Topic Modelling

Max Callaghan

November 22, 2017

Explanation

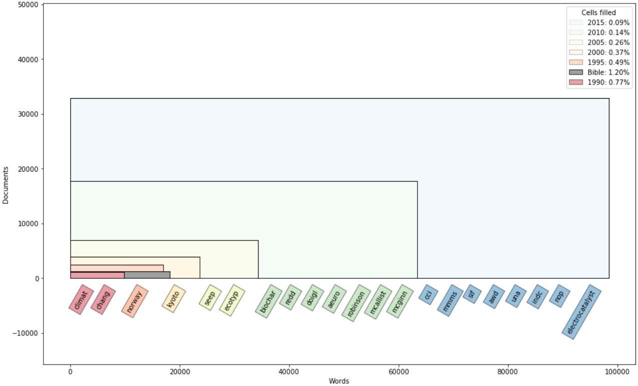
With online archives of documents expanding, we have more information than we can humanly process, but we also have new opportunities for using this information in different ways, or doing "distant reading" (Moretti, 2013).

Topic modelling describes "a suite of algorithms that aim to discover and annotate large archives of documents with thematic information" (Blei et al., 2012)

This annotated corpus of documents allows us "organise and summarise" electronic text collections, so that we can ask questions about how themes are connected and have changed over time (Blei et al., 2012)

Words, Words, Words

- For topic modelling, a collection of documents is a matrix of word occurrences in documents
- (The bag of words assumption)



Non-negative Matrix Factorisation (NMF)

(Lee and Seung, 1999)

$$V_{i\mu}$$

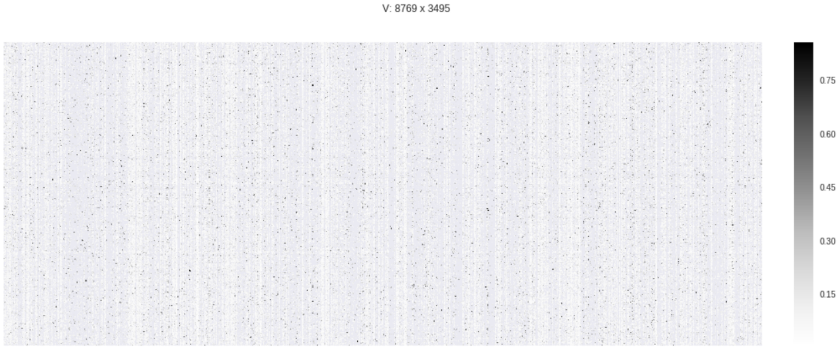


Figure: A Document term matrix of 3495 documents on climate change from the year 2000

Non-negative Matrix Factorisation (NMF)

(Lee and Seung, 1999)

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$$

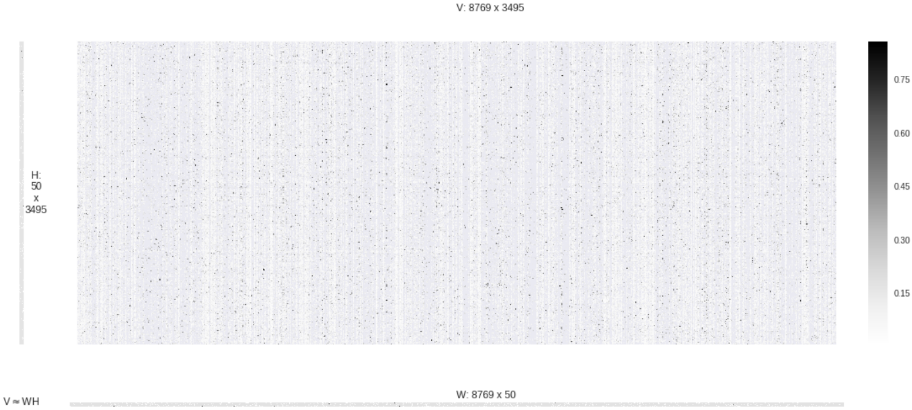
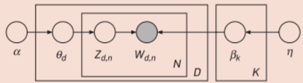


Figure: A topic model of 3495 documents on climate change from the year 2000

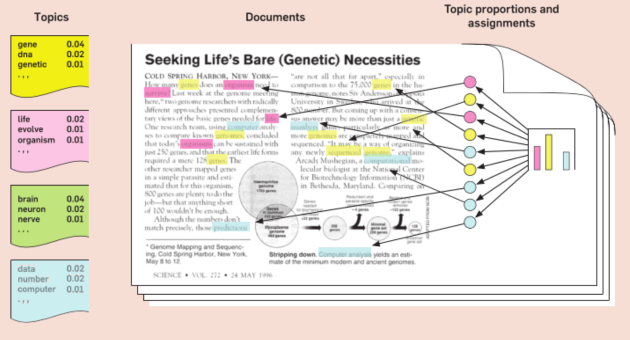
Latent Dirichlet Allocation (LDA)

Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The *N* plate denotes the collection words within documents; the *D* plate denotes the collection of documents within the collection.



- LDA similarly describes documents as distributions of topics, which are distributions of words
- The assumptions about probability are slightly different, but the intuition is the same

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.



APSIS applications - NETs

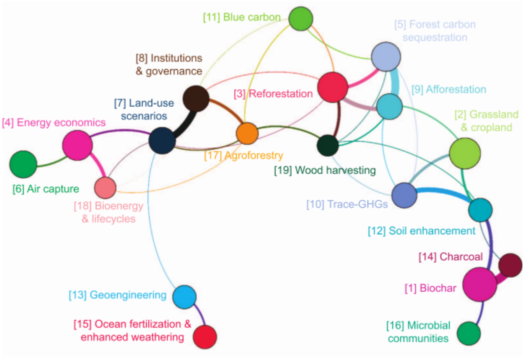


Figure 3. Correlation structure of negative emissions topics. Each node represents a topic, scaled by the marginal topic distribution (table 1); each line represents a positive correlation between two topics. The largest correlation is of 0.24 between biochar and charcoal. Nodes that are proximate to one another are more highly correlated than those which are distant. The visualization is generated from inter-topic correlations using the force-directed algorithm ForceAtlas2 in Gephi (Jacomy *et al* 2014).

Figure: Minx, J. C., Lamb, W. F., Callaghan, M. W.,

Bornmann, L., and Fuss, S. (2017b). Fast growing research on negative emissions.
Environmental Research Letters, 12(3):035007

- What topics exist on negative emissions, how are they connected?

APSIS applications - Cities

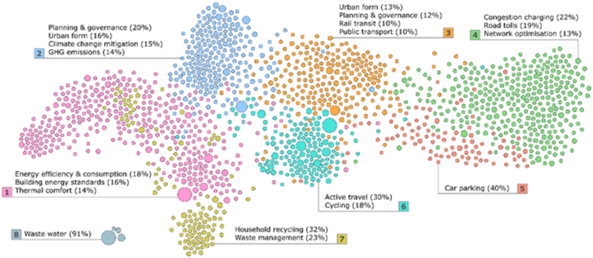


Figure: Lamb, W. F., Callaghan, M. W., Creutzig, F., Khosla, R., and Minx, J. C. (2017). The literature landscape on 1.5[deg]C Climate Change and Cities. *Current Opinion in Environmental Sustainability* (Submitted)

- Topics can be used to characterize bibliographic networks

APSIS applications - Cities

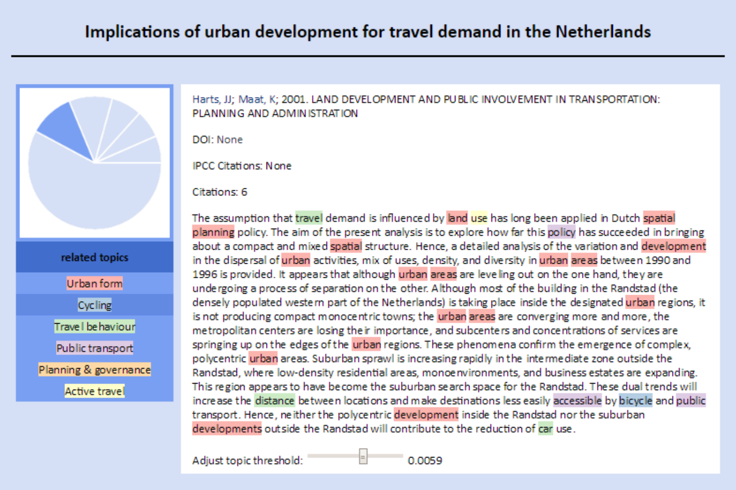


Figure: Annotated document using data from Lamb et al. (2017)

APSIS applications - Cities

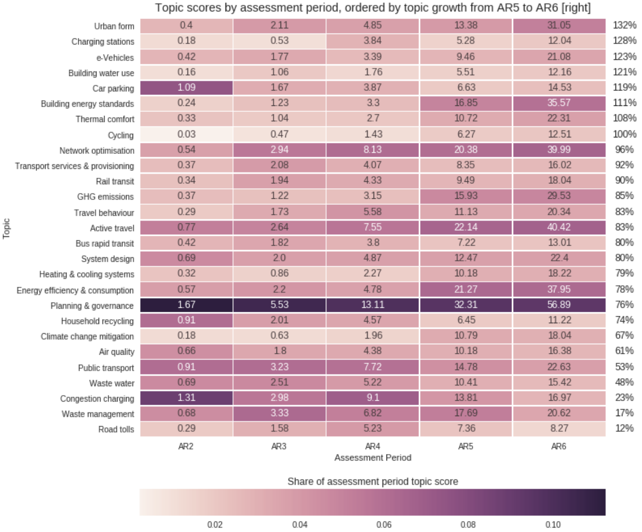


Figure: Topic Growth over time (Lamb et al., 2017)

Extensions

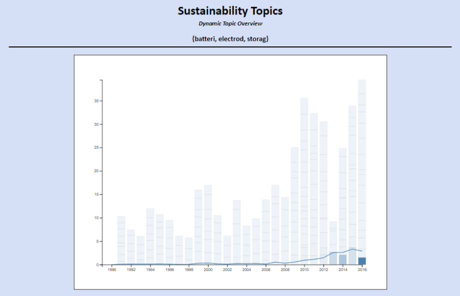
Incorporating additional information

- Place mentions / IPCC references
- Structural topic models (Roberts et al., 2014)

Better modelling topic dynamics

- Emergence / Evolution of topics
- Dynamic landscape of sustainability (Minx et al., 2017a)

Dynamic Topic Models



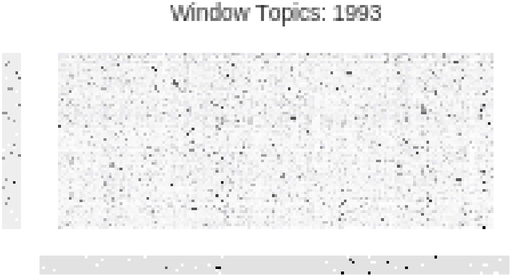
In (Minx et al., 2017a) we apply Greene and Cross (2016)' Dynamic Topic Model

View	0.95	2013	{storag, batteri, electrod}	storag	batteri	electrod	capac	electrochem	charg	high	cycl	cathod	devic
View	1.72	2014	{batteri, storag, electrod}	batteri	storag	electrod	lithium	vehicl	charg	cathod	capac	electrochem	recharg
View	0.67	2015	{batteri, electrod, storag}	batteri	electrod	storag	electrochem	cycl	capac	high	devic	cathod	densiti
View	4.39	2016	{batteri, electrod, storag}	batteri	electrod	storag	electrochem	capac	high	electrolyt	densiti	cathod	anod

Dynamic Topic Models

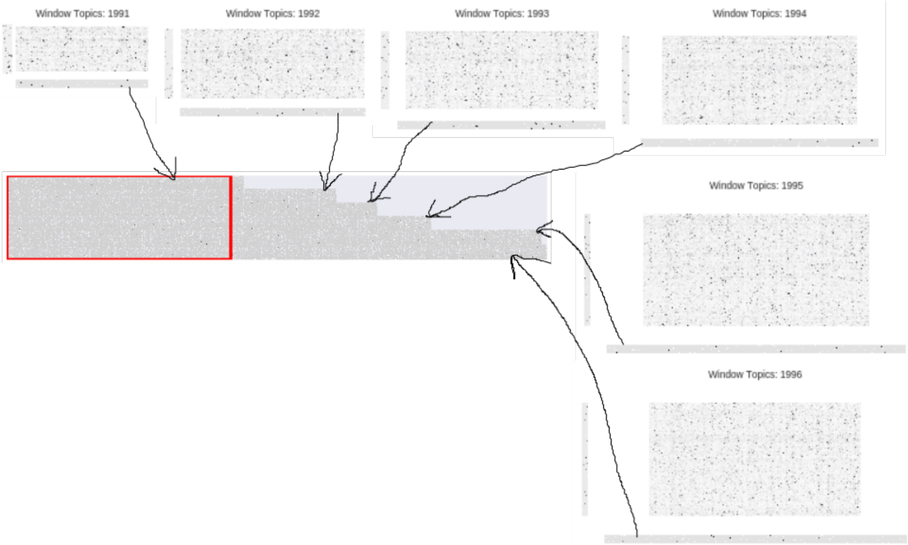
Dynamic Topic Models

Dynamic Topic Models

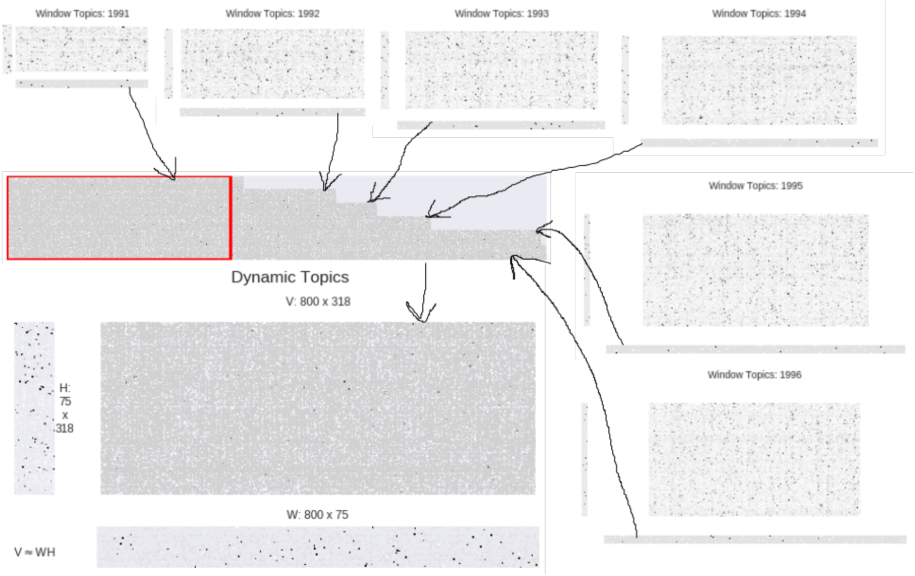


Dynamic Topic Models

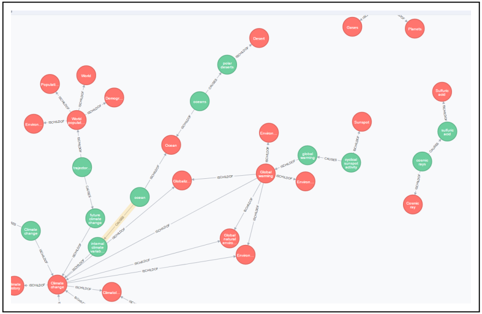
Dynamic Topic Models



Dynamic Topic Models



Other Text Analysis



- Causaly collect causal statements from literature
- They aim to quantify and aggregate the strength of claims

Applications?

- Do we get more consolidated knowledge about causal relationships over time (in some WGs over others)?
- What can we learn about co-benefits and side-effects of different negative emission technologies?

Bibliography

Blei, D., Carin, L., and Dunson, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Greene, D. and Cross, J. P. (2016). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. pages 1–47.

Lamb, W. F., Callaghan, M. W., Creutzig, F., Khosla, R., and Minx, J. C. (2017). The literature landscape on 1.5[deg]C Climate Change and Cities. *Current Opinion in Environmental Sustainability*.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91.

Minx, J. C., Callaghan, M. W., Creutzig, F., Hilaire, J., and Lamb, W. F. (2017a). The dynamic landscape of sustainability science. *Nature Sustainability*.

Minx, J. C., Lamb, W. F., Callaghan, M. W., Bornmann, L., and Fuss, S. (2017b). Fast growing research on negative emissions. *Environmental Research Letters*, 12(3):035007.

Moretti, F. (2013). *Distant Reading*. Verso, London.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.