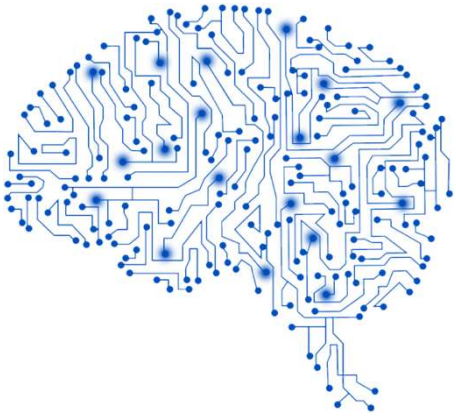
WZB

Text-mining international politics

Session 4
Basics of machine learning and topic models



1

WZB

Supervised machine learning – intuition

- **Basic idea of supervised classification**
  Algorithm 'learns' from (a few) human-coded documents before it automatically classifies (many) 'virgin' texts
- **Achieved along four (iterative) steps:**

1. **Construct a training and a test set from your documents**
   - Human coders apply a coding scheme to two subsets of docs
   - Size depends on doc length, unique language, number of categories etc. but usually a small fraction of the overall corpus is enough
2. **'Learn' classifier function from the training set**
   - Training documents used to find a (statistical) function that best predicts the human-coded categories along document-term frequencies
   - Various algorithms that come with different assumptions (Logistic regression, random forest, neuronal networks, …)
   - The *RTextTools* and *readMe* packages provide common implementations

2

WZB

Supervised machine learning – intuition

3. Validate the classifier in the test set
   - Use the classifier function from the training set to predict the categories of documents in the test set
   - Does you classifier live up to the '*gold standard*' of human coding?
   - If precision is insufficient, go back to step 1: Either your coding scheme has to be re-worked, the training set has to be expanded, or a more suitable fitting function has to be found

4. Classify the 'virgin' texts
   - If *precision* and *recall* are satisfying, you can classify all remaining documents of so-far unknown categories

⇒ Validation part of the method!

⇒ Required size of human-coded sets decreases with lesser categories, more discriminatory language and longer documents

⇒ 'Representative' samples of training and test documents needed (-> crowd sourcing?)

WZB

Unsupervised learning

- General idea ....

  Algorithm 'learns' both categories and categorization from the distribution of characteristics in the supplied data

- ... applied to text analysis
  - Which words tend to co-occur? Which clusters can be optimized? How can documents be distributed over clusters in a statistically optimal way?
  - Researcher does not supply any theoretical categories *a priori* (only abortion criteria, e.g. number of clusters, in some approaches)
  - Results can only be interpreted *ex post*

- Validation
  - Assessing semantic validity requires much contextual knowledge!
  - Models not generalizable beyond the data and parameters supplied!

WZB

A prominent unsupervised approach: *Topic models* (e.g. Blei 2012)

- **Typical application**
  Identification and distribution of abstract 'topics in large amounts of 'documents' without prior knowledge/assumptions on these topics

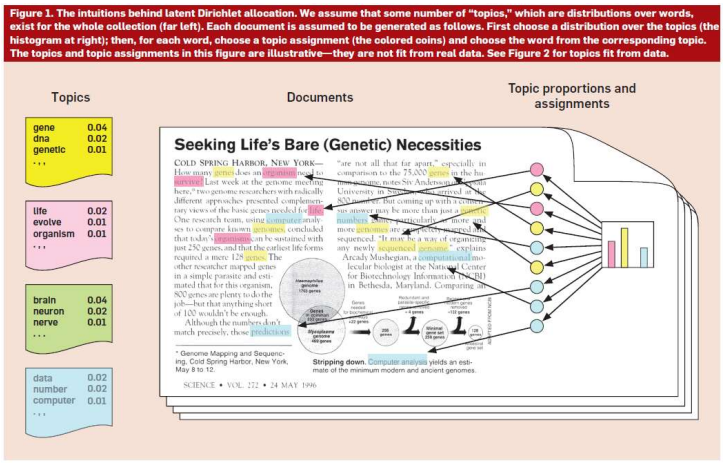
- **Assumptions**
  - Topics are defined by frequency distributions of co-occurring terms
  - Authors generate text by firstly deciding on topic composition (possibly several per text) and only then on the choice of words

- **Estimation**
  - Algorithm reverse-engineers this assumed text creation process by asking: Which latent topic distribution would explain the observed word frequency distribution best?
  - Cluster-analysis on term level, probabilistic distribution of documents

5

WZB

Th logic of topic models presented by their inventor (Blei 2012)



6

WZB

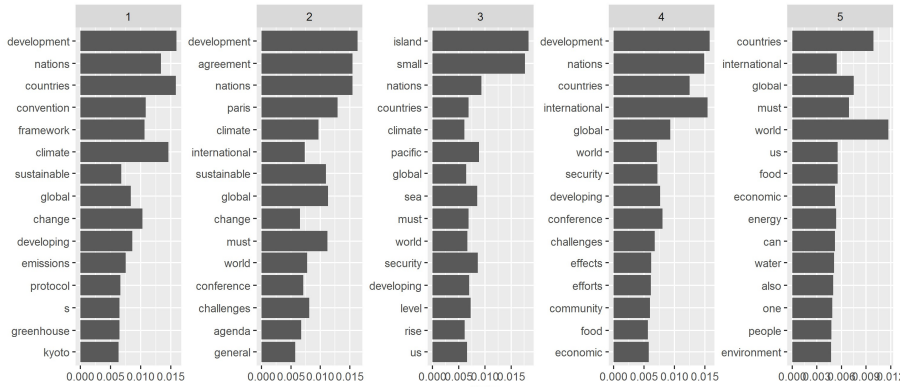
Typical output of topic models (Blei 2012)

Figure 3. A topic model fit to the *Yale Law Journal*. Here, there are 20 topics (the top eight are plotted). Each topic is illustrated with its top-most frequent words. Each word's position along the *x*-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific than "tax."

WZB

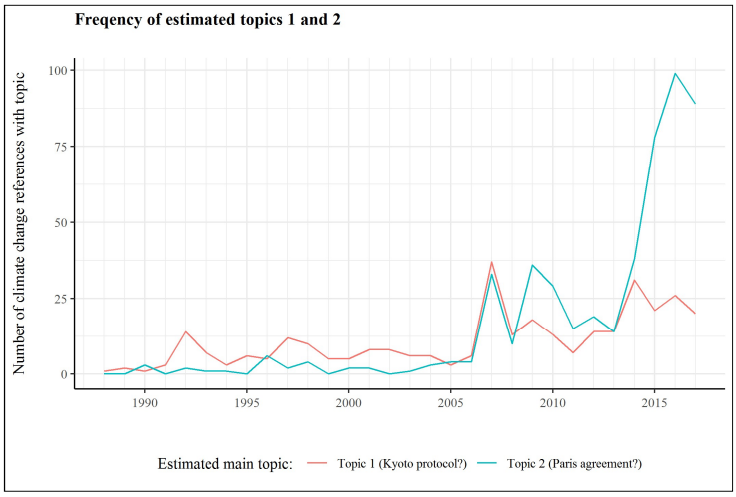
A topic model applied to our running example

- Simple LDA model with five topics estimated from the 100-term windows around UNGD climate change references
- Mild pre-processing: stopwords and country names removed
- *Can you interpret this?*



WZB

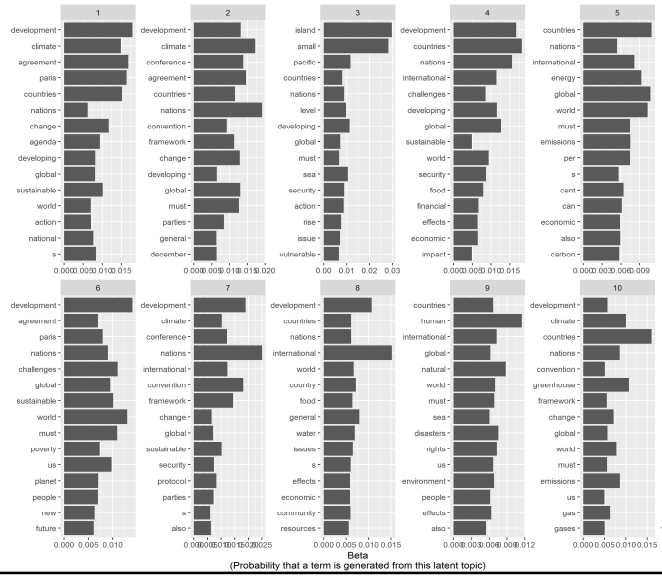
A topic model applied to our running example



9

WZB

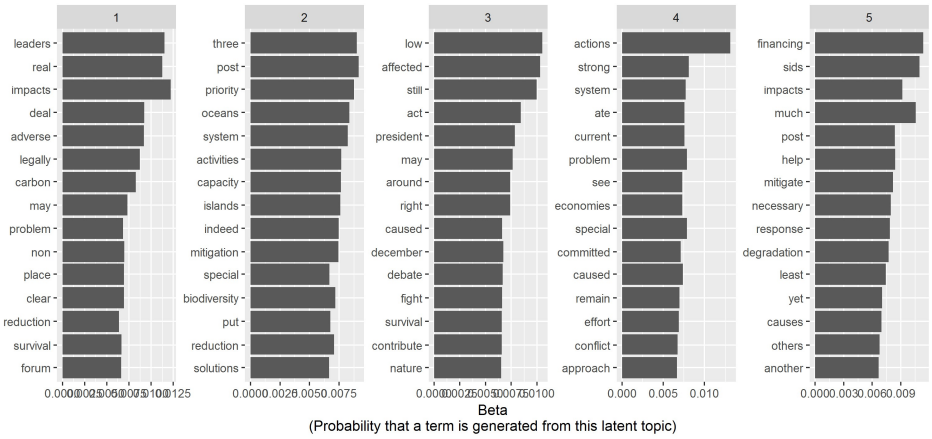
Alternative specification: 10 topics



10

WZB

Alternative specification: Feature selection



WZB

Promises and pitfalls of topic models

- **Useful for …**
  - structuring very large collections w/out priors on what to look for
  - analysing changes in political attention
  - narrow down more targeted samples for other text analyses
- **But caution:**
  - High interpretation demand after the statistical analysis!
  - Results strongly depend on the specific data set, the pre-processing steps and especially model parameters (seed and number of topics)
  - What exactly is a topic (cf. "frame", "issue", "narrative"; "event")?

⇒ **Systematic comparisons across topics only with greatest caution**

- **Recent work invests in more *guided parameter choices* (see the ldatuning package) and more robust models that take *document meta information* (Roberts et al 2014, AJPS) or *varying granularity of topics* (Green and Cross 2017, PA) into account**