

# A Topography of Climate Change Research - Methods

Max Callaghan<sup>1,2</sup>

<sup>1</sup>Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany

<sup>2</sup>School of Earth and Environment, University of Leeds, Leeds LS2 9JT, United Kingdom

Draft current May 13, 2019

## 1 Data

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

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

## 9 Pre-processing

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

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

19 For the calculation of the topic model, documents with missing abstracts are ignored, and the  
20 document term matrix is transformed into a document frequency-inverse document frequency (tf-idf)  
21 matrix, where scores are scaled according to the frequency of their occurrence in the corpus. This  
22 gives more weight to terms which appear in few documents, and less weight to those which appear  
23 in many.

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

## 24 Non-negative Matrix Factorisation

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

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

As demonstrated in Figure ??, 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.

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.

## 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, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python Fabian. *Journal of Machine Learning Research*, 12:2825–2830, 2011.