A Topography of Climate Change Research - Methods

Max Callaghan[1,2]

[1]Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany
[2]School of Earth and Environment, University of Leeds, Leeds LS2 9JT, United Kingdom

Draft current January 3, 2019

Data

This study reproduces the query developed by [1], which is carried out on the Web of Science core collection. Though not exhaustive, it gives a good coverage of the literature in major peer-reviewed journals. Each document is assigned to an assessment period according to the timeline shown in table 1.

We use the references scraped from IPCC assessment reports from [2], and attempt to match these with the results from the web of science. Table [x] shows the percentage of IPCC citations matched in each working group for each assessment report.

name	years
AR1	1988-1989
AR2	1990-1994
AR3	1996-2000
AR4	2001-2006
AR5	2007-2013
AR6	2014-

Table 1: Assessment period time windows

Pre-processing

Data quality in earlier Web of Science results is poorer, and some documents have missing abstracts. In the quantification of the size of the literature and its vocabulary in table [], titles are substituted for abstracts where they are not available. The words of the documents are lemmatized/stemmed, replacing different forms of the same word (i.e. word/words) with a single instance. Commonly occuring words, or "stopwords" are removed, as are all words shorter than 3 characters, and all words containing only punctuation or numbers.

For each period, the documents are transformed into a document-term matrix, each row represents a document, and each column represents a unique word. Each cell contains the number of that column's terms in that document. Only terms which occur more than once are considered.

For the calculation of the topic model, documents with missing abstracts are ignored, and the document term matrix is transformed into a document frequency-inverse document frequency (tf-idf)

matrix, where scores are scaled according to the frequency of their occurence in the corpus. This gives more weight to terms which appear in few documents, and less weight to those which appear in many.

$$tf(t,d) = f_{t,d}, \quad idf(t,D) = \log \frac{N}{|\{d \in D : t \in d|\}} \tag{1}$$

Dynamic Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) is an approach to topic modelling which factorises the term-frequency-inverse document frequency matrix V into the matrices W, the topic-term matrix, and H the document-topic matrix, whose product approximates V:

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \tag{2}$$

As demonstrated in Figure 1, each topic is represented as a set of word scores, and each document a set of topic scores. The combination of the two give the word scores in the document. For clarity in the figure, these are shown as simple counts, but in the model these are scaled according to each term's frequency within the corpus.

In Dynamic Non-negative Matrix Factorisation, proposed by Greene [3], a separate topic model is run for each period. These are then joined through another topic model, which takes the topic-term matrices of the all periods as V, and produces dynamic topics, which describe the window topics according to the words which occur in them. Similar topics across and within time periods are thereby grouped together.

While Greene uses an automatic approach to deciding on topic numbers within time periods, we found the number of topics derived from topic coherence scores (as used by Greene) to be noisy, and instead opt for identifying (subjectively) an optimal number of topics for each window. We do this by comparing topic lists with increasing numbers of topics, where similar topics are automatically placed next to each other. We similarly compare different numbers of Dynamic topics.

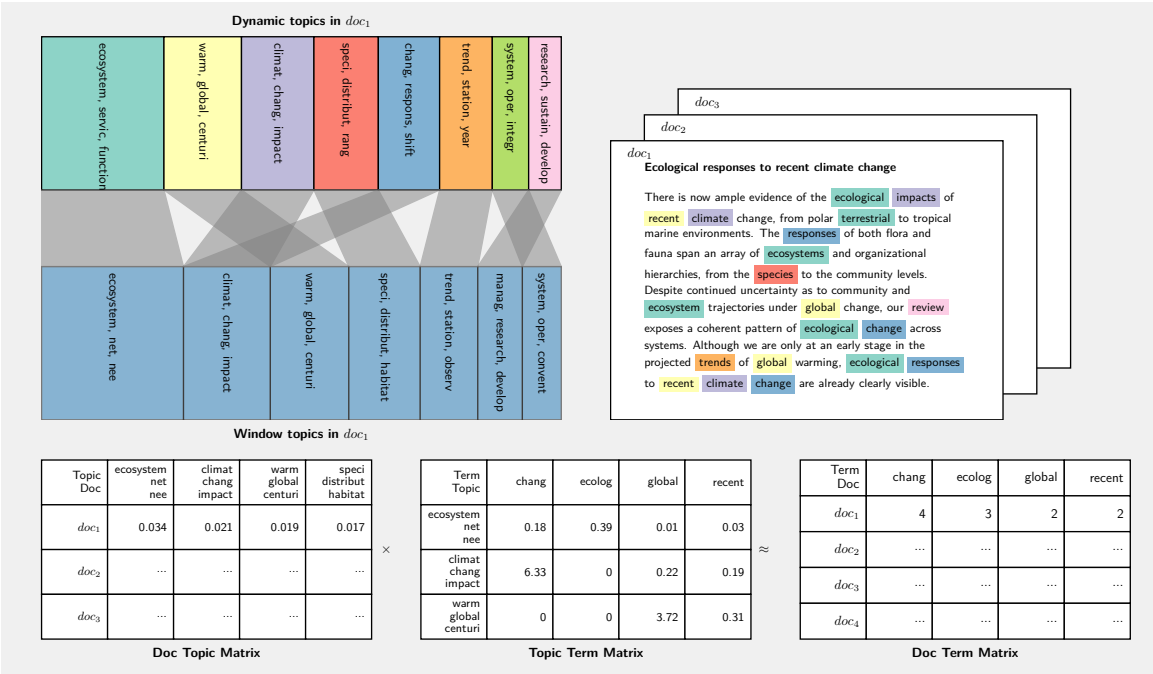
We settle on [x,x,x,x,x] topics for ARs 1 to 6, and [x] dynamic topics.

Topics are calculated using the scikitlearn library [4]

Topic Representation and Newness

To calculate topic representation in IPCC reports we divide each topic's share in the subsample of documents cited by IPCC reports by its share in the whole corpus.

We calculate a dynamic topic's total score as the sum of document-dynamic topic scores (which in turn are made of the product of all document-window topic and window topic-dynamic topic scores). A dynamic topic's window score is the sum of document-dynamic scores considering only documents in the given time window. To represent a dynamic topic's newness, we multiply each assessment period number by the share of it's total score occurring in that window, and take the mean of these scores. A topic in which 100% of documents which make it up occurred in assessment period 1 (6) would thereby receive a score of 1 (6), while a topic evenly distributed across all assessment periods would receive a score of 3.5.



Topic Doc	ecosystem net nee	climat chang impact	warm global centuri	speci distribut habitat
doc_1	0.034	0.021	0.019	0.017
doc_2	...	...	...	...
doc_3	...	...	...	...

×

Term Topic	chang	ecolog	global	recent
ecosystem net nee	0.18	0.39	0.01	0.03
climat chang impact	6.33	0	0.22	0.19
warm global centuri	0	0	3.72	0.31

≈

Term Doc	chang	ecolog	global	recent
doc_1	4	3	2	2
doc_2	...	...	...	...
doc_3	...	...	...	...
doc_4	...	...	...	...

Figure 1: SI Topic make up of a single document

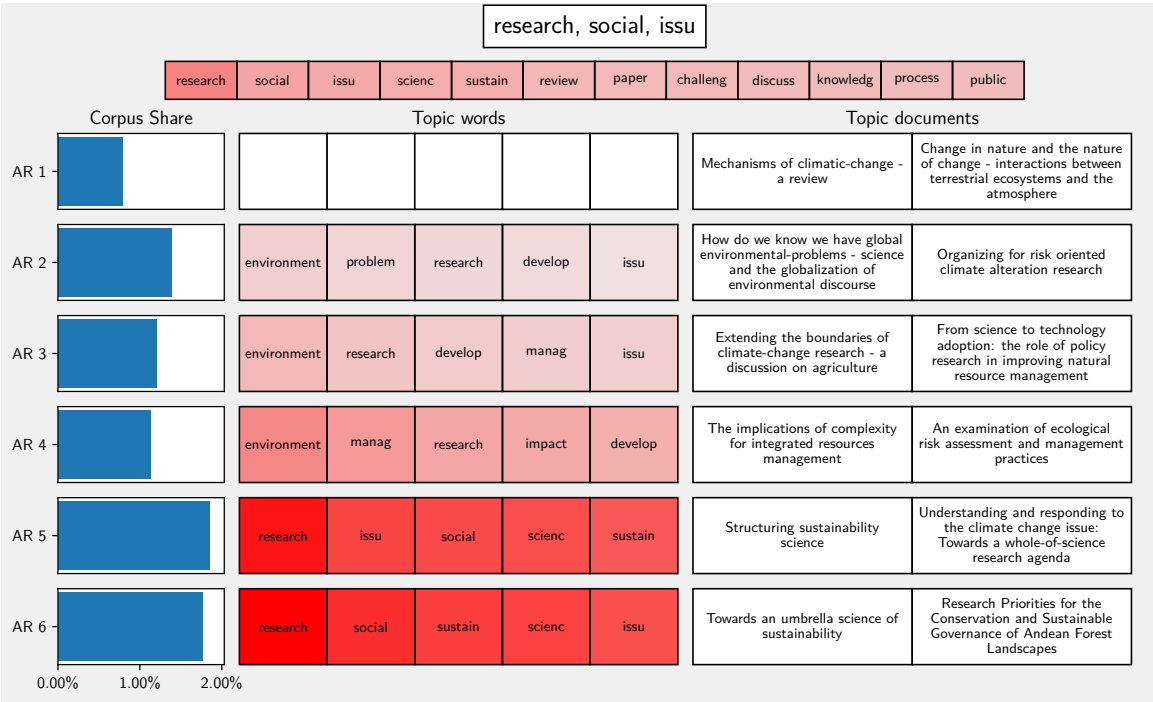


Figure 2: Word and document development of the "Research" dynamic topic

References

[1] Michael Grieneisen and Minghua Zhang. The Current Status of Climate Change Research. *Nature Climate Change*, 1:72–73, 2011.

[2] Jan C. Minx, Max Callaghan, William F. Lamb, Jennifer Garard, and Ottmar Edenhofer. Learning about climate change solutions in the IPCC and beyond. *Environmental Science & Policy*, 2017.

[3] Derek Greene and James P Cross. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. pages 1–47, 2016.

[4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Mattheiu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python Fabian. *Journal of Machine Learning Research*, 12:2825–2830, 2011.