Auto-IA Workshop HMTM Hannover

TOPIC MODELING FOR SOCIAL SCIENTISTS

Model selection and evaluation

Gregor Wiedemann | g.wiedemann@leibniz-hbi.de
Media Research Methods Lab
Leibniz-Institute for Media Research | Hans-Bredow-Institut

Andreas Niekler | aniekler@informatik.uni-leipzig.de
Abteilung Automatische Sprachverarbeitung, Institut für Informatik
Universität Leipzig

This lecture

1. Model selection and evaluation
   1. Manual approaches
   2. Numeric approaches
2. Variants of topic models
   1. HDP
   2. sLDA
   3. ATM, CTM, RTM, STM

Model selection and evaluation

Computer-assisted Text Analysis

- Use of Text Mining is of no end in itself for qualitative data analysis
- instead TM procedures should to be
  – embedded into a methodological framework
  – selected in accordance with the research question
  – applied in a thoughtful systematic analysis workflow
  – run with optimal parameters regarding the data
  – carefully evaluated w.r.t. to the research goal

Compatibility of Text Mining and QDA

- „4 Principles of Automated Text Analysis“ (Justin Grimmer 2013)
  - 1. All Quantitative Models of Language Are Wrong—But Some Are Useful
  - 2. Quantitative Methods Augment Humans, Not Replace Them
  - 3. There Is No Globally Best Method for Automated Text Analysis
  - 4. Validate, Validate, Validate
- Blended reading:
  - systematic combination of distant reading interpretations with close reading validation of findings (Lemke/Stulpe 2015; Lewis et al. 2013)

Model selection and evaluation

- 2 closely linked goals: selection and evaluation of models
  - selection: finding best model and parameters to fit a model with respect to the data
    - reference: models against each other
  - evaluation: procedure to determine / measure the quality of a model
    - reference: absolute quality criteria
- 2 ways of quality assessment / validity check:
  - qualitative evaluation: human judgement on model results
  - numeric optimization: algorithmic judgement

Model selection and evaluatio

terrorist, raf, schmidt, baader, haus, fahndung, jahr, politik, polizei, bka

raf, terrorist, mord, baader, bka, fahndung, buback, ensslin stammheim

- 2 closely linked g … and … nodels
  - selection: finding … ara … del with respect to the dat
    - reference: models against each … er
  - evaluation: procedure to determine / measure the quality of a model
    - reference: absolute quality criteria
- 2 ways of quality as … lidity check:
  - qualitative evaluation … n model results
  - numeric optimization … ent

$f(M_1) = 0.78$
$f(M_2) = 0.56$

Quality criteria

- Objectivity:
  - if model assumptions of the generative process of text origin hold true, algorithmic solution guarantees maximum intersubjectivity
- Validity:
  - model caputures semantic coherence prominent in and relevant for a text collection properly
- Reliability:
  - repeated runs of model inference with same parameters on the same data produce same (or at least similar) results

Challenges

- Topic models ~ semantic clusters of document collections
- Validity:
  - Evaluation of clustering as heuristic instrument is generally hard
  - Example: divide the following 2 lists each in 2 clusters:
    - A) ostrich, **penguin**, **wale**, zebra
    - B) grandson, **granddaughter**, **grandmother**, grandfather
- Reliability:
  - Stochastic process for model inference -> only nearby, not exact solutions!
- Model selection: find model parameters resulting in valid and reliable models

Challenges

- Topic model clusters of document collections
- <u>Validity:</u>
  - Evaluation of clustering as heuristic instrument is generally hard
  - Example: divide the following 2 lists each in 2 clusters:
    - A) ostrich, **penguin**, **wal**
    - B) grandson, **grand** **dmother**, grandfather
- <u>Reliability:</u>
  - Stochastic process for model inference -> only nearby, not exact solutions!
- Model selection: find model parameters resulting in valid and reliable models

Human judgement

Numeric evaluation

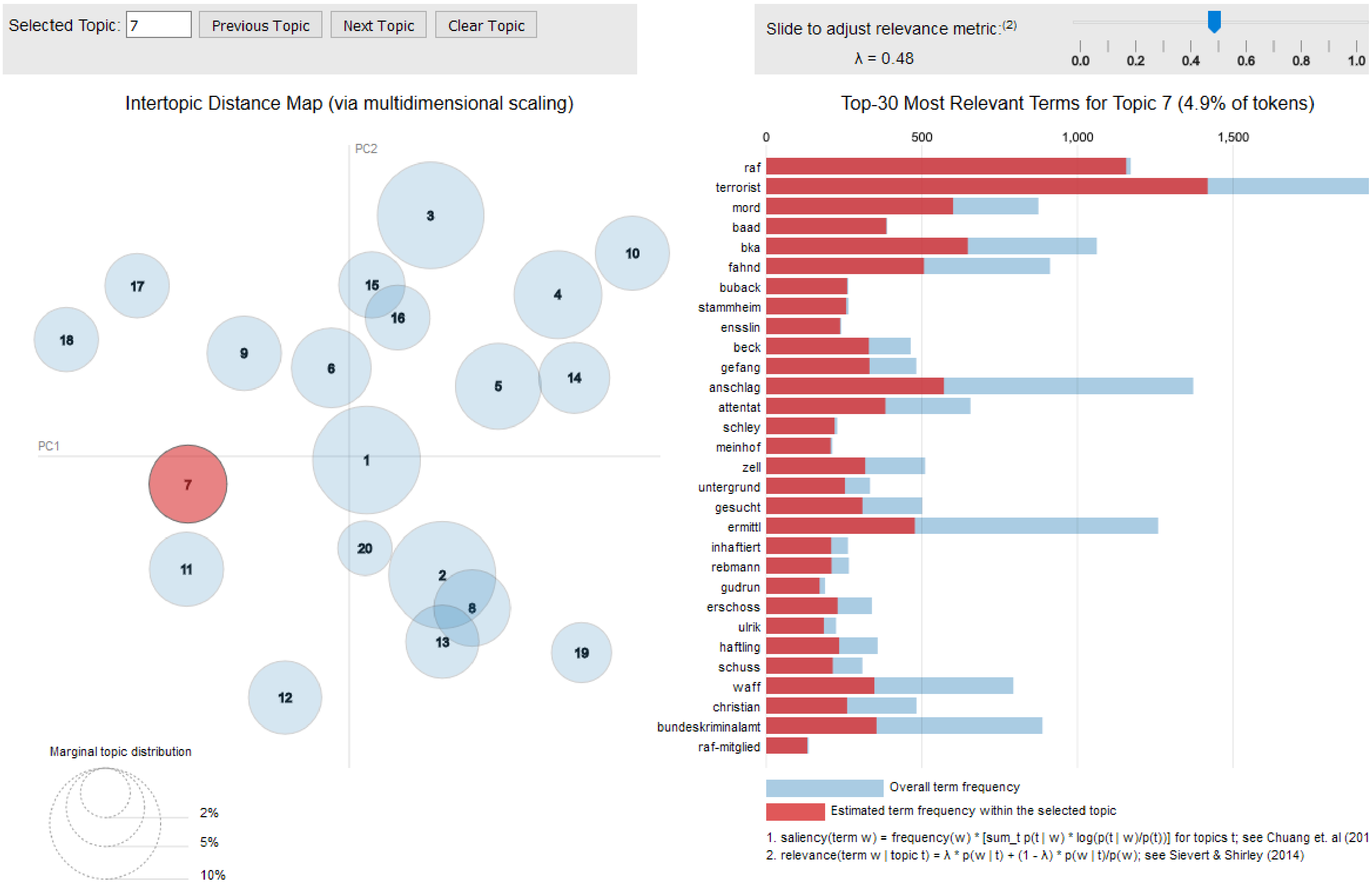
Manual evaluation

- 3 steps proposed by Evans (2014):
  - 1) check semantic coherence of top N terms of each topic: Can you assign a topic label?
  - 2) employ additional numeric measures of topic coherence to identify broad / incoherent topics
  - 3) check if topic distribution over time complies with researcher intuition
- 2 methods introduced by Chang et al. (2009):
  - Word intrusion
  - Topic intrusion
- 1 tool for visual analysisby Sievert (2014): LDAvis
  - Nearness of topics (using PCA)
  - Re-Ranking of topic terms

Manual evaluation

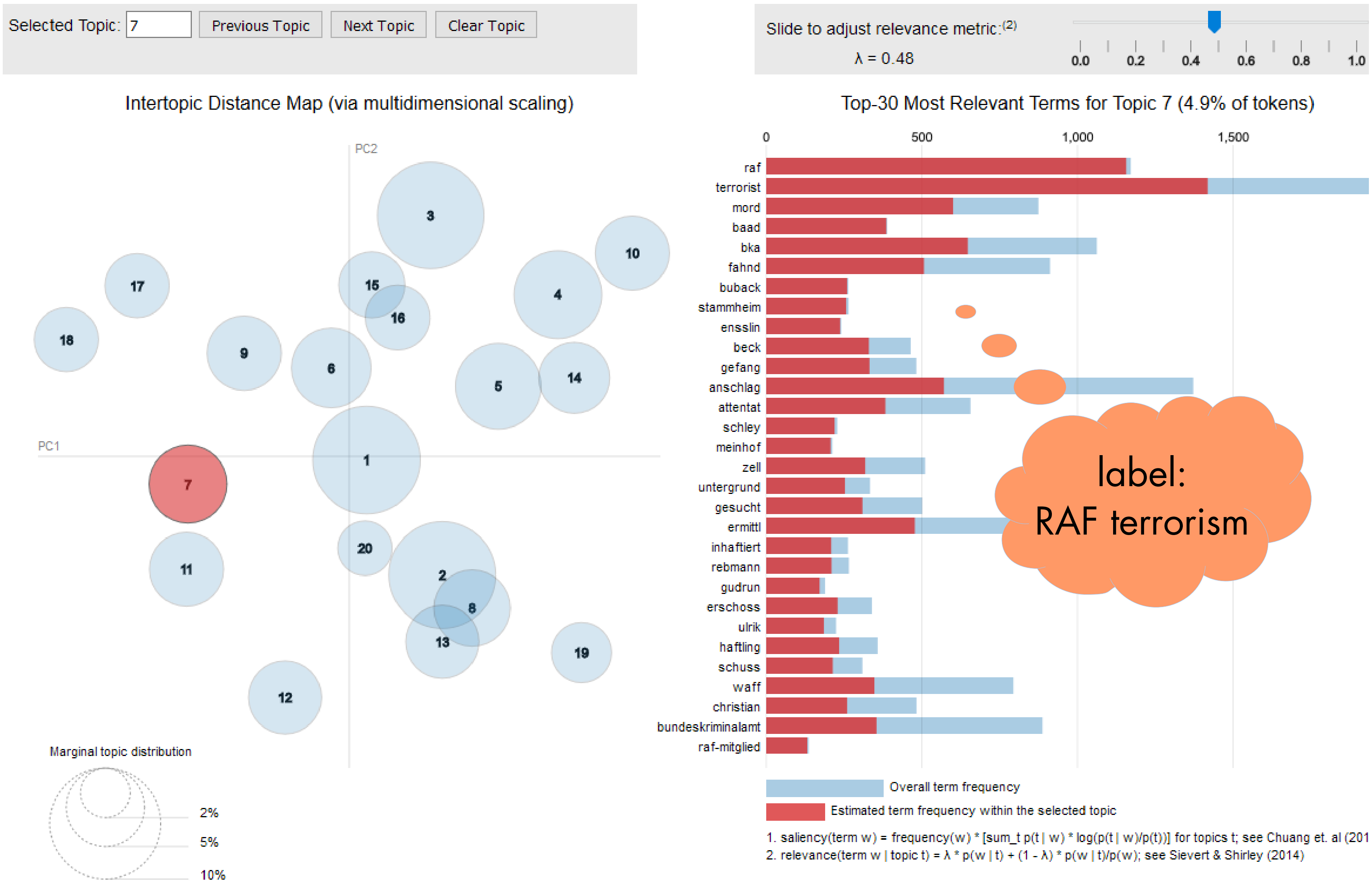
- 3 steps proposed by Evans (2014):
  - 1) check semantic coherence of top N terms of each topic: Can you assign a topic label?
  - 2) employ additional numeric measures of topic coherence to identify broad / incoherent topics
  - 3) check if topic distribution over time co[illegible] with researcher intuition
- 2 methods introduced by Chang et al. (2[illegible]):
  - Word intrusion
  - Topic intrusion
- 1 tool for visual analysisby Sievert (2014): LDAvis
  - Nearness of topics (using PCA)
  - Re-Ranking of topic terms

LDAvis



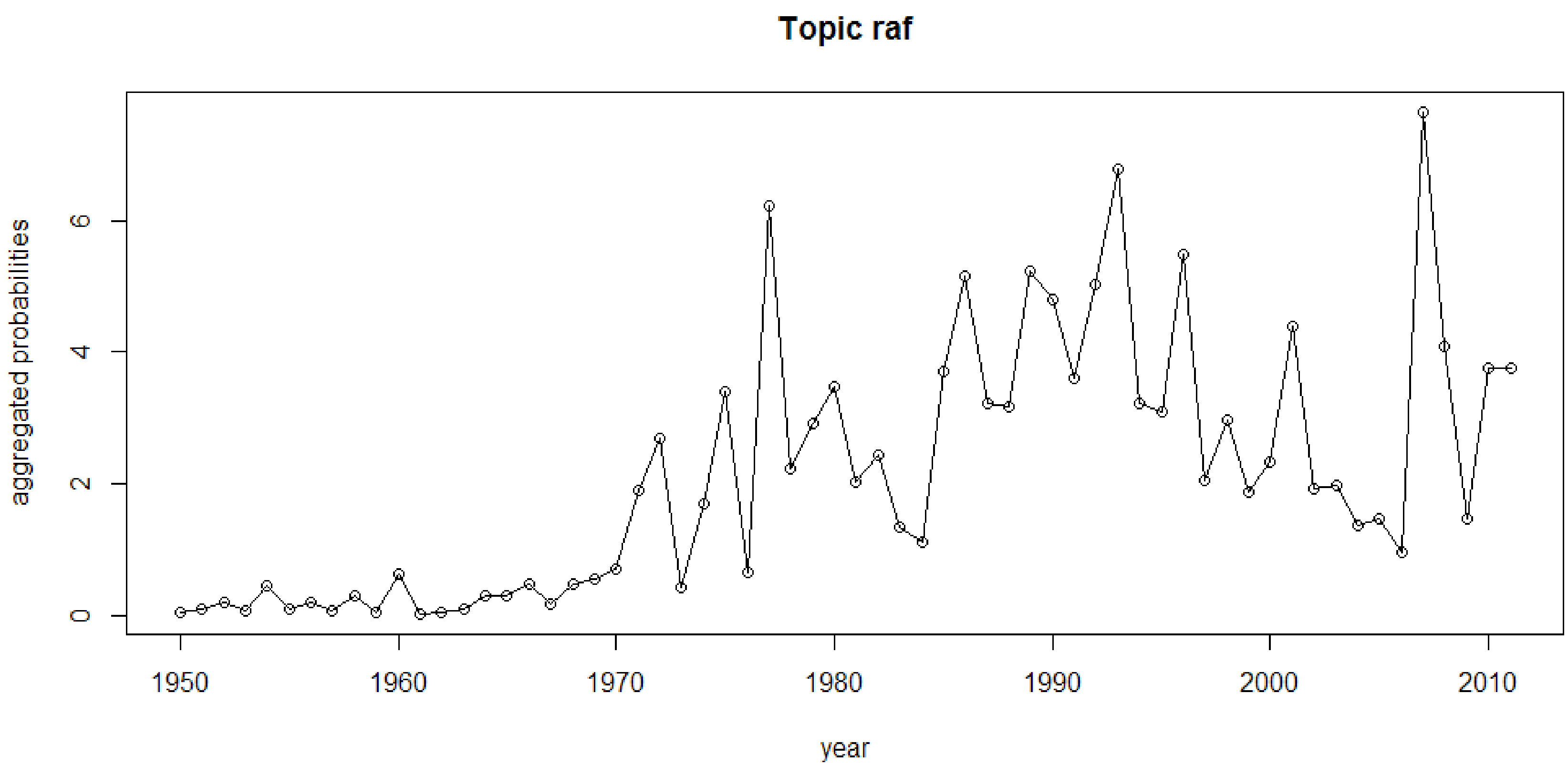
Selected Topic: 7
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.48
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
Top-30 Most Relevant Terms for Topic 7 (4.9% of tokens)
PC2
PC1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Marginal topic distribution
2%
5%
10%
0 500 1,000 1,500
raf
terrorist
mord
baad
bka
fahnd
buback
stammheim
ensslin
beck
gefang
anschlag
attentat
schley
meinhof
zell
untergrund
gesucht
ermittl
inhaftiert
rebmann
gudrun
erschoss
ulrik
haftling
schuss
waff
christian
bundeskriminalamt
raf-mitglied
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (201
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

LDAvis


Selected Topic: 7
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.48
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 7 (4.9% of tokens)
0 500 1,000 1,500
raf
terrorist
mord
baad
bka
fahnd
buback
stammheim
ensslin
beck
gefang
anschlag
attentat
schley
meinhof
zell
untergrund
gesucht
ermittl
inhaftiert
rebmann
gudrun
erschoss
ulrik
haftling
schuss
waff
christian
bundeskriminalamt
raf-mitglied
label:
RAF terrorism
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (201
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

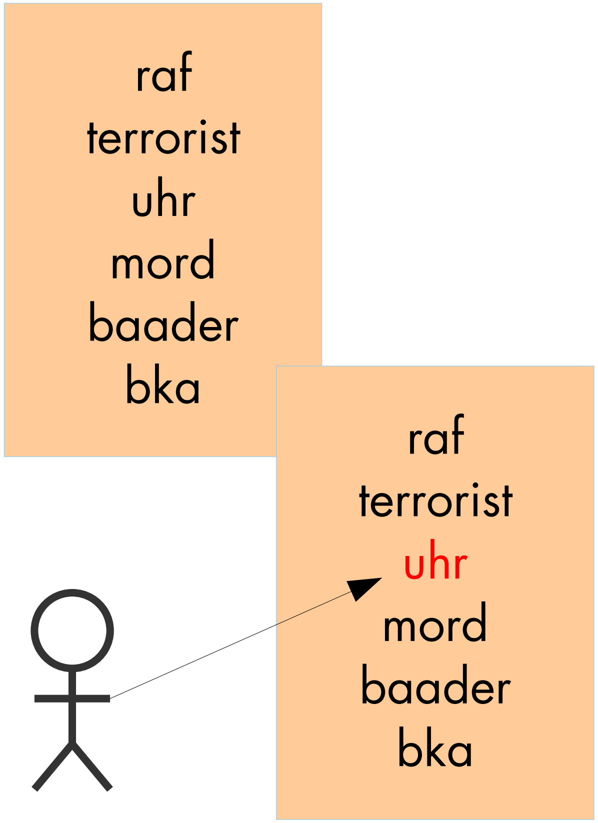
Time series

- Option 1: Aggregate probabilities by time period
- Option 2: Count documents per time period containing a certain topic
- Compliance with researcher intuition?



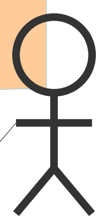
Word intrusion

- Idea Chang et al. (2009):
  - top topic words should represent semantic coherence; coherence could be evaluated by finding inappropriate intruder term
- Experiment:
  - repeat n times for random topic k
    - create top word list L for topic k
    - choose intruder term t not in top words of k, but relevant in other topic
    - put intruder word into shuffled L
    - ask evaluator to find t in L
  - calculate correct guesses / n



Topic intrusion

D4: TERRORISTEN|Vor dem Oberlandesgericht Düsseldorf muß sich die mutmaßliche Terroristin Angelika Speitel wegen Mordes verantworten. Es geht auch um Beteiligung an der Ermordung von Bu-back, Ponto und Schleyer. [...]



T1: raf, terrorist, mord, baader, bka
T2: polizei, daten, burg, vs, hamburg
T3: deutsch, muslim, islamist, anschlag
T4: gericht, jahr, richt, verfahren, vs

- Idea: read document (or at least beginning of it) and find intruded topic from presented list
  - sample document d
  - get 3 most prominent topics in d
  - select 1 topic not prominent in d
  - let user choose suspected intruder from shuffled list

Word / topic intrusion

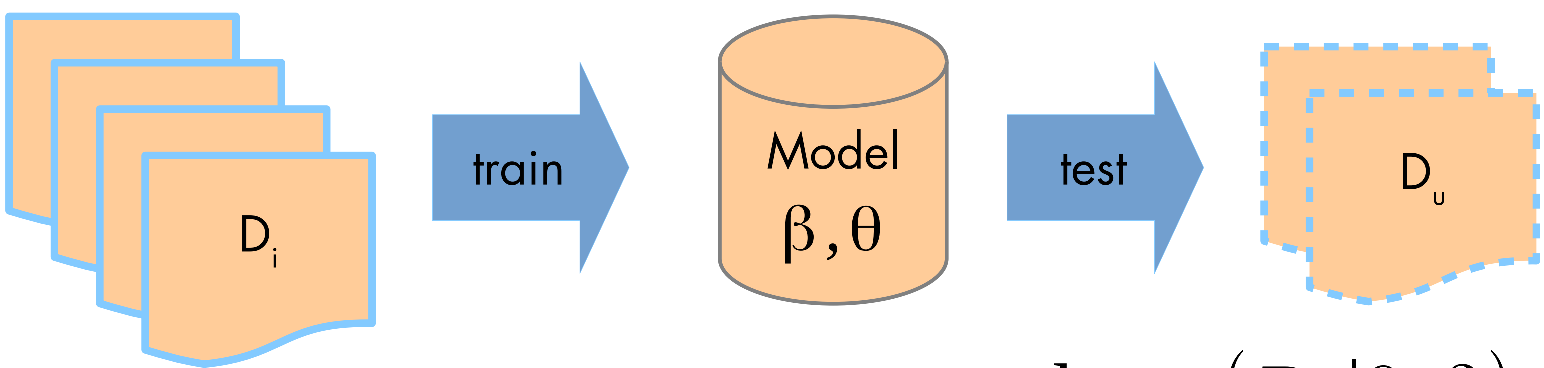
- (Dis-)Advantages:
  - + provide substantial numeric measures in range [0,1]
  - + allows for comparison of models against each other
  - - large effort for substantial evaluations of multiple models
  - - human evaluators perform different in this task
  - - aspired quality not clear (~0.7 cp. to inter-rater reliability in content analysis?)

Numeric evaluation

- Goal: Entirely automatic approaches to judge on model quality
  → determine model quality in one numeric measure
- 3 Approaches:
  - Perplexity (Wallach et al. 2009)
    - How well performs generalization of a learned model to unseen data?
  - Coherence (Mimno et al. 2011)
    - How often do we oberve predicted semantic coherence actually in the data?
  - Reliability (Lancichinetti et al. 2015, Koltsov 2012, ...)
    - How reproducible are model results between repeated inference runs?
- CAUTION: None of them replaces careful manual inspection!

Perplexity

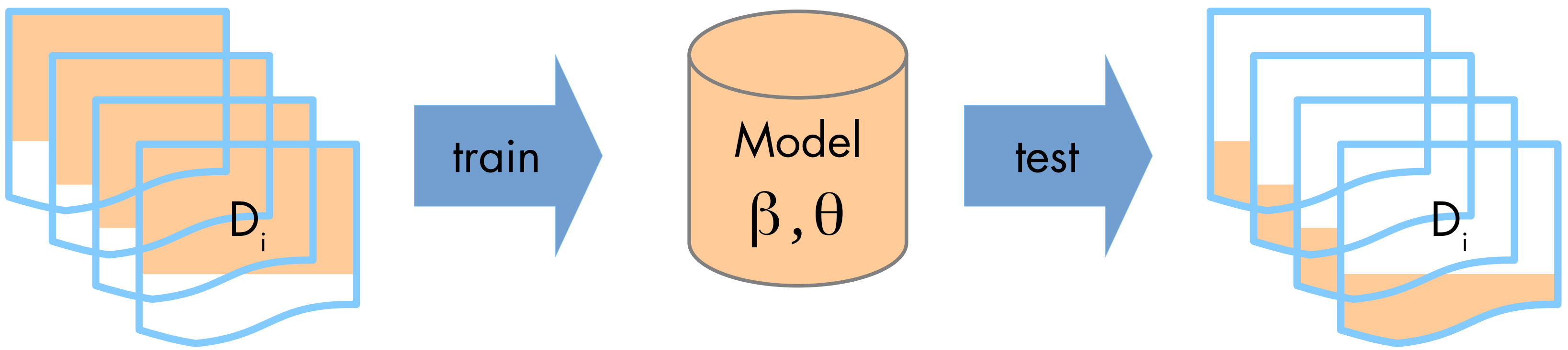
- Perplexity:
  - surprise of the model, when presented with new data -> a.k.a „Held-out Log Likelihood“
  - What is the probality of the words in a test documents under the pre-trained model?
- Assumption:
  - The lower the perplexity, the better model captures semantic coherence in the collection



$$2^{-\log p(D_u|\theta,\beta)}$$

Perplexity

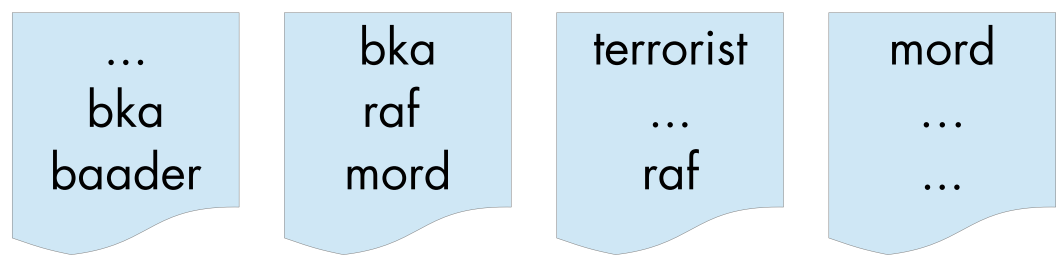
- Variant: Document completion
  - Use X % of document content for training and remaining 100 – X % for testing



$$2^{-\log p\left(split\left(D_i,2\right)|\theta,\beta\right)}$$

Topic Coherence

- Chang et al. (2009):
  - large user studies with word intrusion and topic intrusion
  - low perplexity does not correspond well with user perecption of coherent topics
- Mimno et al. (2011):
  - Idea: measure co-occurrence of highly-probable topic terms in documents instead of perplexity
  - The higher the coherence, the better the model captures actual semantics
  - Higher correlation of the coherence measure with user perception than perplexity



$$C(k, V^k) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} log \left(\frac{D(v_n^k, v_l^k) + 1}{D(v_l^k)} \right)$$

k – topic k

V^k – top N words of topic k

D(t)– number of documents containing t

Perplexity / coherence

- (Dis-)Advantages:
  - + provide substantial numeric measures in range
  - + allows for comparison of models against each other
  - + coherence allows for assessment on single topics k, and entire models → mean of coherence(k) for all k in 1:K
  - – no bounded value range → no absolute comparison
  - – high coherence seems to correlate with low priors → overfitting
  - – as single optimization goal they miss the goal of inference of good models too!

Reliability (a.k.a. stability)

- numerous local optima of $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})$
- solution possible only via stochastic inference
- → random sampling → results near optimal solution, but varying in probability space
- depending on
  – parameter initialization
  – sampling strategy
  between repeated inference may vary greatly
- quality criterion in social science: determine reliability of measurement instruments!

Reliability

- Idea:
  - compare pairs of models of repeated inference runs
  - measure similarity of results, e.g. how many topics can be reproduced reliably
- Challenges:
  - identification of matching topic pairs ← no stable identifier due to stochastic inference process
  - definition of similarity: when are topics considered „equal“
- State-of-the-art:
  - problem identified in social science (e.g. Lancichinetti in 2015)
  - Approaches:
    - Roberts et al. 2016 (STM): Spectral Clustering for initialization + Random Seed fixation, fully reproducible
    - Maier et al. 2018 (LDA): LL-Co-occurrence Clustering for initialization, increased stability
    - Rieger 2020 (LDA): ldaProtoype, select the LDA run from N = 100 runs with highest mean pairwise similarity

Reliability

- 2 types of approaches two compare two models $m_1 = (\beta_{1:K}, \theta_{1:D})$ and $m_2 = (\beta'_{1:K}, \theta'_{1:D})$
- Approach I: matching topics topic-term-distributions β:
  - choose similarity measure SIM and define similarity threshold s
  - similarity measures:
    - Kullback-Leibler-Divergence (KLD), Jensen-Shannon-Divergence (JSD) (Koltsov 2012)
    - Cosine Similarity on top N topic words (Niekler 2015)
  - for each β_k find most similar β'_k where $SIM(\beta_k, \beta'_k) > s$

Reliability

- compare two models $m_1 = (\beta_{1:K}, \theta_{1:D})$ and $m_2 = (\beta'_{1:K}, \theta'_{1:D})$
- Approach II: matching topics by document-topic-distributions θ (Lancichinetti 2015):
  - topic distribution given document $p(k|d)$ from θ and θ' cannot be compared directly due to unknown topic matching → Idea:
    - **compare** $\mathbf{p(d|k)}$ because document indexes are fixed and known
    - $p(d|k)$ can be obtained via Bayes' Rule: $\mathbf{p(d|k) = p(k|d) * p(d) / p(k)}$
  - calculate manhattan distance on p(d|k) and p(d|k') for all topic pairs from m_1 and m_2 and match least distant topics (best match)
  - correct for distance obtained by randomly sampled topic distribution over documents
  - Reliability score = average chance corrected manhattan distance between matched pairs

Reliability

- (Dis-)Advantages Approach I (comparing β and β'):
  - + provides measure in range [0;1]
  - + follows analysts intuition of comparing semantic coherence of terms
  - + especially cosine distance concentrating on top topic terms
  - – measuring similarity with KLD or JSD for comparing probability distributions (information loss) is less intuitive
  - – measures need to assume theresholds for similarity -> high influence on reliability score
  - – cosine measure also need parameter N for top topic words to match

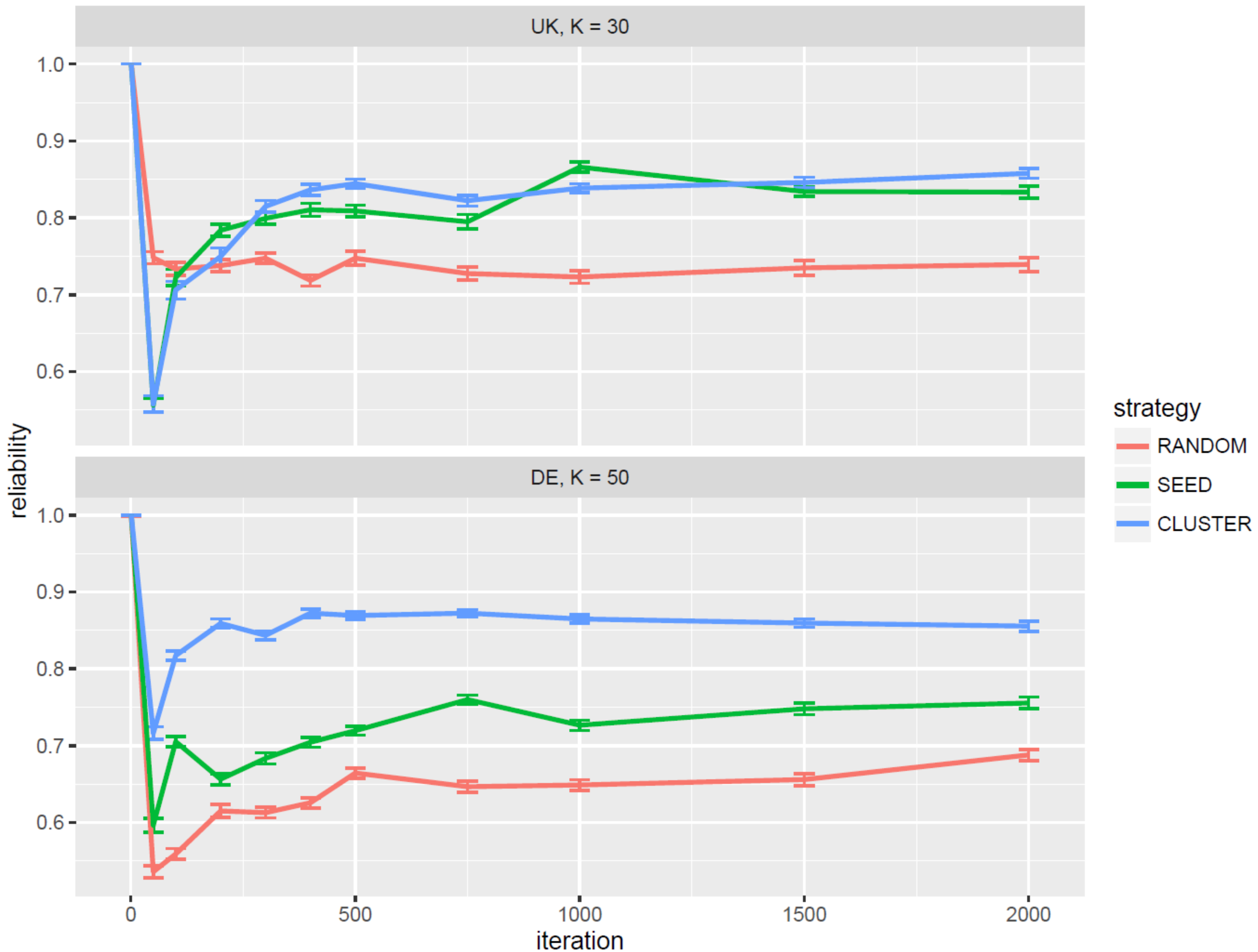
Reliability

- (Dis-)Advantages Approach II (comparing θ and θ'):
  - + provides measure in range [0;1]
  - + does not rely on thresholds for comparison
  - + chance correction
  - +/– considers unequal importance of topics by weighting distance with overall topic probability
  - – no bipartite matching of pairs from m_1 and m_2 guaranteed
  - – Manhattan distance on probabilities less intuitive
  - – rather conservative scoring of reliability

Increasing reliability

- Reducing K:
  - lowered number of topics → more stable clusterings
- Fixed initialization of topic assigments to words $\mathbf{z}$ before/during Gibbs sampling
  - Seed: random, but fixed initialization (does this really improve model reliability?)
  - Clustered: using term co-occurrence clusters as informed prior for fixed initialization
    - Variant 1: „Topic Mapping“ by Lancichinetti et al. 2015: initialize $\mathbf{z}$ by term co-ooccurrence clusters of documents in the collection; run Topic Model inference for only 1 iteration → Reliability = 1 (but: it is actually not longer topic modeling…)
    - Variant 2: run Topic Model inference for N iterations → Reliability < 1, but still improved (next slide)
  - Change sampling process as suggested by Koltsov 2012: force sampled topic for word w onto its left and right neighbors -> co-occurring words tend to have same topic; But:no straightforward implementation on bag-of-words representations in R
- Pragmatic approach:
  - Leave out unreliable / in incoherent topics in final analysis

Increasing reliability



Best Practice Suggestion (Maier et al. 2018)

- General advice:
  - avoid model selection solely based on numeric evaluation measures (!correspondence with human judgement)
  - make theoretically sound selections insteand and check manually
- Workflow:
  - 1. Preprocessing: clean documents/remove biolerplate, lowercase, remove punctuation, remove stop words, remove infreqent terms (df(w) < 0.5 % document frequency), lemmatization/stemming
  - 2. (initialize topic assignments for LDA)
    - a) set seed, or
    - b) cluster terms by their co-occurrence statistics
  - 3. Compute a variety of models with different parameters K, alpha, (fix eta = 1 / K)
    - 1. for each K, select model with alpha wih highest topic coherence
    - 2. select model with best interpretable K topics (use LDAvis as helper tool)
  - 4. Validate selected model
    - rank words: term probability + lambda relevance score (LDAvis) → interprete semantic coherence → label
    - rank topics: topic probability + rank1 (background vs major topics), coherence (compared to other topics)
    - read N documents for each topic with highest topic probability
    - check reliability to repeated inference runs
  - 5. Final analysis: time series, cross-sectional analysis
    - leave out uninterpretable models
    - leave out unreliable models

Automatische Sprachverarbeitung

END