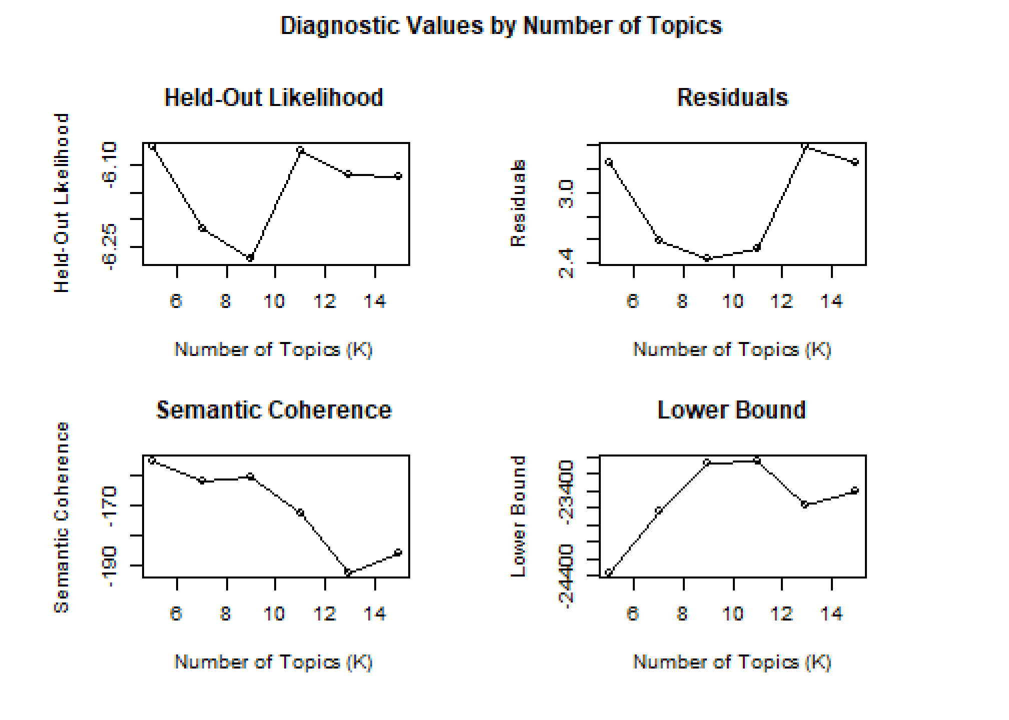
- `searchK()` command creates diagnostics for different topic numbers
    - different form `k=0` with attempt to find optimum
    - may require conversion of dfm

```
stm_obj← convert(gadarian_dfm,"stm")
ksearch←searchK(stm_obj$documents,stm_obj$vocab,
                     K = seq(5, 15, by = 2), max.em.its = 15)
```

```
## Beginning Spectral Initialization
##       Calculating the gram matrix...
##       Finding anchor words...
##       .....
##       Recovering initialization...
##       .............
## Initialization complete.
##
......................................................................................
...........................
## Completed E-Step (0 seconds).
## Completed M-Step.
```

Technical: Different numbers of topics (k)

```
plot(ksearch)
```



Validation

Technical: fit and coherence (same k)

- `selectModel()` evaluates different models for the same number of topics
- `manyTopics()` as combination of `searchK()` and `selectModel()` that allows evaluating different models across multiple number of topics
- see stm Vignette for more options

Validation

Manual validation

"Practitioners typically assume that the latent space is semantically meaningful. [...] However, whether the latent space is interpretable is in need of quantitative evaluation." - Chang et al 2009 (NIPS)

Two tests:

- **word intrusion**: identify a spurious word inserted into topic
- **topic intrusion**: identify a topic that is not associated with the document

→ implemented in `oolong`: overview vignette

Other types of topic models

Other types of topic models

- models that implement **alternative algorithms**
  - LDA (Latent Dirchlet Allocation)
  - LSA / LSI: Latent Semantic Analysis / Latent Semantic Indexing
  - NMF: Non-Negative Matrix Factorization
- **dynamic topic models**: variation of topics over time
  - trends of word usage within topics
  - Blei Lafferty 2007
- **correlated topic models**: non-random topic distributions for documents
  - explicit modeling of topics that co-occur
  - Blei Lafferty 2007
- **(semi-)supervised topic models**: manipulation of topics
  - introducing supervision into topic selection

(Semi)supervised topic models

A reviewer 2 comment: **You should not use topic models for classification**

↔ But what with all the great advantages of topic models?

→ **Semi-supervised topic models** allow you to nudge the model towards including certain topics, defined by keywords

- **incorporating information about documents** ↔ **incorporating information about topics**
  - →different use of pre-knowledge than stm, dynamic topic models etc.

(Semi)supervised topic models

Concept

Two aspects:

- improving **topic-word distributions**: topics prefer to generate words related to seed set
- improving **document-topic distributions**: model prefers topic for which seed words exist
- **But** we only nudge the model: Jagarlamudi et. al 2012, 205
    - *"importantly, we only encourage the model to follow the seed sets and do not force it. So if it has compelling evidence in the data to overcome the seed information then it still has the freedom to do so."*

(Semi)supervised topic models

Semi-supervised Topic models in R

- **Seeded LDA**
  - Watanabe, Kohei, and Yuan Zhou. “Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches.” Social Science Computer Review, February 21, 2020
- **Keyword Assisted Topic Models (keyATM)**
  - Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. “Keyword Assisted Topic Models.” ArXiv:2004.05964 [Cs, Stat], April 13, 2020
  - includes implementations for covariates similar to stm, and dynamic implementation for time variation
- **implementations beyond R**
  - CorEx: Gallagher et al
  - original SeededLDA: Jagarlamudi et al

→ **increasingly sophisticated branch of topic models**

Summary

Summary

Pros & Cons

Advantages

- no need for training data
- few decisions (topic number, model selection, ...)
- often intuitive results
- exciting developments with supervised topic models

Disadvantages

- topics are instable
  - random initialization
- topics are often not coherent
  - e.g. mixing of concepts
  - e.g. clustering of technical terms
- controversial as classification method

Summary

A pragmatic approach

"Still, excitement about the use of topic models for discovery needs to be tempered with skepticism about how often the unexpected juxtapositions LDA creates will be helpful, and how often merely surprising. **A poorly supervised machine learning algorithm is like a bad research assistant.** It might produce some unexpected constellations that show flickers of deeper truths; but it will also produce tedious, inexplicable, or misleading results. "

(Schmidt 2012, *Words Alone: Dismantling Topic Models in the Humanities*)

Summary

Additional resources on topic models

- structuraltopicmodel.com
- stm Vignette
- Roberts et.al. (2013)
- Roberts et.al.(2014)
- Roberts et.al. (2016)
- additional packages, e.g. `stmcorrViz`, `stminsights`, `oolong`, ...

Thank you! - Questions?