Day 10: Topic Models

ME314: Introduction to Data Science and Big Data Analytics

LSE Summer School

14 August 2018

Day 10 Outline

Topic models

Latent Dirichlet allocation (LDA)

Beyond Latent Dirichlet Allocation

Correlated and Dynamic Topic Models

Structural Topic Model

Recap

- Quantitative text analysis always requires:
  1. Construction of a quantitative matrix from textual features
  2. A quantitative or statistical procedure applied to that matrix
  3. Summary or interpretation of the results of that procedure
- Yesterday, we focused on two main statistical procedures
  1. Dictionary approaches
  2. Supervised approaches
- Today we move on to unsupervised methods
- Note that we will still need to make many of the same feature selection decisions as we did yesterday...

Topic models

Intro

- Topic models are algorithms for discovering the main "themes" in an unstructured corpus
- They require no prior information, training set, or labelling of texts before estimation
- They allows us to automatically organise, understand, and summarise large archives of text data.
  - Uncover hidden themes.
  - Annotate the documents according to themes.
  - Organise the collection using annotations.

What is a topic?

- **Google definition**: “a matter dealt with in a text, discourse, or conversation; a subject.“
- **Topic model definition**: a probability distribution over a fixed word vocabulary
- Consider a simple vocabulary: gene, dna, genetic, data, number, computer
- When speaking about **genetics**, you will:
  - frequently use the words “gene”, “dna” & “genetic”
  - infrequently use the words “data”, “number” & “computer”
- When speaking about **computation**, you will:
  - frequently use the words “data”, “number” & “computation”
  - infrequently use the words “gene”, “dna” & “genetic”

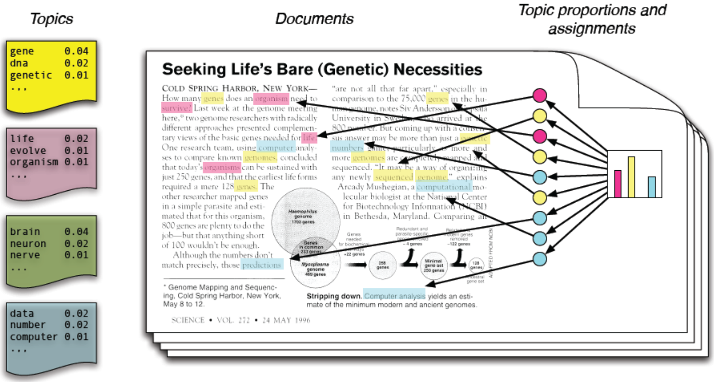
Topic	gene	dna	genetic	data	number	computer
Genetics	0.4	0.25	0.3	0.02	0.02	0.01
Computation	0.02	0.01	0.02	0.3	0.4	0.25

A motivating example

- Data: UK House of Commons' debates (PMQs)
  - $\approx$ 30000 parliamentary speeches from 1997 to 2015
  - $\approx$ 3000 unique words
  - $\approx 2m$ total words
- Note that I have already made a number of sample selection decisions
  - Only PMQs ($\approx$ 3% of total speeches)
  - Removed frequently occurring & very rare words
  - All words have been 'stemmed'
- Results of a 30-topic model: ▸ Link

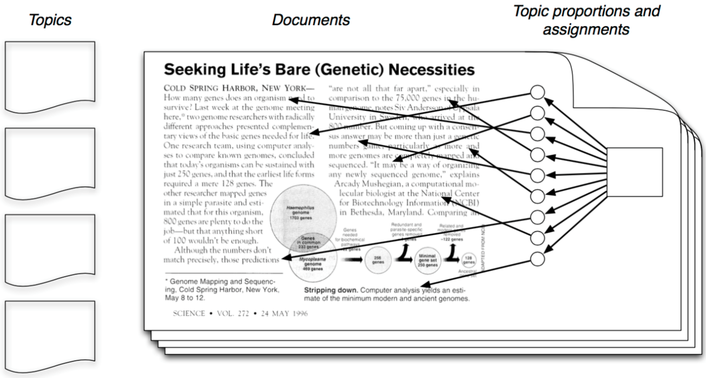
Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation (LDA)

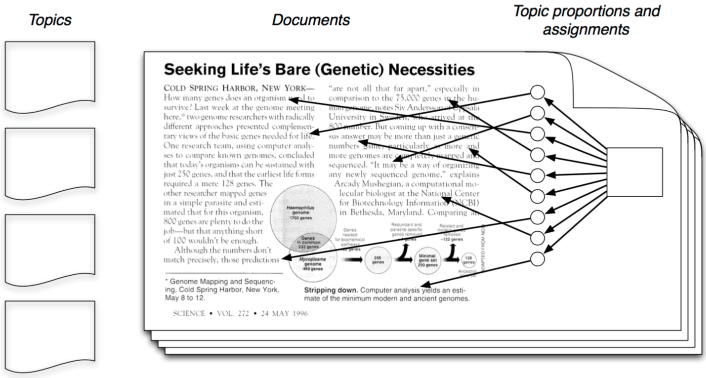


- Each topic is a distribution over words
- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics

Latent Dirichlet allocation (LDA)

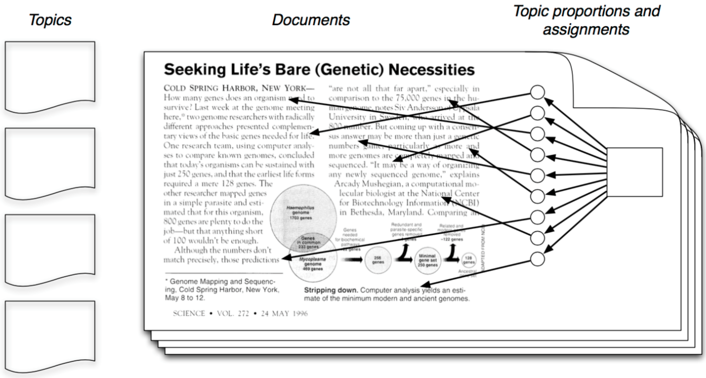


Latent Dirichlet allocation (LDA)



- In reality, we only observe the documents
- The other structure are hidden variables

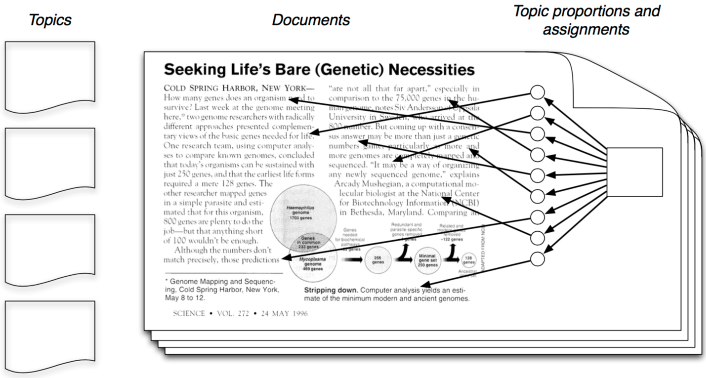
Latent Dirichlet allocation (LDA)



- Our goal is to infer the hidden variables
- I.e., compute their distribution conditioned on the documents

$$p(topics, proportions, assignments|documents)$$

Latent Dirichlet allocation (LDA)



- Topic modelling allows us to extrapolate backwards from a collection of documents to infer the "topics" that could have generated them.

Latent Dirichlet Allocation

- The LDA model is a Bayesian mixture model for discrete data where topics are assumed to be uncorrelated
- LDA provides a generative model that describes how the documents in a dataset were created
- Each of the K *topics* is a distribution over a fixed vocabulary
- Each document is a collection of words, generated according to a multinomial distribution, one for each of K topics
- Inference consists of estimating a posterior distribution over the parameters of the probability model from a combination of what is observed (words in documents) and what is hidden (topic and word parameters)

Latent Dirichlet Allocation: Details

- For each document, the LDA generative process is:
  1. randomly choose a distribution over topics (a multinomial of length K)
  2. for each word in the document
     2.1 Probabilistically draw one of the K topics from the distribution over topics obtained in step 1, say topic k (each document contains topics in different proportions)
     2.2 Probabilistically draw one of the V words from β_k (each individual word in the document is drawn from one of the K topics in proportion to the document's distribution over topics as determined in previous step)
- The goal of inference in LDA is to discover the topics from the collection of documents, and to estimate the relationship of words to these, *assuming this generative process*

LDA generative model

How to generate

1. Term distribution β for each topic is drawn:

$$\beta_k \sim \text{Dirichlet}(\eta)$$

   β_k describes topic k: it gives probability that each word occurs in a given topic

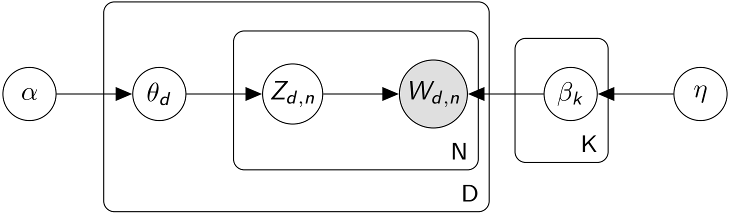
2. proportions θ of the topic distribution for the document are drawn by

$$\theta_d \sim \text{Dirichlet}(\alpha)$$

   θ_d describes topic d: it gives probability that each topic occurs in a given document

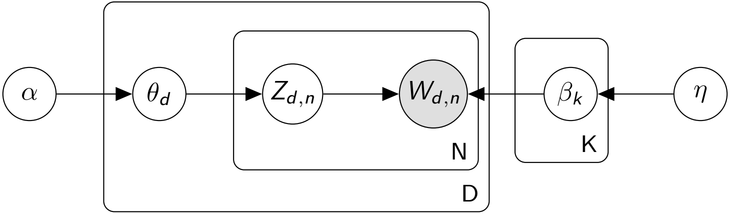
3. For each of the N words in each document
   - choose a topic $z_i \sim \text{Multinomial}(\theta)$
   - choose a word $w_i \sim \text{Multinomial}(p(w_i|z_i, \beta))$

LDA as a graphical model



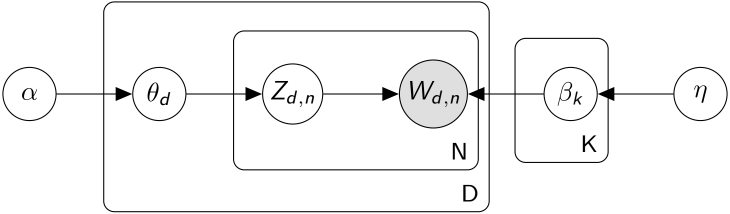
- Encodes assumptions
- Connects to algorithms for computing with data

LDA as a graphical model



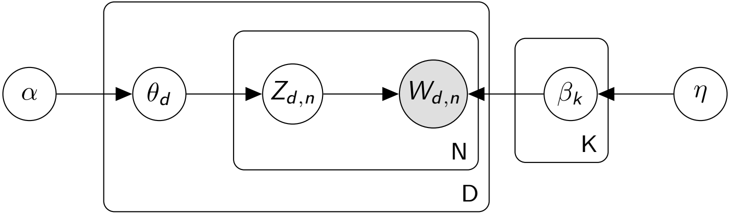
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

LDA as a graphical model



- α proportions parameter (corpus level)
- η topic parameter (corpus level)
- β_k probability distribution over words (topic level)
- θ_d topic proportions (document level)
- $Z_{d,n}$ topic assignment (word level)
- $W_{d,n}$ observed word (word level)

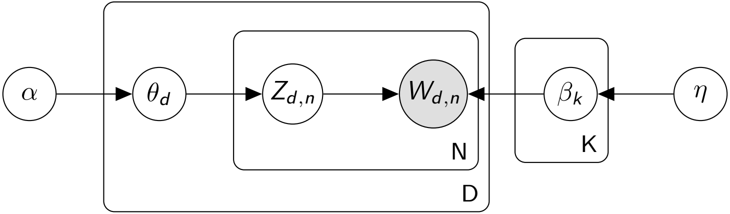
LDA as a graphical model



- $\beta_k \sim \text{Dirichlet}(\eta)$
- $\theta_d \sim \text{Dirichlet}(\alpha)$
- $Z_{d,n} \sim \text{Multinomial}(\theta_d)$
- $W_{d,n} \sim \text{Multinomial}(p(w_i|z_i, \beta_k))$

Note: η and α govern the *sparsity* of the draws from the dirichlet. As they $\rightarrow 0$, the multinomials become more sparse.

LDA as a graphical model

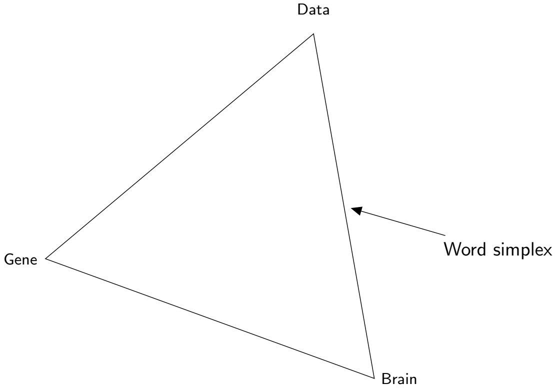


- This joint defines a posterior, $p(\theta, z, \beta | w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior distribution over these parameteres to perform the task at hand $\rightarrow$ information retrieval, document similarity, exploration, and others.

The Dirichlet distribution

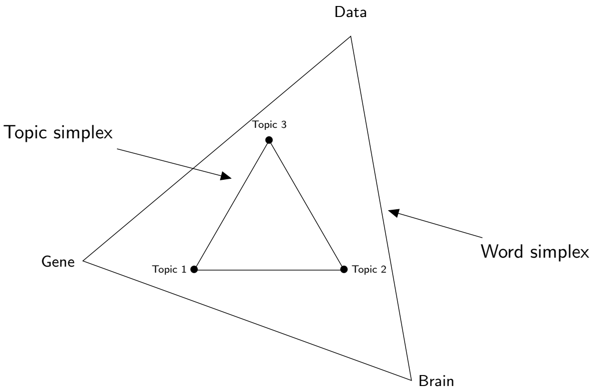
- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one
- The Dirichlet is used twice in LDA:
  - The topic proportions (θ) are a K dimensional Dirichlet
  - The topics (β) are a V dimensional Dirichlet.
- Estimation is performed using collapsed Gibbs sampling and/or Variational Ecpectation-Maximization (VEM)
- Fortunately, for us these are easily implemented in `R`

Latent Dirichlet allocation (LDA)



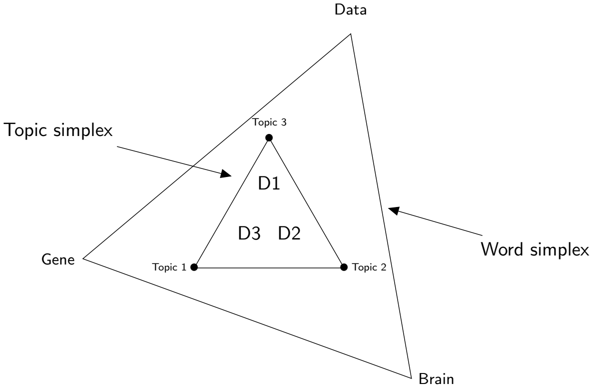
- Imagine a corpus consisting of only three words
- The word simplex describes the possible probabilities of the multinomial distribution over these three words

Latent Dirichlet allocation (LDA)



- We can locate topics within the word-simplex
- Each topic represents a different distribution over words
- (The smaller is η, the more sparse the topics, the closer they will be to the word-simplex lines)

Latent Dirichlet allocation (LDA)

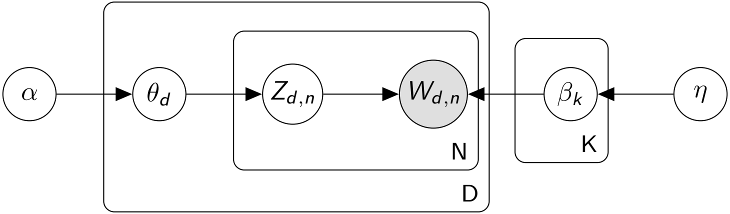


- The topic simplex describes the possible probabilities of the multinomial distribution over these topics
- We can locate documents within the topic-simplex
- Each document is a mixture of topics
- (Smaller $\alpha \rightarrow$ sparser documents $\rightarrow$ documents will be closer

Why does LDA "work"?

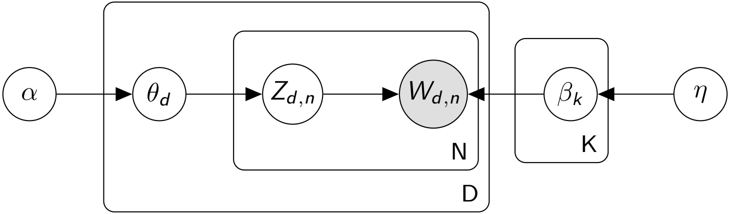
- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible. (α)
  2. For each topic, assign high probability to as few terms as possible. (η)
- These goals are at odds.
  - Putting a document in a single topic makes (2) hard: All of its words must have probability under that topic.
  - Putting very few words in each topic makes (1) hard: To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.

LDA summary



- LDA is a probabilistic model of text. It casts the problem of discovering themes in large document collections as a posterior inference problem.
- It lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure.

LDA summary

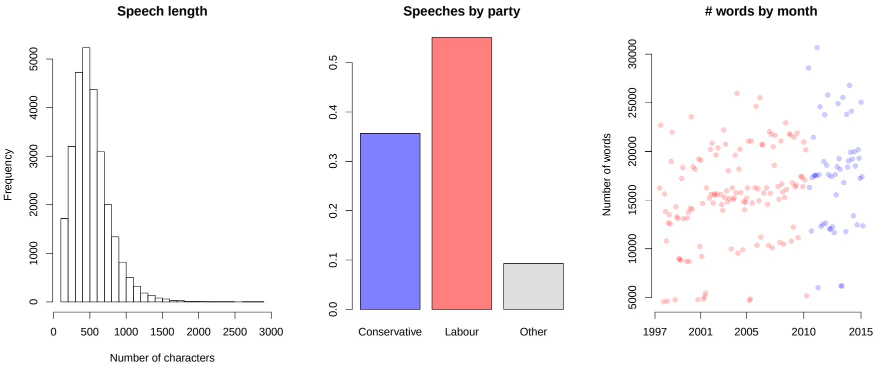


- LDA is a simple building block that enables many applications.
- It is popular because organizing and finding patterns in data has become important in the sciences, humanities, industry, and culture.
- Further, algorithmic improvements let us fit models to massive data.

LDA example

- Data: UK House of Commons' debates (PMQs)
  - $\approx$ 30000 parliamentary speeches from 1997 to 2015
  - $\approx$ 3000 unique words
  - $\approx 2m$ total words
- Note that I have already made a number of sample selection decisions
  - Only PMQs ($\approx$ 3% of total speeches)
  - Removed frequently occurring & very rare words
  - All words have been 'stemmed'
- Estimate a range of topic models ($K \in \{20, 30, ..., 100\}$) using the `topicmodels` package

LDA example


Speech length
Frequency
0
1000
2000
3000
4000
5000
0
500
1000
1500
2000
2500
3000
Number of characters
Speeches by party
0.0
0.1
0.2
0.3
0.4
0.5
Conservative
Labour
Other
words by month
Number of words
5000
10000
15000
20000
25000
30000
1997
2001
2005
2010
2015

Implementation in R (via quanteda)

```r
## Create corpus
speechCorpus <- corpus(pmq$Speech, docvars = pmq)

## Create DFM
speechDFM <- dfm(speechCorpus,
      remove = stopwords("en"), stem = T)

## Remove very infrequent words
speechDFM <- dfm_trim(speechDFM, min_termfreq = 5)

## Convert for usage in 'topicmodels' package
tmDFM <- convert(speechDFM, to = 'topicmodels')

## Set topic count and estimate LDA using Gibbs sampling
K <- 60
ldaOut <- LDA(tmDFM, k = K, method = "Gibbs",
                control = list(seed = 123))

## Save (because it can take a long time to run again!)
save(ldaOut, file = paste0("ldaOut",K,".Rdata"))
```

LDA example

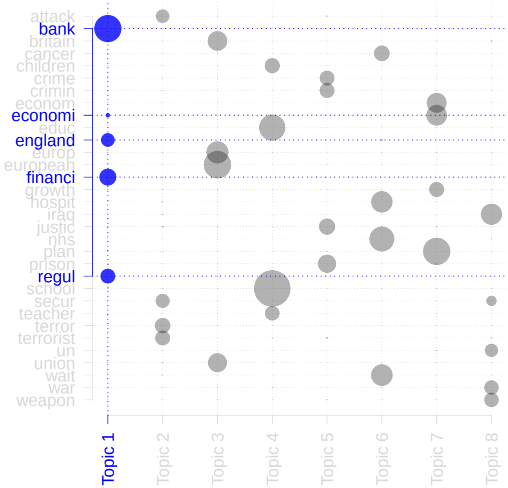
We will make use of the following score to visualise the posterior topics:

$$\text{term-score}_{k,v} = \hat{\beta}_{k,v} \log \left(\frac{\hat{\beta}_{k,v}}{\left(\prod_{j=1}^{K} \hat{\beta}_{j,v}\right)^{\frac{1}{K}}} \right) \tag{1}$$

This formulation is similar to the TFIDF term score, where

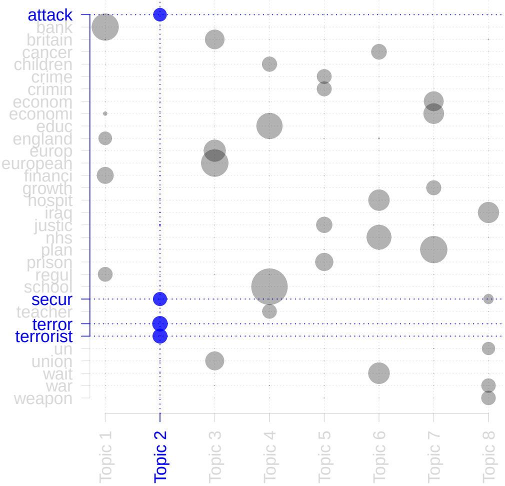
- the first term, $\hat{\beta_{k,v}}$, is the probability of term v in topic k and is akin to the term frequency
- the second term is akin to the document frequency (i.e. it down-weights terms that have high probability under all topics)

LDA example



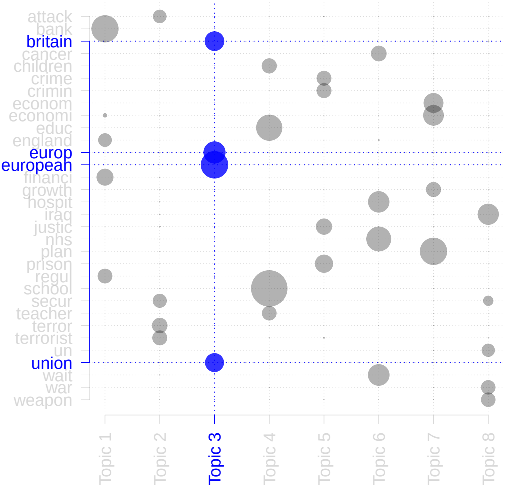
attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

LDA example

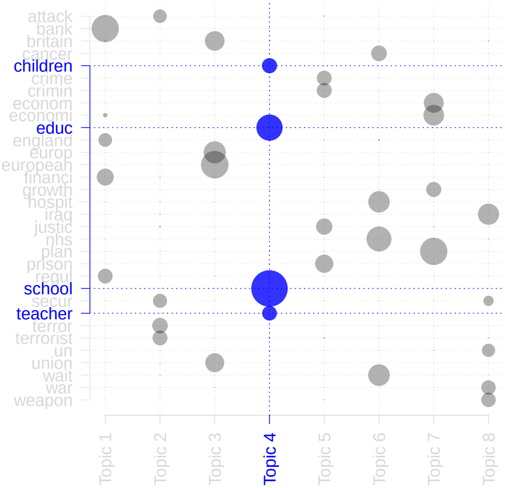


attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

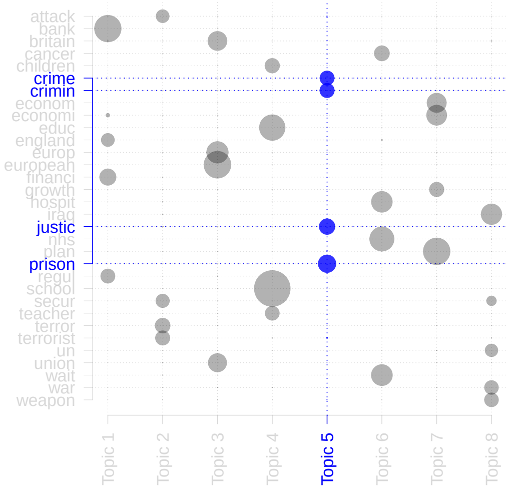
LDA example


attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

LDA example

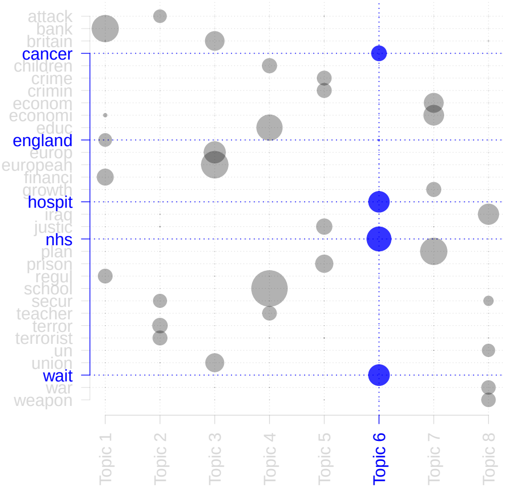

attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

LDA example

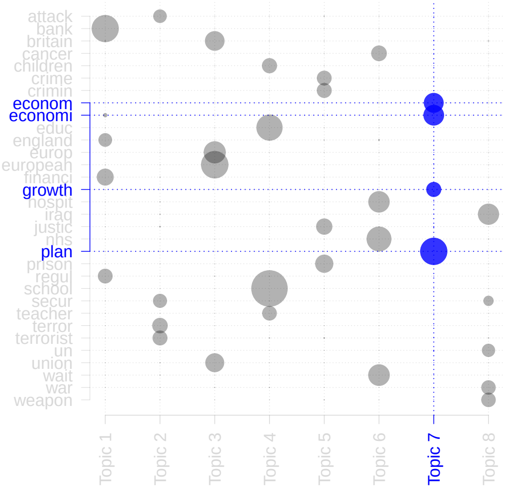


attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

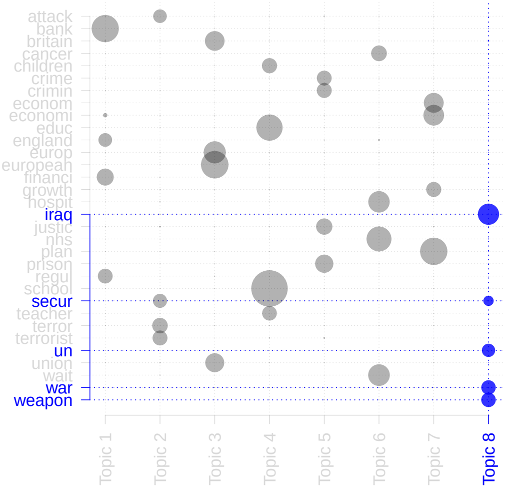
LDA example


attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

LDA example


attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

LDA example


attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

LDA example

Topic 1	Topic 2	Topic 3	Topic 4
bank	terror	european	school
financi	terrorist	europ	educ
regul	secur	britain	children
england	attack	union	teacher
crisi	protect	british	pupil
fiscal	agre	referendum	class
market	act	constitut	parent

Topic 5	Topic 6	Topic 7	Topic 8
prison	nhs	plan	iraq
justic	wait	economi	weapon
crimin	hospit	econom	war
crime	cancer	growth	un
releas	patient	grow	resolut
court	list	longterm	iraqi
sentenc	health	deliv	saddam

Evaluating LDA performance

How can we tell how well a given topic model is performing?

- How well does the model predict held-out data?
- Standard approach: Ask which words the model believes will be in a given document and comparing this to the document's actual word composition (normally by splitting texts in half)
- Problems:
  - Prediction is not always important in exploratory or descriptive tasks. We may want models that capture other aspects of the data.
  - There tends to be a negative correlation between quantitative diagnostics such as these and human judgements of topic coherence!

Evaluating model performance

Chang, Jonathan et al. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in neural information processing systems*.

Uses human evaluation of:

- whether a topic has (human-identifiable) semantic coherence: word intrusion, asking subjects to identify a spurious word inserted into a topic
- whether the association between a document and a topic makes sense: topic intrusion, asking subjects to identify a topic that was not associated with the document by the model

Finding:

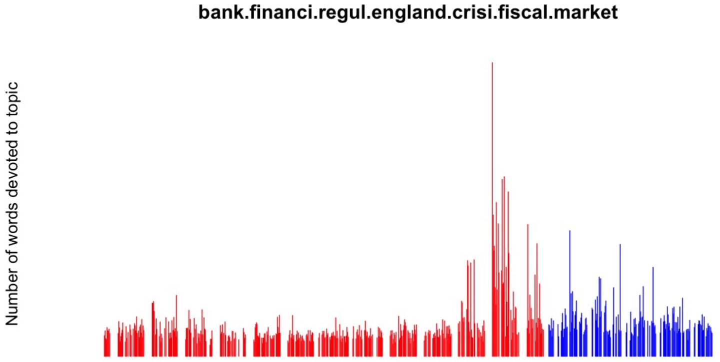
> *"Topic models which perform better on held-out likelihood may infer less semantically meaningful topics."*

Evaluating LDA performance

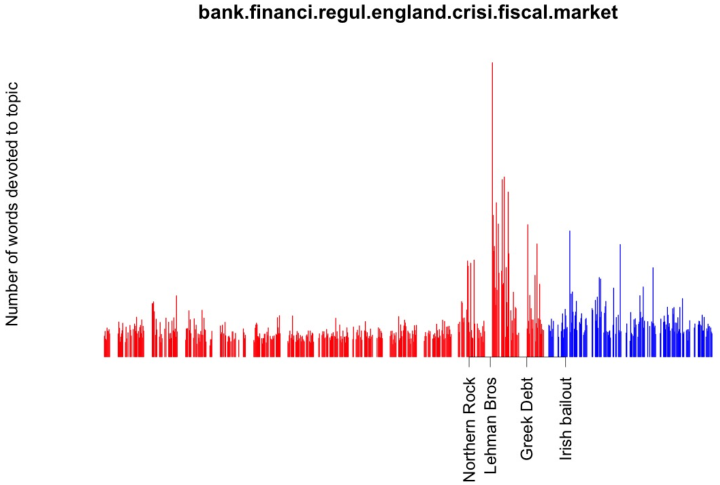
- *Semantic validity* – does a topic identify a coherent groups of texts that are internally homogenous but distinctive from other topics?
- *Predictive validity* – how well does variation in topic usage correspond to known events?
- *Construct validity* – how well does our measure correlate with other measures?

Here, we will focus on semantic and predictive validity. Why?

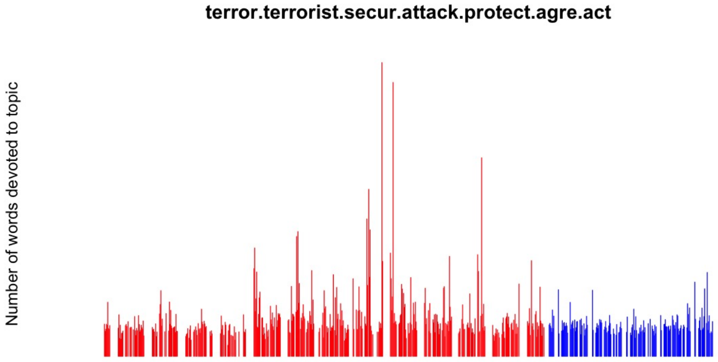
Predictive validity



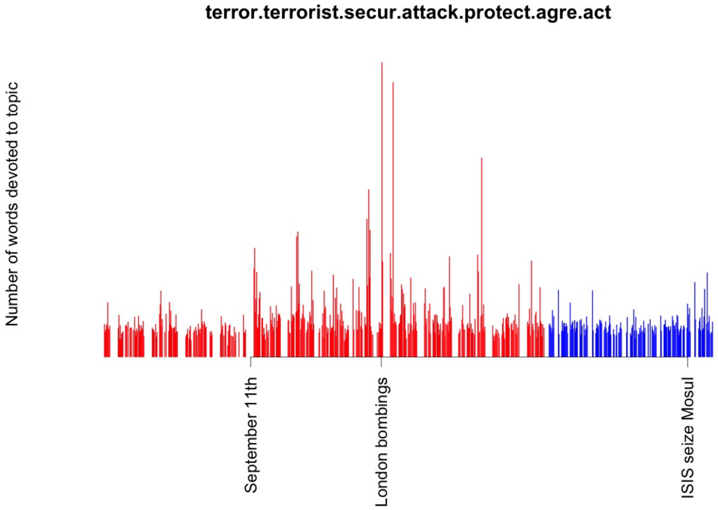
Predictive validity



Predictive validity



Predictive validity



Semantic validity

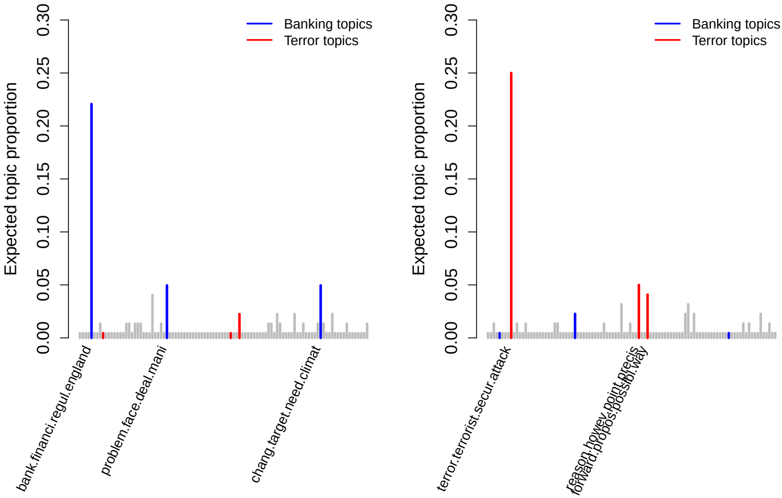
Consider the following texts:

```
The reforms that we are bringing into the banking system will include greater competition in
banking. We will have a judgment from the European Commission soon, which we are supporting, that
will allow more competition in British banking. As for the restructuring of the banking system and
whether there should be investment banks on one side and retail-only banks on the other, the right
hon. Gentleman must remember that Northern Rock was effectively a retail bank and it collapsed.
Lehman Brothers was effectively an investment bank without a retail bank and it collapsed. The
difference between retail and investment banks is not the cause of the problem. The cause of the
problem is that banks have been insufficiently regulated at a global level and we have to set the
standards for that for the future. We will be doing that at the G20 Finance Ministers summit in a
few weeks' time.
```

```
The purpose of this coming before the House is for the Home Secretary to advise us that, in her
view, there is an exceptional terrorist threat a grave terrorist threat that either has occurred or
is occurring and that the need for action is urgent, but that it has not been possible to assemble
the necessary evidence to lay charges within the 28 days. It will then be for the House to vote on
the commencement order and agree that an exceptional terrorist incident has occurred. It is not the
business of the House to interfere in the individual case, but it should be able to vote simply on
whether an exceptional and grave terrorist threat has occurred. Given that the right hon. Gentleman
and others have referred to the Civil Contingencies Act 2004 in discussing this issue, I would hope
that he understands this is exactly the same problem that has to be faced in respect of that
Act.
```

We will call these the banking and terrorism texts.

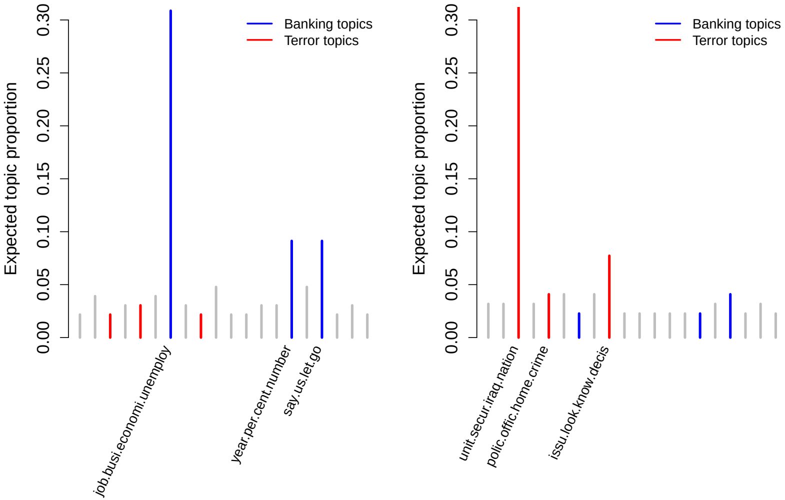
Semantic validity



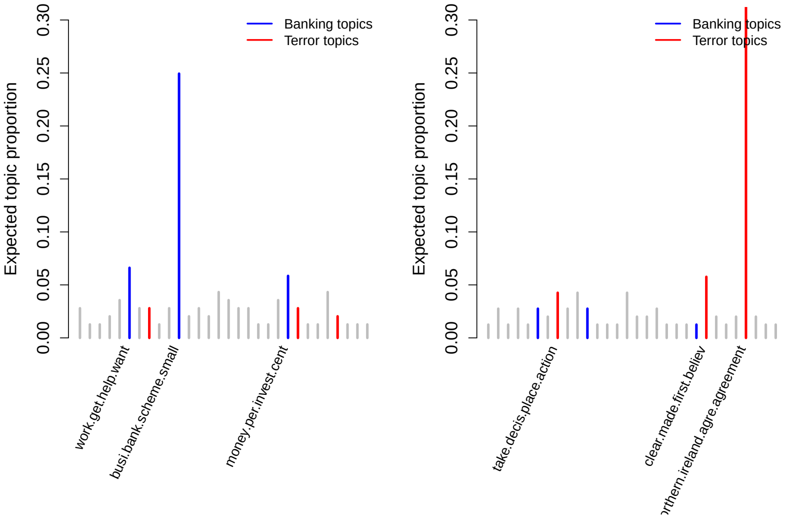
Semantic validity

- These plots suggest that our model is picking up at least some properties that we would intuitively expect to see in this particular corpus
- However, they do not help us to choose between the different models that we have estimated
- In other words, how should we pick K?

Which K? 20...

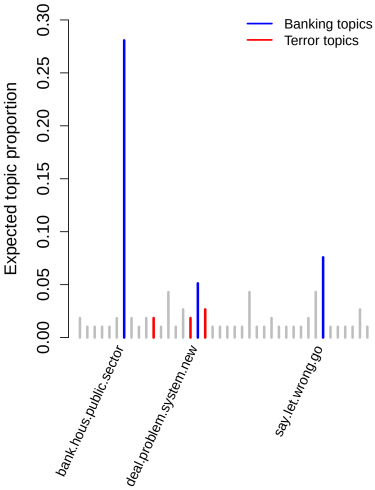

Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
job.busi.economi.unemploy
year.per.cent.number
say.us.let.go
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
unit.secur.iraq.nation
polic.offic.home.crime
issu.look.know.decis

Which K? 30...

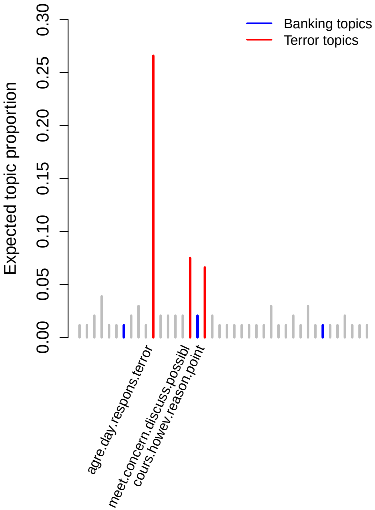


Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
work.get.help.want
busi.bank.scheme.small
money.per.invest.cent
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
take.decis.place.action
clear.made.first.believ
orthern.ireland.agre.agreement

Which K? 40...

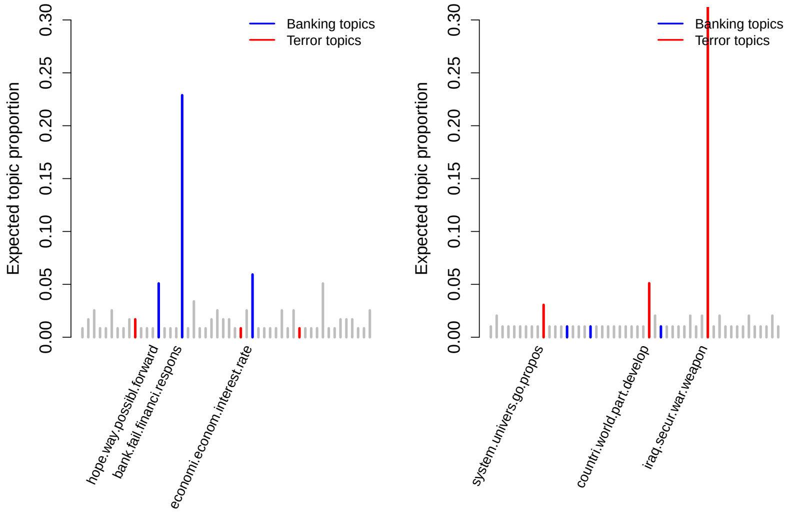


Banking topics
Terror topics
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
bank.hous.public.sector
deal.problem.system.new
say.let.wrong.go



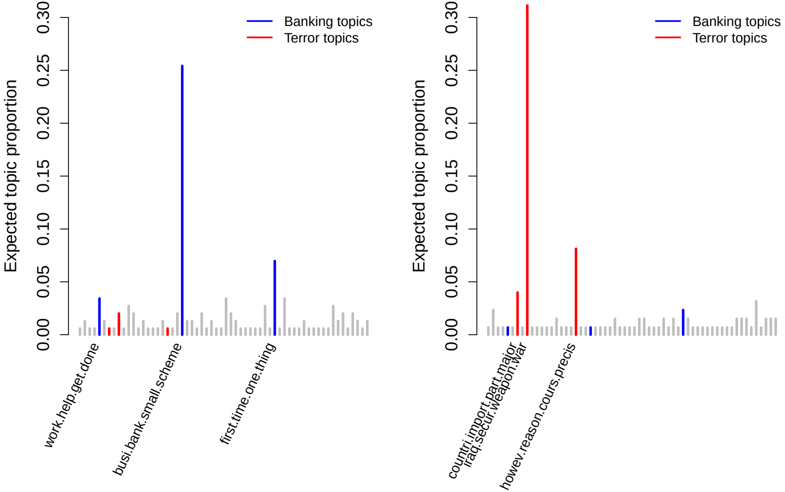
Banking topics
Terror topics
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
agre.day.respons.terror
meet.concern.discuss.possibl
cours.howev.reason.point

Which K? 50...

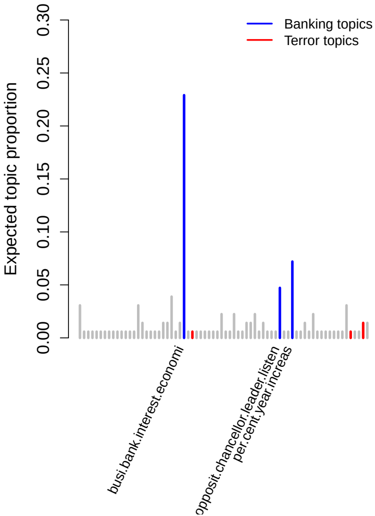


Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
hope.way.possibl.forward
bank.fail.financi.respons
economi.econom.interest.rate
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
system.univers.go.propos
countri.world.part.develop
iraq.secur.war.weapon

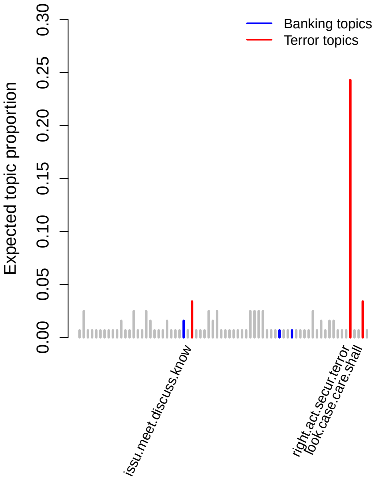
Which K? 60...


Banking topics
Terror topics
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
work.help.get.done
busi.bank.small.scheme
first.time.one.thing
Banking topics
Terror topics
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
countri.import.part.major
iraq.secur.weapon.war
howev.reason.cours.precis

Which K? 70...

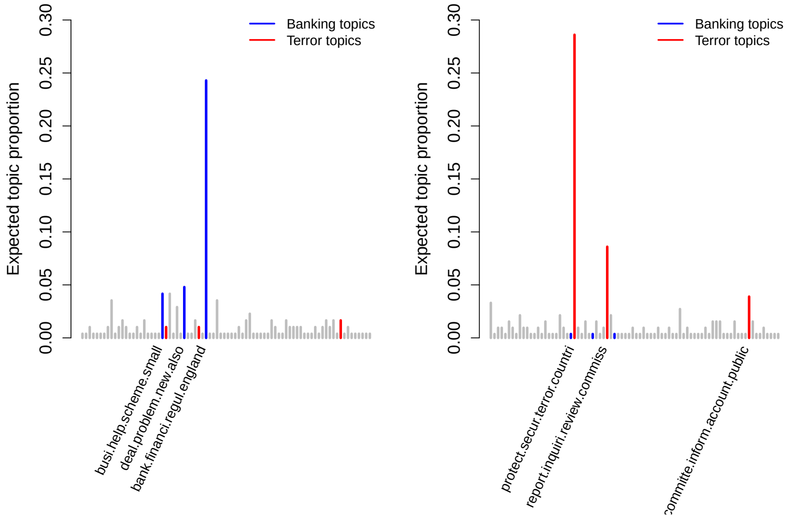


Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
busi.bank.interest.economi
opposit.chancellor.leader.listen
per.cent.year.increas

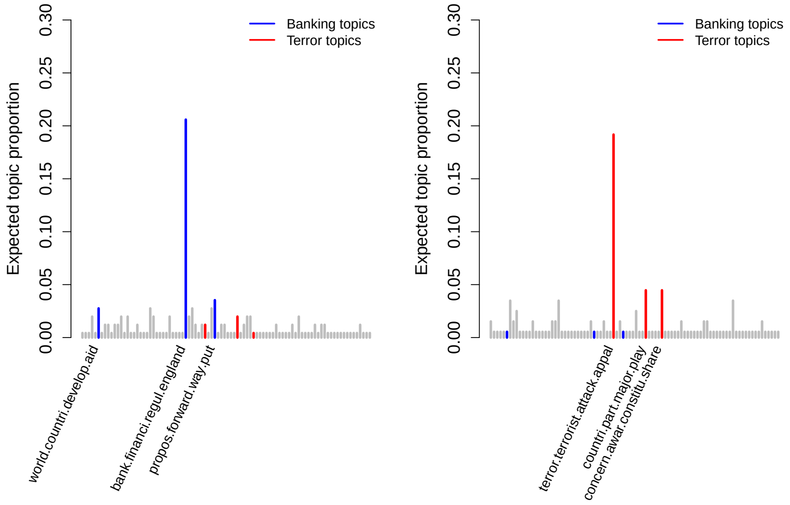


Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
issu.meet.discuss.know
right.act.secur.terror
look.case.care.shall

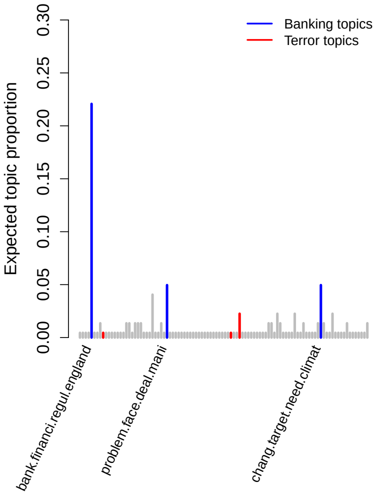
Which K? 80...

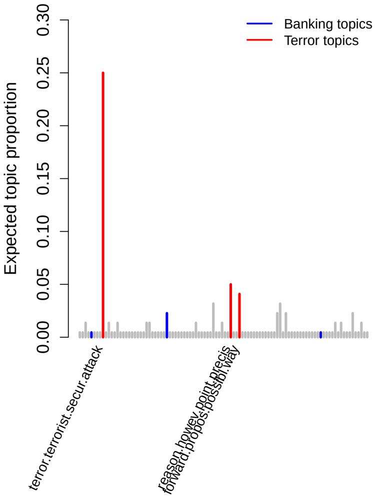

Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
busi.help.scheme.small
deal.problem.new.also
bank.financi.regul.england
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
protect.secur.terror.countri
report.inquiri.review.commiss
committe.inform.account.public

Which K? 90...


Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
world.countri.develop.aid
bank.financi.regul.england
propos.forward.way.put
Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
terror.terrorist.attack.appal
countri.part.major.play
concern.awar.constitu.share

Which K? 100...


Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
bank.financi.regul.england
problem.face.deal.mani
chang.target.need.climat


Expected topic proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Banking topics
Terror topics
terror.terrorist.secur.attack
reason.howev.point.precis
forward.propos.possibl.way

Semantic validity (revisited)

- Semantic validity requires that the topics are coherent and meaningful.
- We hope that texts assigned to a given topic are homogenous
- We hope that texts from different topics are distinctive
- This is a strong test of a topic model, but requires human input, which can be costly
- One option here would be to crowdsource the validation task to online workers
- Another option is to mercilessly exploit a class of participants

Semantic validity (revisited)

Related?

Unrelated

NEXT COMPARISON.

Comparisons completed: 0

Text one:

That is total complacency about one month s figures when the Prime Minister has had five years of failure under this Government. Under this Prime Minister we are a country of food banks and bank bonuses; a country of tax cuts for millionaires while millions are paying more. Is not his biggest broken promise of all that we are all in it together?

Text two:

This is totally desperate stuff because the Prime Minister has nothing to say about the cost of living crisis. That is the reality, and his reshuffle had nothing to do with the country and everything to do with his party. After four years of this Government, we have a recovery that people cannot feel, a cost of living crisis that people cannot deny, and a Prime Minister whom people cannot believe.

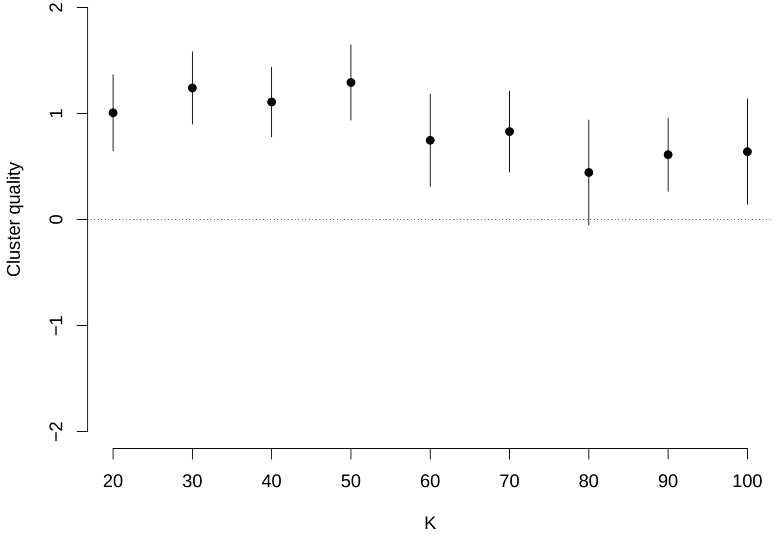
Semantic validity (revisited)

- Sample pairs of speeches from the posterior distribution
  - 5 pairs from the same topic, for each topic
  - 5 pairs from different topics, for each topic
- Randomly present to human coders, asking whether they are:
  - closely related (3)
  - loosely related (2)
  - unrelated (1)
- Calculate the 'Cluster Quality' for the topic by regressing

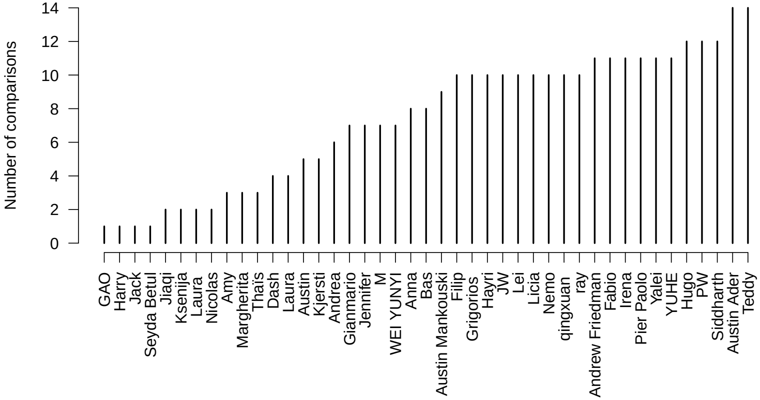
$$Related_{ik} = \alpha + \beta_k * SameTopic_{ik} + \gamma \tag{2}$$

- β_k is an estimate of the cluster quality of topic model k
  - i.e. the difference between relatedness of same-topic and different-topic pairs
- γ is a coder fixed-effect (Why?)
- Repeat for each value of K

Semantic validity (revisited)



Semantic validity (revisited)



An application

- Once we are happy with the topic model we have estimated, we can use the posterior distribution in various ways
  - Visualisation
  - Information retrieval
  - Corpus exploration
  - Similarity
  - Dimensionality reduction
- In this example, we can use the posterior distribution of document-topic proportions to ask: Which MPs are most active at asking questions in each topic?

$$MPAttention_{i,k} = \frac{MPWords_{i,k}}{\sum_1^K MPWords_{i,k}} \tag{3}$$

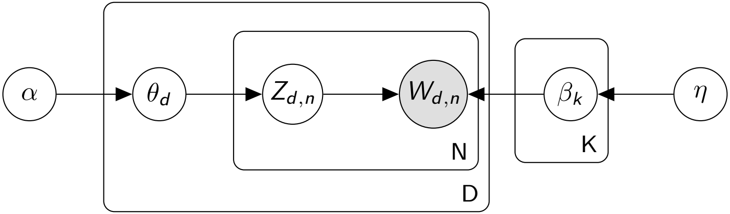
An application

bank.financi.regul	terror.terrorist.secur	european.europ.britain	school.educ.children
Christopher Gill	Shahid Malik	David Lock	Christine Butler
Malcolm Wicks	Parmjit Gill	William Cash	Melanie Johnson
Tom Greatrex	George Mudie	Alistair Darling	Julie Kirkbride
Alasdair Morgan	Jonathan Djanogly	Denzil Davies	Sam Gyimah
Nick Herbert	James Brokenshire	David Heathcoat−Amory	Malcolm Moss
Karl McCartney	Tobias Ellwood	David Wilshire	Paul Clark
Donald Gorrie	Brian Wilson	David Davis	Ian Liddell−Grainger
Justin Tomlinson	John Maples	Giles Radice	Michael Heseltine
John Townend	Seamus Mallon	Ann Winterton	Stephen Hammond
Howard Flight	Ann Keen	Jenny Jones	Chris Pond
Derek Foster	Stephen Barclay	Dale Campbell−Savours	Ivan Henderson
Lindsay Roy	Pat McFadden	Jacob Rees−Mogg	Derek Conway

prison.justic.crimin	nhs.wait.hospit	plan.economi.econom	iraq.weapon.war
Jack Lopresti	Julia Goldsworthy	Chloe Smith	Alan Howarth
Kevin McNamara	Seema Malhotra	Conor Burns	Chris Smith
Alan Clark	Michael Penning	Donald Gorrie	Tony Worthington
Chris Skidmore	Nick Hurd	Guy Opperman	Terry Davis
Charles Walker	Virendra Sharma	Karen Bradley	George Foulkes
Jeremy Wright	Tim Farron	Neil Carmichael	Jonathan Sayeed
Tess Kingham	Bill Esterson	Wayne David	Melanie Johnson
Sarah Champion	John Penrose	Michael Colvin	Denzil Davies
Philip Davies	Malcolm Chisholm	Michael Ellis	Paul Stinchcombe
Kali Mountford	Grant Shapps	Anne Milton	Adam Price
Mike Wood	Marion Roe	John Stevenson	Kevin Hughes
Lynda Waltho	Mike Thornton	Sarah Newton	Tony Benn

Beyond Latent Dirichlet Allocation

LDA summary



- LDA is a simple topic model.
- It can be used to find topics that describe a corpus.
- Each document exhibits multiple topics.
- There are several ways to extend this model.

Extending LDA

- LDA can be embedded in more complicate models, embodying further intuitions about the structure of the texts.
- E.g., it can be used in models that account for syntax, authorship, word sense, dynamics, correlation, hierarchies, and other structure.
- The data generating distribution can be changed. We can apply mixed-membership assumptions to many kinds of data.
- E.g., we can build models of images, social networks, music, purchase histories, computer code, genetic data, and other types.
- The posterior can be used in creative ways.
- E.g., we can use inferences in information retrieval, recommendation, similarity, visualization, summarization, and other applications.

Extending LDA

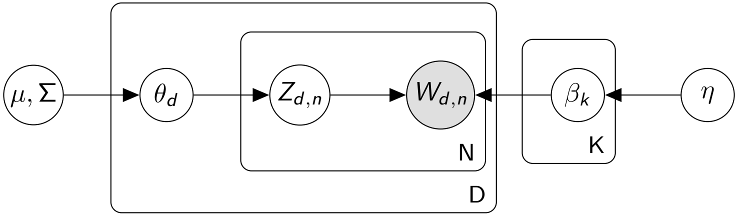
- These different kinds of extensions can be combined.
- To give a sense of how LDA can be extended, we'll look at several examples of major extensions.
- We will discuss
  - Correlated topic models
  - Dynamic topic models
  - Structural topic models

Correlated and Dynamic Topic Models

Correlated topic models

- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about fossil fuels is more likely to also be about geology than about genetics.
- The logistic normal is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- Re-parameterise so that the (log of the) parameters of the topic-proportions multinomial are drawn from a multivariate Gaussian distribution

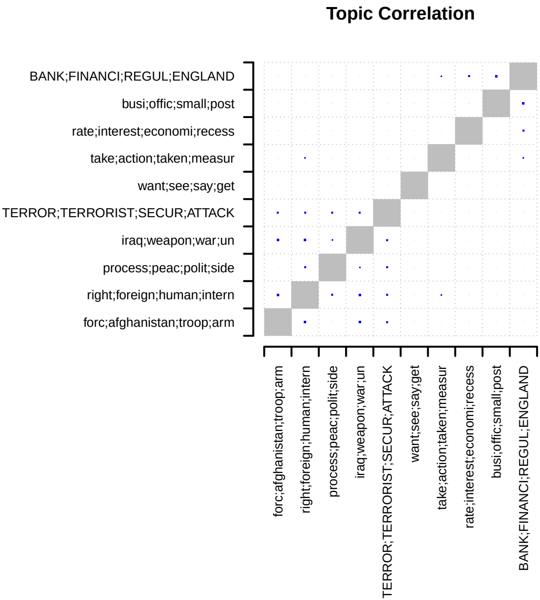
Correlated topic models



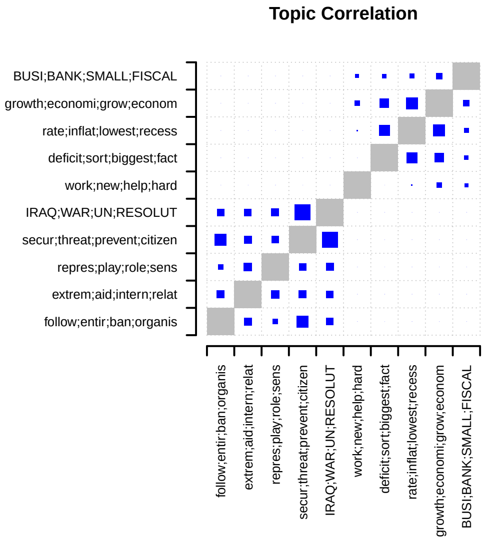
where the first node is logistic normal prior.

- Draw topic proportions from a logistic normal.
- This allows topic occurrences to exhibit correlation.
- Provides a "map" of topics and how they are related
- Provides a better fit to text data, but computation is more complex

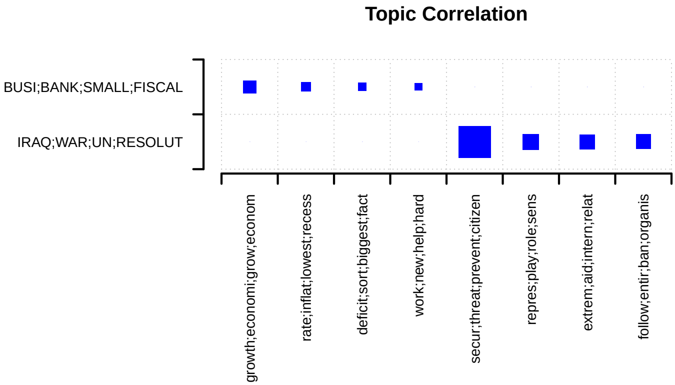
LDA topic correlation



CTM topic correlation



CTM topic correlation



Dynamic topic models

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- We may want to track how language changes over time.
  - How has the language used to describe neuroscience developed from "The Brain of Professor Laborde" (1903) to "Reshaping the Cortical Motor Map by Unmasking Latent Intracortical Connections" (1991)
  - How has the language used to describe love developed from "Pride and Prejudice" (1813) to "Eat, Pray, Love" (2006)
- Dynamic topic models let the topics drift in a sequence.

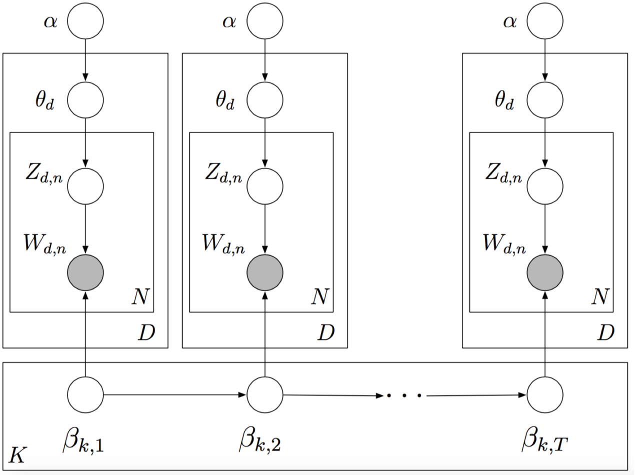
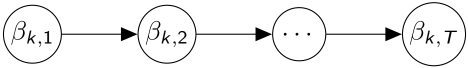


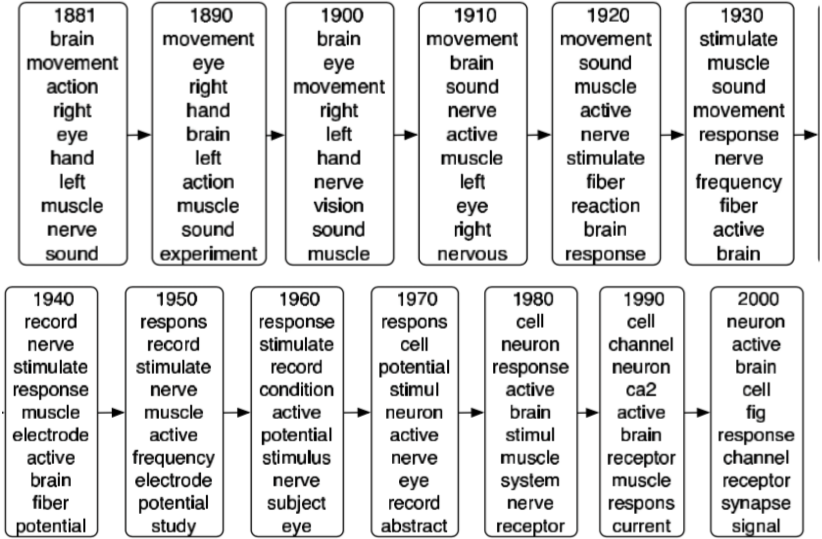
Plate (K) is topics drift through time.

Dynamic topic models



- Use a logistic normal distribution to model topics evolving over time.
  - The kth topic at time 2 has evolved smoothly from the kth topic at time 1
- As for CTMs, this makes computation more complex. But it lets us make inferences about sequences of documents.

Dynamic topic models



Summary: Correlated and dynamic topic models

- The Dirichlet assumption on topics and topic proportions makes strong conditional independence assumptions about the data.
- The correlated topic model uses a logistic normal on the topic proportions to find patterns in how topics tend to co-occur.
- The dynamic topic model uses a logistic normal in a linear dynamic model to capture how topics change over time.
- What's the catch? These models are harder to compute.

Structural Topic Model

Structural Topic Model

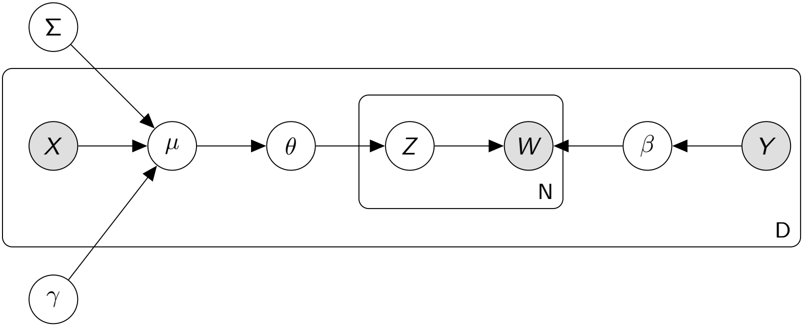
- Typically, when estimating topic models we are interested in how some covariate is associated with the prevalence of topic usage (Gender, date, political party, etc)
- The Structural Topic Model (STM) allows for the inclusion of arbitrary covariates of interest into the generative model
- The addition of covariates provides structure to the prior distributions
  1. Benefit 1: improves the estimation of the topics by allowing documents to share information according to the covariates (known as 'partial pooling' of parameters)
  2. Benefit 2: the relationship between covariates and latent topics is most frequently the estimand of interest, so we should include this in the estimation procedure

Structural Topic Model

How does it differ from LDA?

- As with the CTM, topics within the STM can be **correlated**
- **Topic prevalence** is allowed to vary according to the covariates X
  - Each document has its own prior distribution over topics, which is defined by its covariates, rather than sharing a global mean
- **Topical content** can also vary according to the covariates Y
  - Word use *within* a topic can differ for different groups of speakers/writers

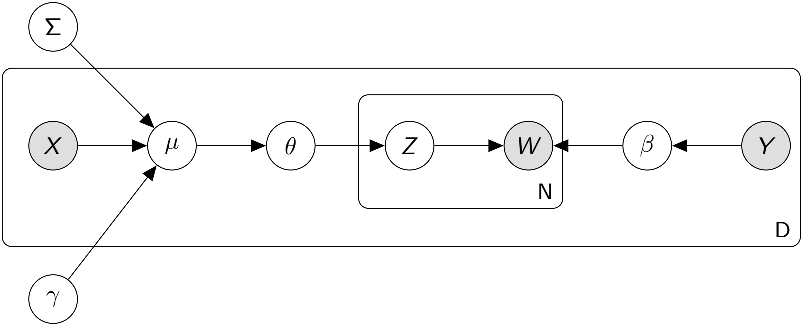
Structural topic model



Topic prevalence model:

- Draw topic proportions from a logistic normal generalised linear model based on covariates X
- This allows the expected document-topic proportions to vary by covariates, rather than from a single shared prior

Structural topic model



Topical content model:

- The β coefficients, which indicate the distribution over words for a given topic, are allowed to vary according to the covariates Y
- This allows us to estimate how different covariates affect the words used *within a given topic*

Structural Topic Model – example

- In the legislative domain, we might be interested in the degree to which MPs from different parties represent distinct interests in their parliamentary questions
- We can use the STM to analyse how topic prevalence varies by party

```

## Set topic count and estimate STM
K <- 60
stmOut <- stm(
         documents = speechDFM,
         data= docvars(speechDFM),
         prevalence = ~party,
         content = ~party,
         K = K,
         seed = 123)
```

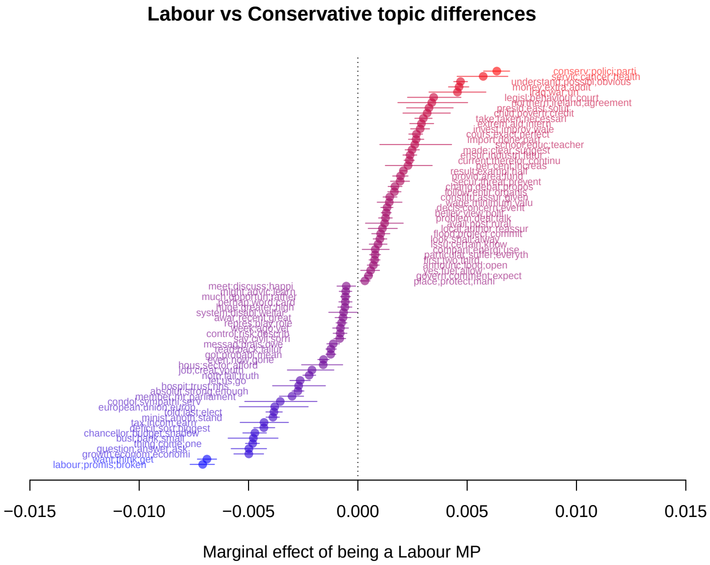
Structural Topic Model – example

- Specify a linear model with:
  - the topic proportions of speech d, by legislator i as the outcome
  - the party of legislator i as the predictor

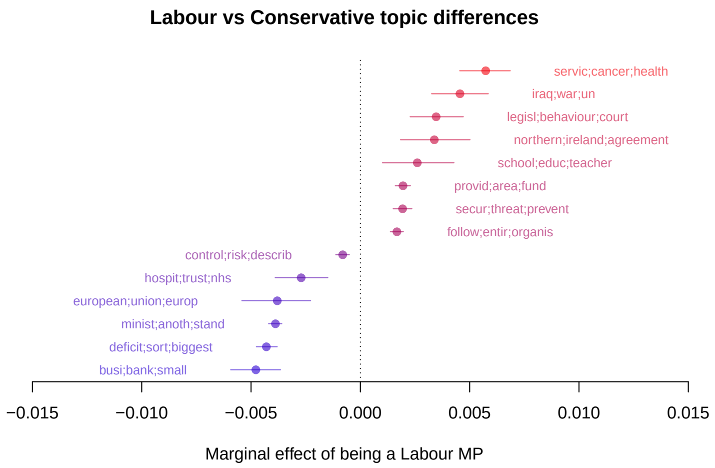
$$\theta_{dk} = \alpha + \gamma_{1k} * \text{labour}_{di} \tag{4}$$

- The γ_k coefficients give the estimated difference in topic proportions for Labour and Conservative legislators for each topic

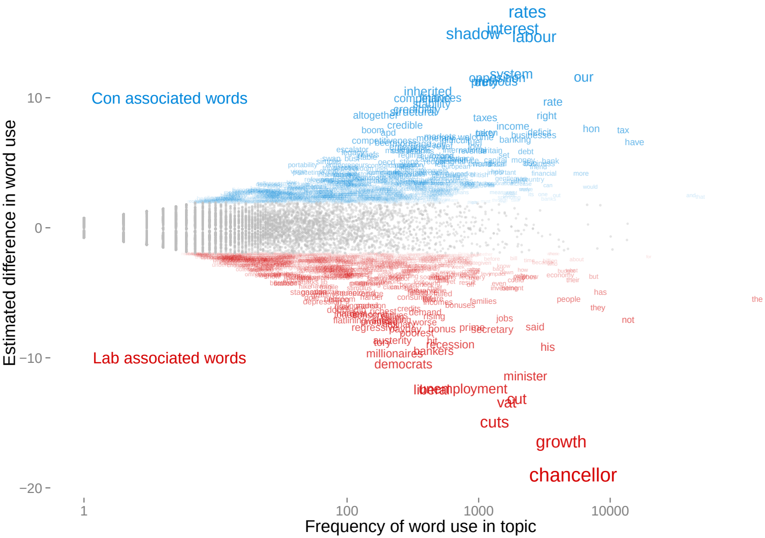
Topic prevalence



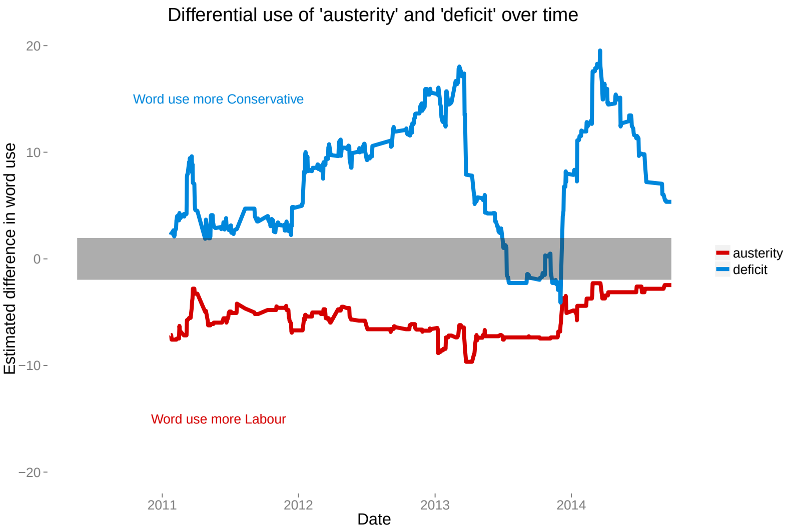
Topic prevalence



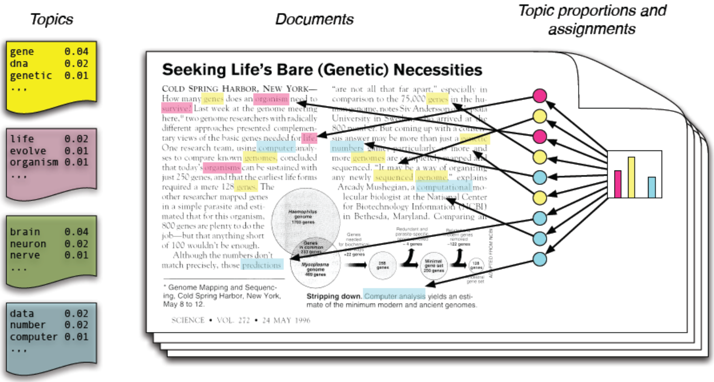
Topical content



Topical content

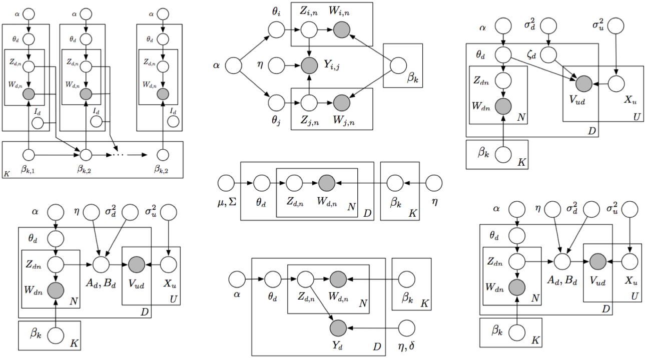


Summary



- LDA assumes that there are K topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.

Summary



Topic models can be adapted to many settings

- relax assumptions
- combine models
- model more complex data

Implementations of topic models in R

Incomplete list:

- `topicmodels`
- `lda`
- `stm`
- `mallet`