A Topography of Climate Change Research

Max Callaghan

May 3, 2017

To contribute to evidence-based policies on tackling climate change, the IPCC aims to comprehensively assess the relevant scientific literature (IPCC, 2013). With the size of this literature currently at least [x] times larger than at the time of the IPCC's first assessment report (Houghton et al., 1990), this task has become impossible without the aid of machine-reading. We collect over 300,000 abstracts from Web of Science (WoS) and Scopus, and develop a topic model in order to give an overview of this unmanageably large corpus. We find that new topics, such as [biochar] can be identified as they emerge, an application with a high potential for increasing the relevance and comprehensiveness of further IPCC reports.

The size of the scientific literature on climate change has expanded rapidly over the lifetime of the IPCC. While the first assessment report had around 7,000 articles to assess, 5,000 new articles are now published every month, bringing the total size of the literature to well over half a million papers, (Figure 1). The increase in volume, velocity, and variety of content to be assessed has turned the task of the IPCC into a 'Big Literature' challenge. To ask questions about the literature *at scale*, we now need to employ computational techniques involving natural language processing.

Topic models are one such technique. A topic model learns the latent topics that structure a large corpus of documents, by leveraging the systematic co-occurrence of words across documents. Topics are distributions of words, and the topic mixture of each document explains the words observed in that document. This means that topic models can aid the understanding of large corpuses, and of the place of individual documents within them, by showing a document or corpus as a combination of 100 or so intelligible topics, rather than combinations of thousands of words.

The topic model presented here is a rough map of climate change research since 1985. It shows a broad outline of the topics that make up this research and how they relate to each other, and demonstrates how this has changed over time. This relief map does not replace the maps drawn by assessment-makers-as-cartographers for policymakers described in Edenhofer and Kowarsch (2015). Rather, it supplements it by reducing the exploration time necessary and acting as a reference, to compare how the assessment reflects the literature. [Rather, it provides a vital anchor.]

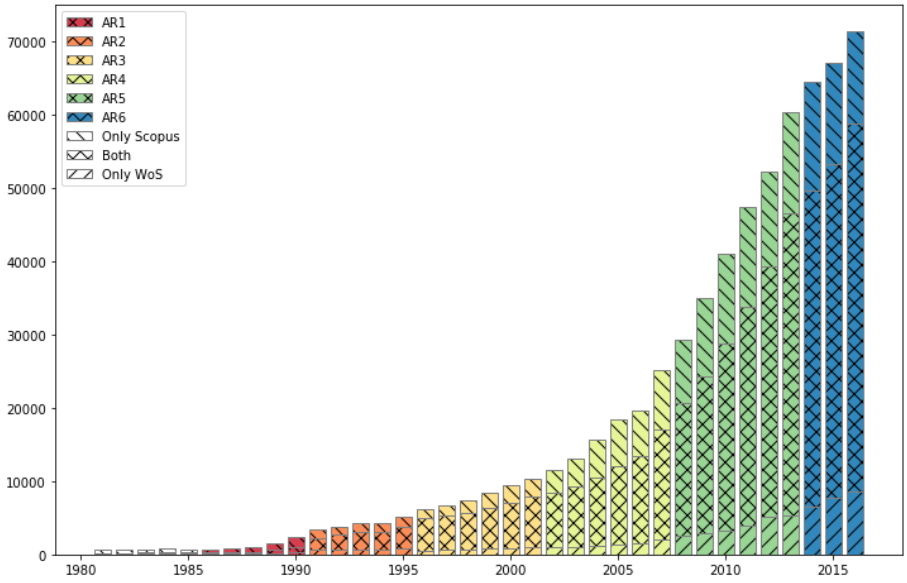

Figure 1: Growth in relevant literature in WoS and Scopus

While topic modelling has been employed to answer specific questions about small aspects of climate literature, e.g. (e.g. Minx et al., 2017; Grubert and Siders, 2016), this is the first application of topic models to gain an overview of the entire field.

1 Introduction

- Literature exploding, IPCC not keeping up (Minx et al., 2016)
- No systematic way of selecting references, no comprehensive assessment
- First step towards this is a map provided by topic models.
- A map shows the places which the assessment makers need to navigate, leaving them to decide course. Computer assisted not computer decided.
- What is the topic structure of climate change research? How has it changed over time?

2 Methodology

- Model selection: NMF (Lee and Seung, 1999)

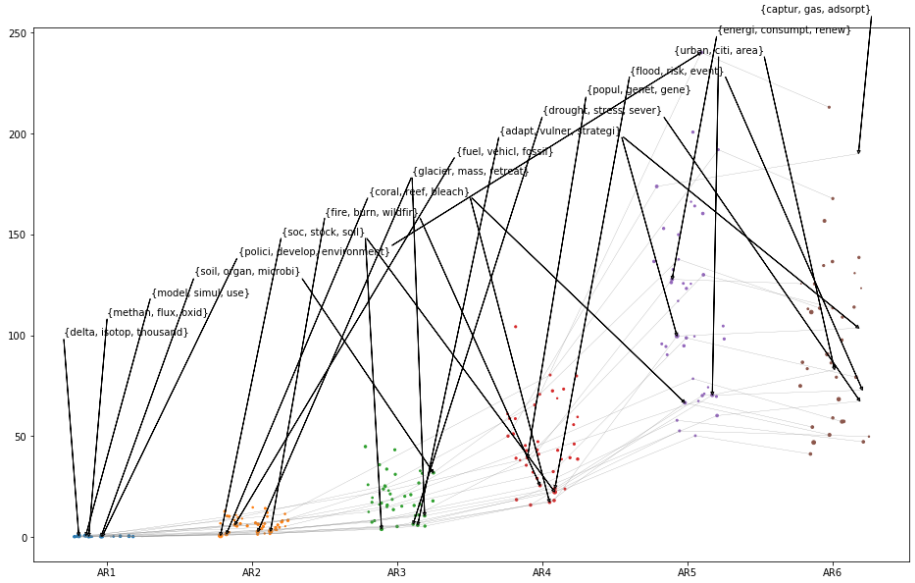

Figure 2: Topic growth over time. The 3 topics in each assessment period that grew by the largest amount are labelled

AR1

topic__title	pchange
{delta, isotop, thousand}	40144%
{methan, flux, oxid}	27808%
{model, simul, use}	23694%
{soil, organ, microbi}	23110%
{polici, develop, environment}	21927%

AR2

topic__title	pchange
{soc, stock, soil}	12638%
{fire, burn, wildfir}	10893%
{coral, reef, bleach}	7619%
{glacier, mass, retreat}	7543%
{fuel, vehicl, fossil}	4897%

- How does it work? Advantages: Simple, scalable: better results than with other solutions, if only because it was possible to iterate with large document collections. LDA can be better but relies on tuning hyperparameters, hard to do with such big corpus
- Topic model browser Chaney and Blei (2012)
- Human Validation:
- Compare topic space to keyword space. Reduces dimensionality, overcomes non-standardisation
- Which topics get the most citations?

3 Data

- Queries: use Grieneisen and Zhang (2011), or take the best bits of Grieneisen and Zhang (2011) and Haunschild et al. (2016)?
- Sources: WoS, Scopus or both?
- Preprocessing: Remove punctuation, numbers, common, uncommon words, stemming

4 Results

- The biggest topics are x and y
- These topics have grown, these have decreased
- X keywords fit into Y topics like so...
- Finding similar documents works across topics, rather than just by keywords, so better.
- Similar documents more or less likely to be across disciplinary boundaries ??
- Model selection validated by these measurements.

5 Conclusion

- A very simple topic model provides an overview of the whole landscape.
- This allows researchers / assessment makers to identify areas that have grown recently
- Topic models aid document discovery, have the potential to contribute to more comprehensive assessments.

Figure 3: Topic structure of climate change literature [network plot]

Figure 4: Focus on [biochar?] showing document with highlighted words

- Next steps for research: using topic models to assess the assessment process: find gaps etc.

List of Figures

References

Chaney, A. and Blei, D. (2012). Visualizing Topic Models. *Icwsm*, pages 419–422.

Edenhofer, O. and Kowarsch, M. (2015). Cartography of pathways: A new model for environmental policy assessments. *Environmental Science and Policy*, 51:56–64.

Grieneisen, M. and Zhang, M. (2011). The Current Status of Climate Change Research. *Nature Climate Change*, 1:72–73.

Grubert, E. and Siders, A. (2016). Benefits and applications of interdisciplinary digital tools for environmental meta-reviews and analyses. *Environmental Research Letters*, 11(9):093001.

Haunschild, R., Bornmann, L., and Marx, W. (2016). Climate Change Research in View of Bibliometrics. *PLoS ONE*, 11(7):1–19.

Houghton, J. T., Jenkins, G. J., and Ephraums, J. J. (1990). *Climate Change The IPCC Scientific Assessment*, volume 1.

Figure 5: Model validation graph, showing error for different topic numbers, feature numbers

Figure 6: Some relation of topics to other features of dataset: e.g. most interdisciplinary journals and least, or so...

IPCC (2013). Principles governing IPCC work.

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., Lamb, W. F., Garard, J., and Edenhoffer, O. (2016). Learning about climate change solutions.

Minx, J. C., Lamb, W. F., Callaghan, M. W., Bornmann, L., and Fuss, S. (2017). Fast growing research on negative emissions.