Topic models: evaluation

William Lowe

Hertie School

14th October 2020

Plan

→ Evaluation

→ Statistical model evaluation

→ Human model evaluation

→ Statistical topic evaluation

→ Human topic evaluation

→ Case study

Evaluation

There are two main modes of evaluation:

- → Statistical
- → Human / substantive

and two natural levels

- → The models as a whole
- → The topics it creates

Overall message: These are not yet well aligned

- → We will emphasize substance and topics

Held-out likelihood

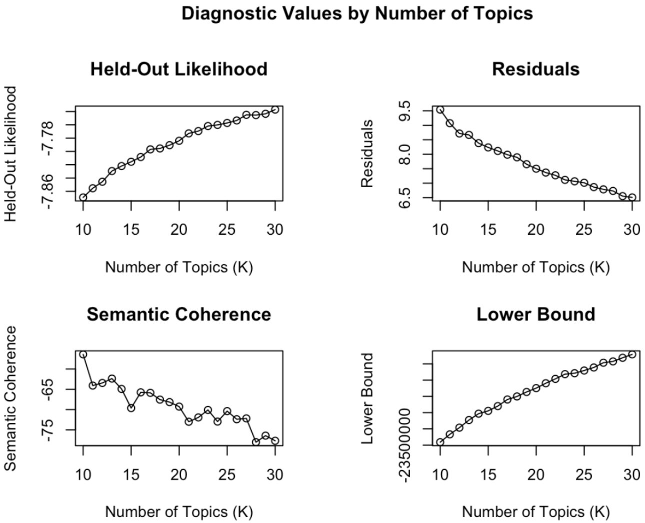
Since documents are assumed to be bags of words, then we can

- set aside some proportion of each document
- fit a topic model to the remainder
- ask how probable the held out parts are under the model

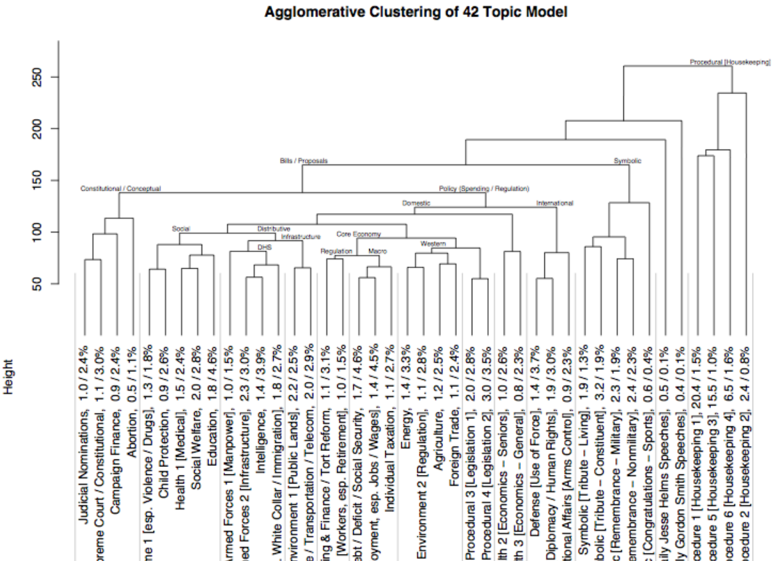
The stm package calls this 'heldout likelihood by document completion'

- Returns the average log probability of the heldout documents' words

Choice of K


Diagnostic Values by Number of Topics
Held-Out Likelihood
Held-Out Likelihood
-7.86
-7.78
10
15
20
25
30
Number of Topics (K)
Residuals
Residuals
6.5
8.0
9.5
10
15
20
25
30
Number of Topics (K)
Semantic Coherence
Semantic Coherence
-75
-65
10
15
20
25
30
Number of Topics (K)
Lower Bound
Lower Bound
-2350000
10
15
20
25
30
Number of Topics (K)

Construct validity



Choice of k

> *The results presented in this paper ... assume there are 43 topics present in the data. I varied the number of assumed topics from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issues being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground.*
>
> *(Grimmer, 2010)*

Choice of K

> *The results presented in this paper … assume there are 43 topics present in the data. I varied the number of assumed topics from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issues being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground.*
>
> *(Grimmer, 2010)*

We can be realists or anti-realists about topics

- → Anti-realism: topics are ‘lenses’
- → Realism: topics are real discourse units, e.g. themes, categories, etc.

Choice of k

> *The results presented in this paper … assume there are 43 topics present in the data. I varied the number of assumed topics from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issues being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground.*
>
> *(Grimmer, 2010)*

We can be realists or anti-realists about topics

→ Anti-realism: topics are 'lenses'

→ Realism: topics are real discourse units, e.g. themes, categories, etc.

We can *try* to be realists about the conditional independence assumption

→ Once we know the topic indicator, remaining word variation is just random → unpredictable

That's seldom true for mundane linguistic reasons

Topic coherence, exclusivity, and frex

Semantic coherence

→ Regularly co-occurring words should be in topics together

Exclusivity

→ High precision words make for *well-separated* topics

$$\frac{\beta_w^{(j)}}{\sum_{k \neq j} \beta_w^{(k)}}$$

frex

→ A weighted average of exclusivity and simple frequency, favouring exclusivity

These are heuristic measures, so don't take them too seriously.

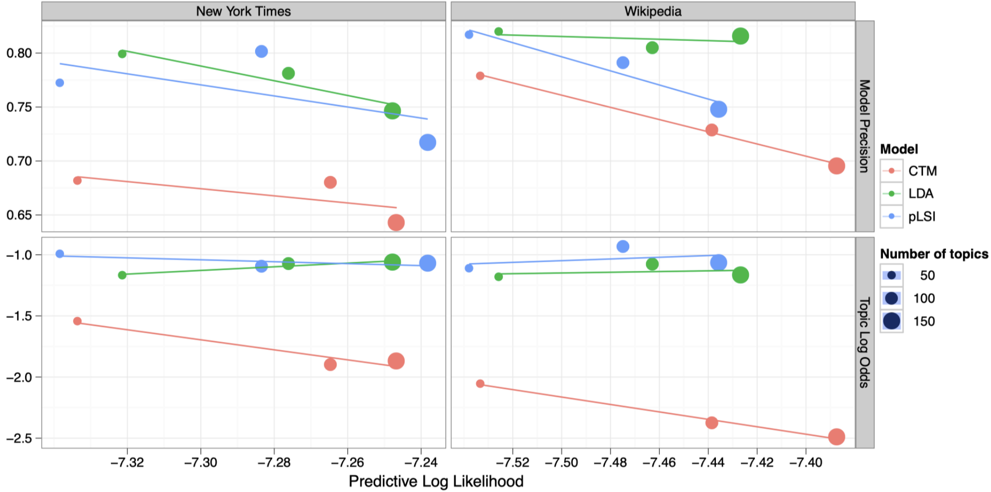
Humans in the loop

Experiments by Chang et al. (2009) using

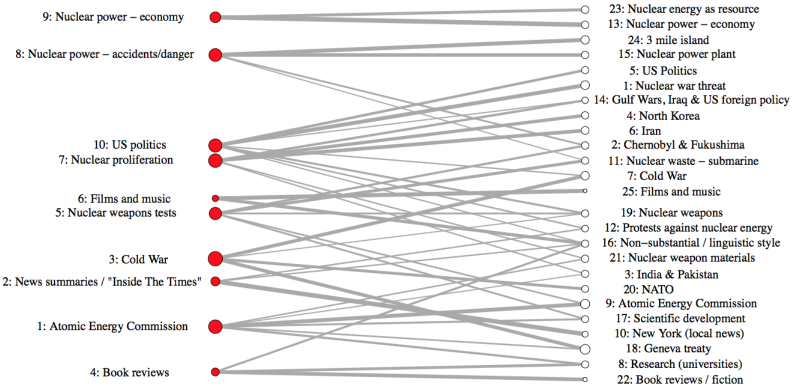
→ word intrusion: “which of the following words does not belong?”
→ topic intrusion: “which of these topics do not make sense for this passage?”

Unfortunately, these are measures that covary *negatively* with the previous statistical measures

Chang et al. 2009



Gamson and Modigliani redux


9: Nuclear power – economy
8: Nuclear power – accidents/danger
10: US politics
7: Nuclear proliferation
6: Films and music
5: Nuclear weapons tests
3: Cold War
2: News summaries / "Inside The Times"
1: Atomic Energy Commission
4: Book reviews
23: Nuclear energy as resource
13: Nuclear power – economy
24: 3 mile island
15: Nuclear power plant
5: US Politics
1: Nuclear war threat
14: Gulf Wars, Iraq & US foreign policy
4: North Korea
6: Iran
2: Chernobyl & Fukushima
11: Nuclear waste – submarine
7: Cold War
25: Films and music
19: Nuclear weapons
12: Protests against nuclear energy
16: Non–substantial / linguistic style
21: Nuclear weapon materials
3: India & Pakistan
20: NATO
9: Atomic Energy Commission
17: Scientific development
10: New York (local news)
18: Geneva treaty
8: Research (universities)
22: Book reviews / fiction

from van Atteveldt et al. (MS)

Case study

The real timeline of Baerg and Lowe (2020)

→ Paper and basic results sent to ACL conference workshop (wins prize)
→ Conference paper expanded to political science length journal article and sent out to review
→ (repeat several times)
→ Reviewers do not believe topics measure what we say they measure
→ Embark on a manual validation exercise (random kwics etc.)
→ …
→ Publish

Disagreements were all about measurement validity.

Nobody ever asked for the heldout log likelihood, coherence, exclusivity, or choice of K.

Evaluation

Recommendations if you feel the need to topic model:

- Don't take the statistical evaluation measures very seriously
- Work at getting interpretable substance right
- Take what you need from the model (we threw away some topics and aggregated some others)
- You can always fit on a random sample and apply to the remainder of your documents
- Your idea of a topic is hard to communicate, even to your friends, so don't expect any machinery to intuit it
- Make sure it all replicates!

References

Baerg, N. & Lowe, W. (2020). 'A textual taylor rule: Estimating central bank preferences combining topic and scaling methods'. *Political Science Research and Methods*, *8*(1), 106–122.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. & Blei, D. M. (2009). 'Reading tea leaves: How humans interpret topic models'. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 288–296.

Grimmer, J. (2010). 'A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases'. *Political Analysis*, *18*(1), 1–35.