A Topography of Climate Change Research

September 20, 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 almost two orders of magnitude larger than at the time of the IPCC's first assessment report, this task has become impossible without the aid of machine-reading. We collect over 400,000 abstracts from Web of Science (WoS) and Scopus, and develop a topic model in order to give an overview of this unmanageably large corpus. This overview shows us the distribution and development of topics across the literature, and allows us to identify topics with greater and lesser representation in 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 5,500 articles to assess, nearly 5,000 new articles are now published every month, bringing the total size of the literature to close to 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 (Minx, 2017). To ask questions about the literature *at scale*, we now need to apply computational techniques to the analysis of large document collections.

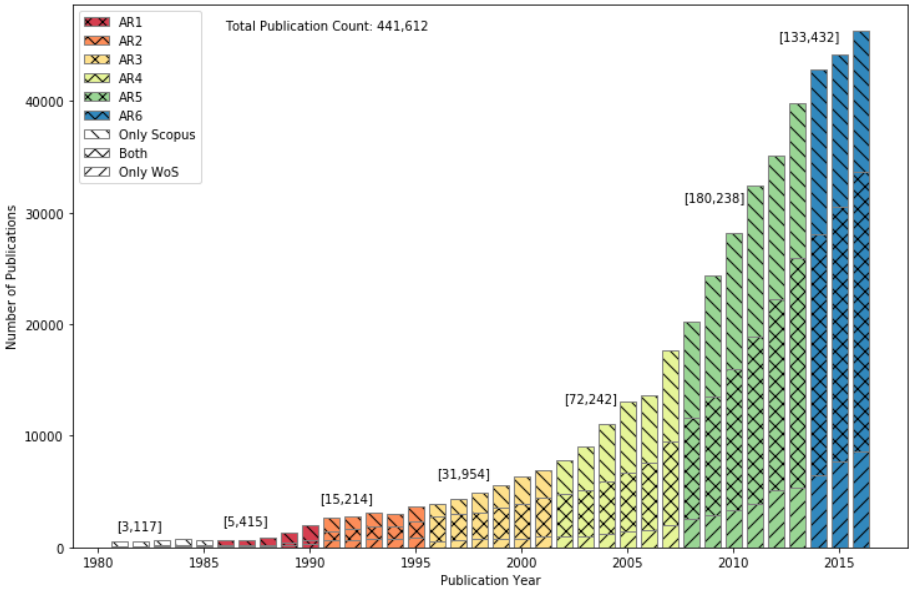

Figure 1: Growth in relevant literature in WoS and Scopus

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.

Assessment-makers like the IPCC have been described as cartographers for policymakers (Edenhofer and Kowarsch, 2015). As such their purpose is, summarising available scientific knowledge, to describe the problem and solution space of a policy issue. 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. Such a map sheds light on the terrain of knowledge about a policy issue, making an

overview of an unmanageably large and diverse landscape possible. This overview allows both for the production of policy pathways that are well informed by science, and, as demonstrated in this paper, the assessment of the comprehensiveness with which these pathways reflect the landscape.

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.

Results

From here on in, the paper is more of an outline, with a mixture of bullet points about results that exist or I would still like to generate

Topic Model

- Figure 2 shows the structure of the topic model with 100 topics, with each node coloured according to the IPCC working group in which the highest proportion of the topic's documents are cited.
- Topics that systematically co-occur in documents are linked, and the resulting network is displayed with links weighted according to topic correlation score. The strength of links is greater where two topics are categorised as being in the same working group. This relationship is statistically significant (see SI)
- Some big/interesting topics are x and y
- These topics have grown at particularly interesting times (Figure 3)
- These topics are better covered in IPCC, these are less well covered 4

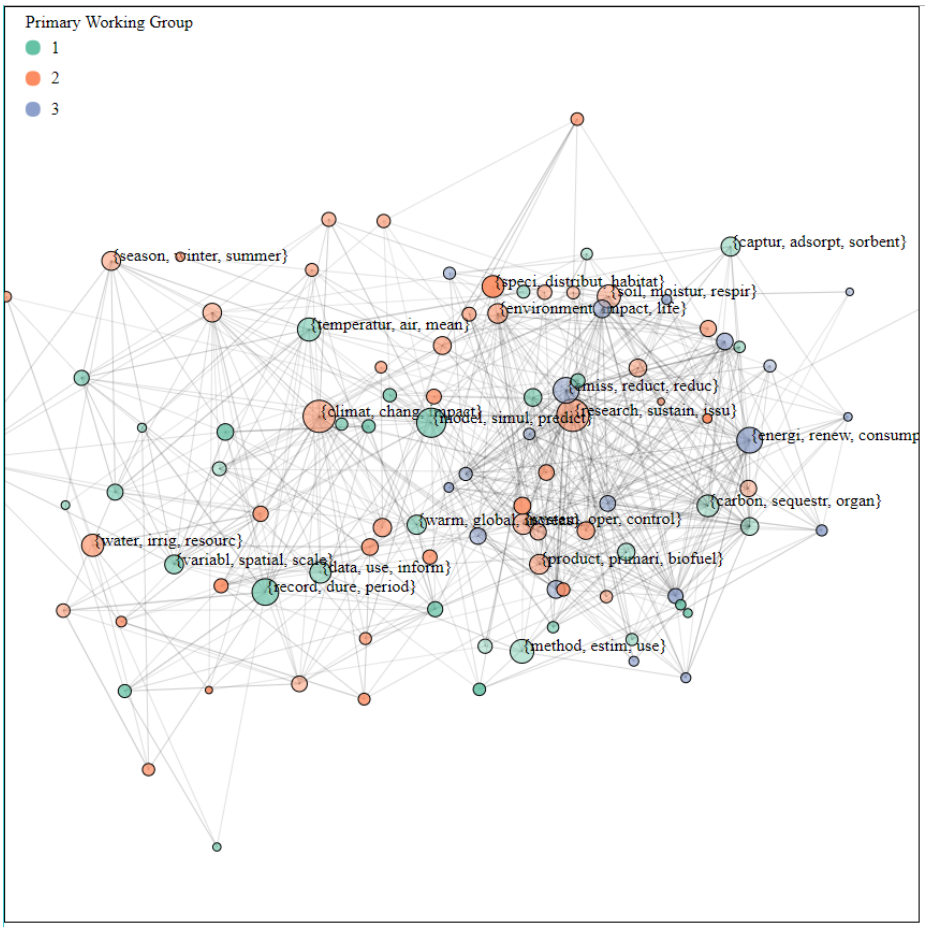

Figure 2: Topic structure of climate change literature

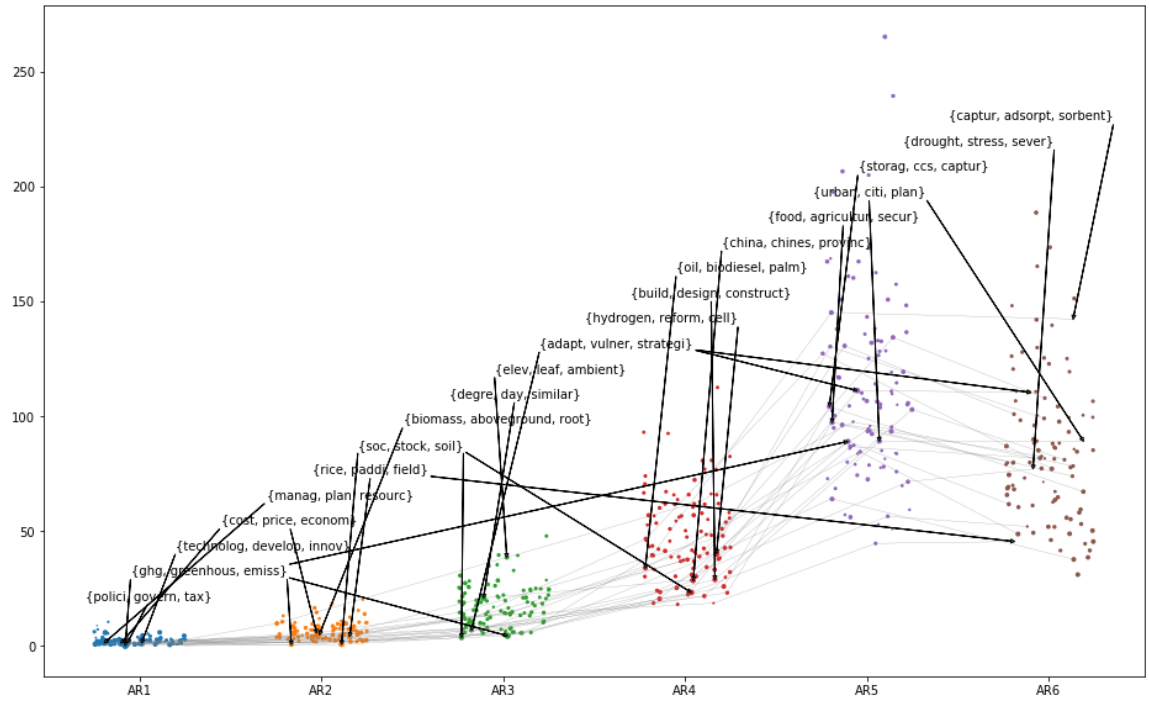

Figure 3: Topic growth over time. The 3 topics in each assessment period that grew by the largest amount are labelled [Example figure]

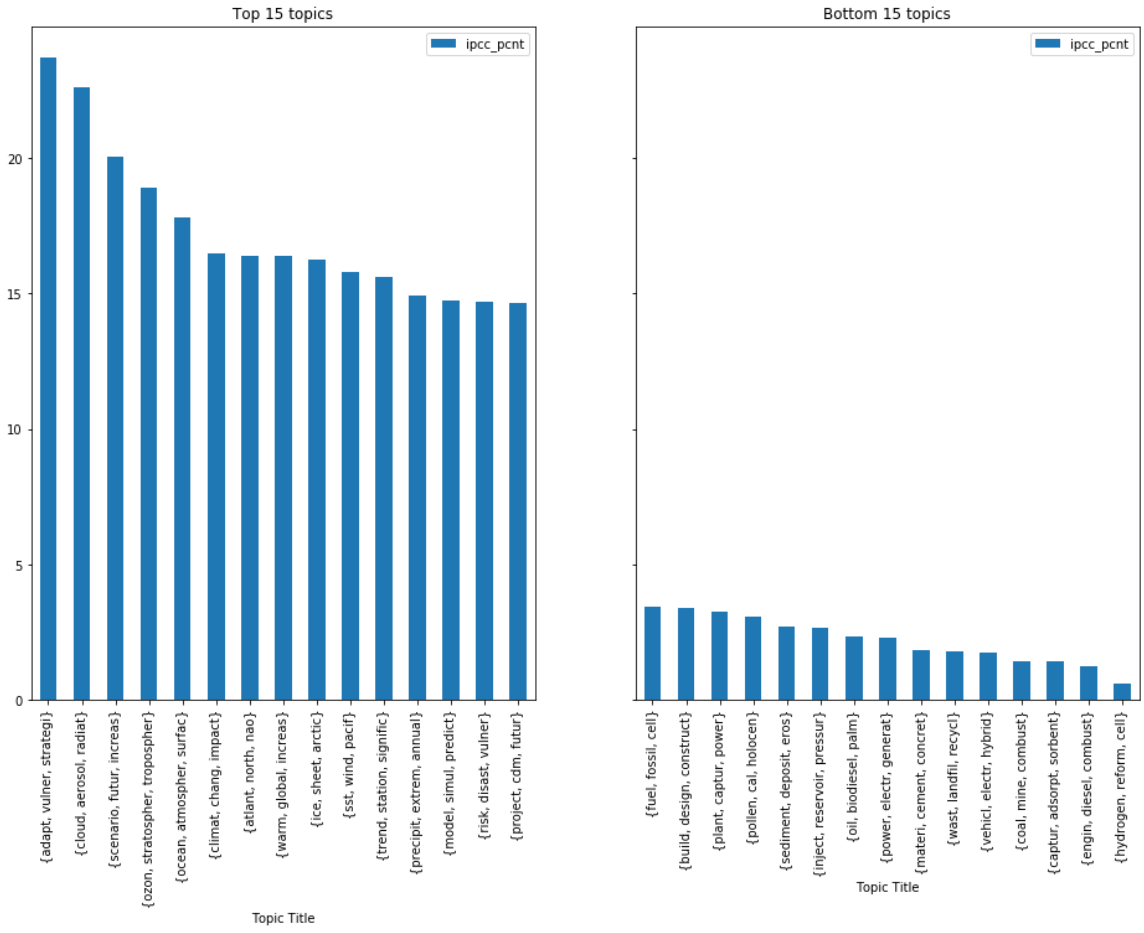

Figure 4: IPCC references by topic. The bars show the percentage of each topic that has been matched to an IPCC reference

AR1		AR2		AR3		AR4		AR5		AR6	
title	pchange	title	pchange	title	pchange	title	pchange	title	pchange	title	pchange
polici, govern, tax	255%	rice, paddi, field	513%	ghg, green-hous, emiss	639%	hydrogen, reform, cell	618%	adapt, vulner, strategi	397%	drought, stress, sever	8%
ghg, green-hous, emiss	250%	soc, stock, soil	492%	degre, day, similar	566%	soc, stock, soil	534%	ghg, green-hous, emiss	334%	rice, paddi, field	2%
technolog, develop, innov	240%	cost, price, econom	398%	soc, stock, soil	432%	build, design, con-struct	417%	food, agri-cultur, secur	320%	urban, citi, plan	-0%
cost, price, econom	220%	biomass, above-ground, root	387%	elev, leaf, ambient	294%	oil, biodiesel, palm	402%	urban, citi, plan	311%	adapt, vulner, strategi	-1%
manag, plan, resourc	191%	ghg, green-hous, emiss	354%	adapt, vulner, strategi	278%	china, chines, provinc	384%	storag, ccs, captur	303%	captur, adsorpt, sorbent	-2%

Table 1: The top 5 topics in each assessment period by percentage growth since the last assessment period

AR1		AR2		AR3		AR4		AR5		AR6	
title	pchange	title	pchange	title	pchange	title	pchange	title	pchange	title	pchange
tree, stand, ring	-11%	sediment, deposit, eros	91%	ice, sheet, arctic	101%	concentr, atmo-spher, air	87%	cloud, aerosol, radiat	47%	technolog, develop, innov	-39%
soc, stock, soil	-15%	drought, stress, sever	84%	food, agri-cultur, secur	100%	methan, oxid, wetland	86%	ice, sheet, arctic	45%	coal, mine, combust	-40%
coal, mine, combust	-19%	sea, level, coastal	84%	coal, mine, combust	75%	elev, leaf, ambient	63%	sst, wind, pacif	42%	fuel, fos-sil, cell	-40%
dioxid, carbon, atmo-spher	-23%	lake, level, diatom	83%	warm, global, increas	73%	rice, paddi, field	49%	atlant, north, nao	35%	hydrogen, reform, cell	-41%
solar, radiat, irradi	-31%	glacier, mass, retreat	61%	dioxid, carbon, atmo-spher	67%	degre, day, similar	37%	ozon, strato-spher, tropo-spher	25%	oil, biodiesel, palm	-42%

Table 2: The bottom 5 topics in each assessment period by percentage growth since the last assessment period

Figure 5: Focus on [biochar?] showing document with highlighted words

Figure 6: Model validation graph, showing error for different topic numbers, feature numbers

- Some interesting topic correlations are x and y; well fitting documents to both include x and y
-

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.
- AR5 seemed to have less comprehensive coverage of x topics. This was a particular issue in WG y.
- This may not be an issue, there could be good reasons for this, but these should be made transparent.
- For the next assessment report, x topics may require particular attention.
- The emerging topic on CCS resonates with a growing recognition of the importance of negative emissions and the lack of understanding about how they could fill their role. This will be of particular importance for the IPCC special report on 1.5 degrees.

Figure 7: Some relation of topics to other features of dataset: e.g. most interdisciplinary journals and least, or so...

Methodology

- Topic modelling in general: reducing large matrix of documents to words to two smaller matrices of topics x words and topics x documents.
- Model selection: NMF (Lee and Seung, 1999)
- How does it work? Advantages: Simple, scalable: better results than with other solutions
- Topic model browser Chaney and Blei (2012)
- Merging with IPCC citation dataset - caveats...
- Network explanation
- Regression of network score on dummy variable for same

1 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

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.

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. (2017). Big Literature in Climate Change.

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