Topic Detection and Tracking System

Research Project 2nd Semester

Group 5: Daniel Brüggemann, Yannik Hermey,
Carsten Orth, Darius Schneider, Stefan Selzer

Agenda

- Introduction
- Latent Dirichlet Allocation (LDA)
- Non-Negative Matrix Factorization (NMF)
- Visualization
- Outline - Future Work

Introduction - Topic

- Increase in text data
- Manual handling impossible
- Mining algorithms as a solution for topic detection
- Dynamic data handling
  - Tracking of changes
  - Streaming data
  - Hardly any extensions

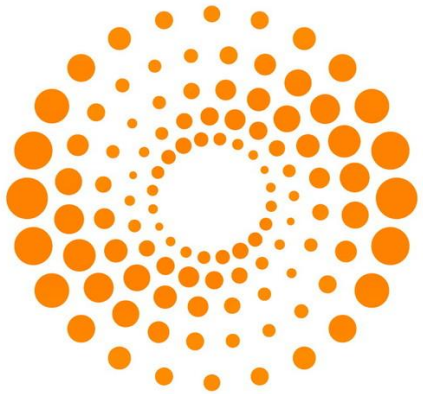
Introduction - Goals

- Topic Extraction
  - Track topics over time
  - Recognize emerging and fading topics
- Topic Extraction Techniques
  - Latent Dirichlet Allocation (LDA) ➜ Dynamic Topic Detection
  - Non-Negative Matrix Factorization (NMF)
- Interactive visualization for the topics over time
  - Show corresponding terms and documents

    ranked by significance

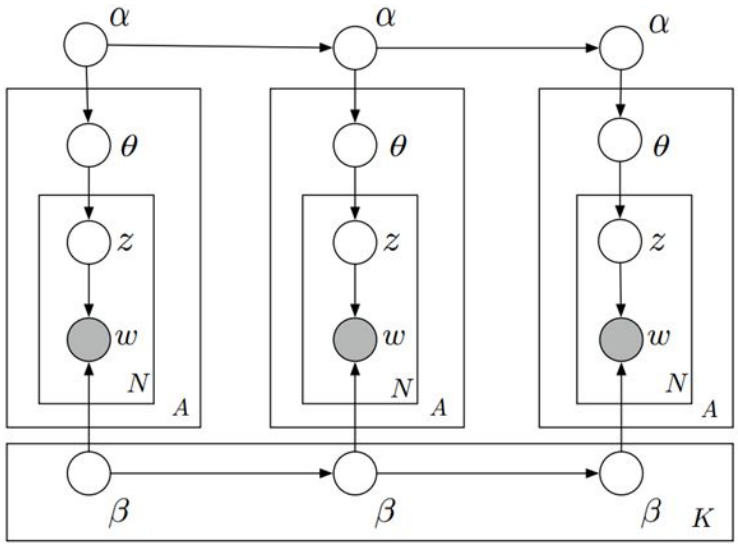
Introduction - Document Collection

- Reuters Corpus RCV1
- August 1996 - August 1997
- ca. 800.000 documents
- ca. 100.000.000 words
- ca. 400.000 terms
- 103 annotated topics in corpus



Dynamic LDA - Blei & Lafferty

- Latent Dirichlet Allocation
- Bag-of-words, no semantics
- Frequency of occurencies of words given by a vocabulary
- Probabilistic topic model
- Word distribution per topic
- Topic distribution per document
- Dynamic: extended by distribution over time steps



Dynamic LDA - Document Collection

- First 2 months (August and September 1996) of Reuters Corpus RCV1
- 83.650 documents (ca. 10% of documents overall)
- 42 days → 6 time steps a 7 days
- 12807 - 14997 documents per time step, almost evenly distributed
- 104 topics extracted
  - 103 annotated topics in corpus
  - 1 for unlabeled documents
- Words in documents overall : 16.467.261

Dynamic LDA - Preprocessing

- Generate vocabulary containing meaningful words :

Distinct terms	308.854
NER category removals	133.976
Lemmatization removals	17.372
Regex cleanup removals	18.962
Spellcheck cleanup removals	6.768
Stopword removals	574
Final vocabulary size	131.202

Dynamic LDA - Detected Events

- Extracted topics reveal events from August and September 1996

Child abuse in Belgium	Tropical storm Edouard	Peace talks in Palestina	Kurdish war in Iraq
child	storm	israel	iraq
police	hurricane	peace	iraqi
woman	north	israeli	iran
death	wind	netanyahu	kurdish
family	west	minister	turkey
girl	mph	palestinian	northern
dutroux	mile	arafat	arbil

Dynamic LDA - Topic Evaluation

- No match of LDA topics with RCV1 annotated topics assigned by f-score matrix
- Best assignment :
  - F-score : 0.288 , precision : 0.432 , recall : 0.216
  - RCV1 GSPO : Sports
  - LDA Topic:

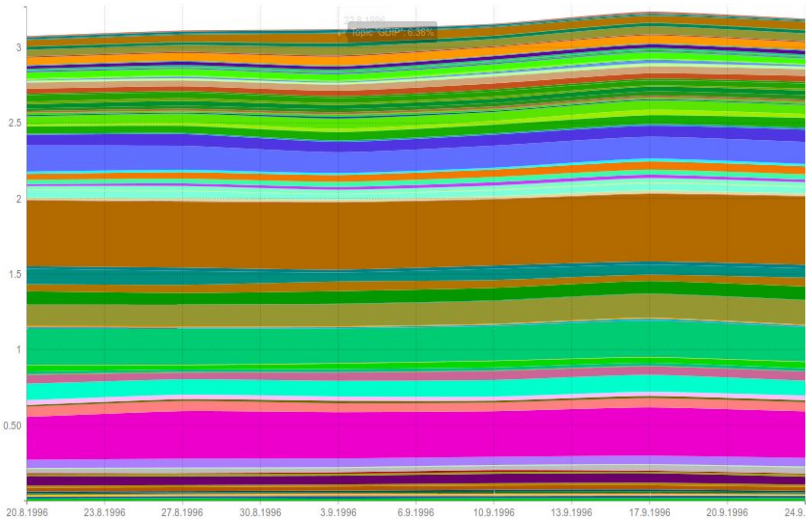
party minister election government political opposition parliament president prime leader vote poll former national coalition democratic deputy candidate power ruling

Dynamic LDA - Topic Rivers

- Topic rivers for August and September 1996 of RCV1 corpus

Dynamic Topic Model

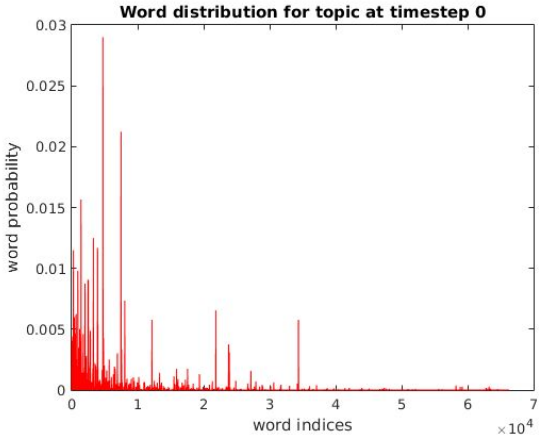
RCV1 annotated topics

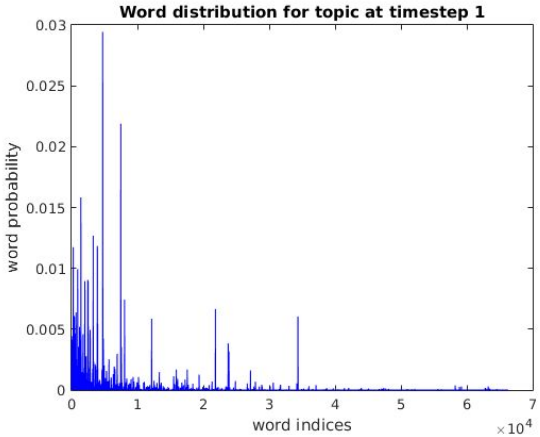


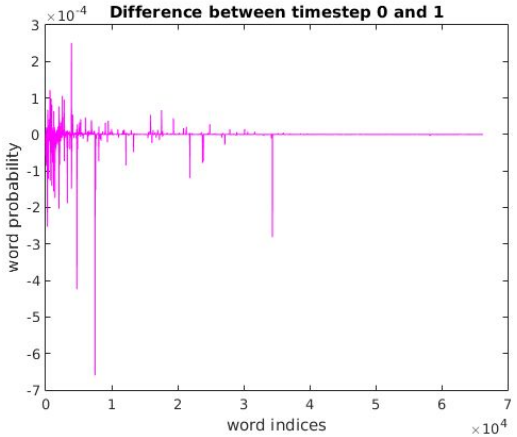
Dynamic LDA - Word Distributions for Iraq Topic

- Example word distribution at two timesteps:

  iraq iraqi force northern kurdish official united iran war troops baghdad
  iranian kdp attack gulf leader puk military fight faction ...





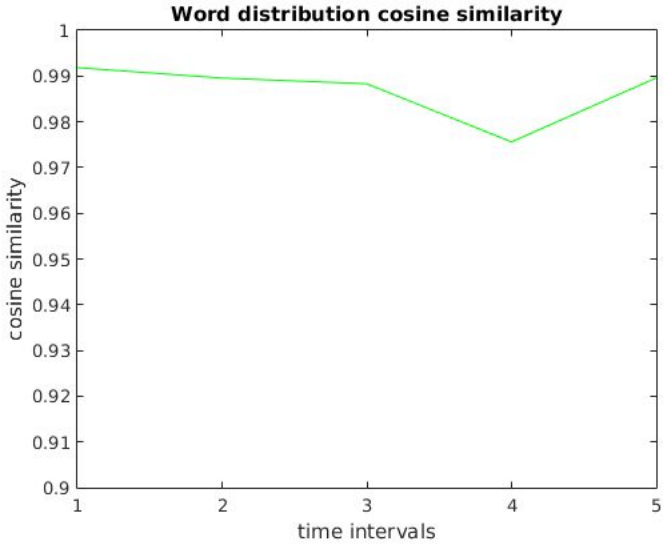


Dynamic LDA - Topic Emergence for Iraq Topic

- Cosine similarity used as similarity measure for topic's word distributions at each time step.
- Topic emerges into a new one, if difference of similarities of two time step exceeds a threshold.

  → Topic turning point

- Threshold for given data is 0.01.



Dynamic LDA - Top Words for Iraq Topic

- Changes of the top words at the topic turning points

week 1	week 4	week 5
iraq	iraq	iraq
missile	missile	gulf
attack	gulf	kuweit
saudi	iraqi	military
iraqi	military	missile
military	kuweit	iraqi
gulf	attack	united

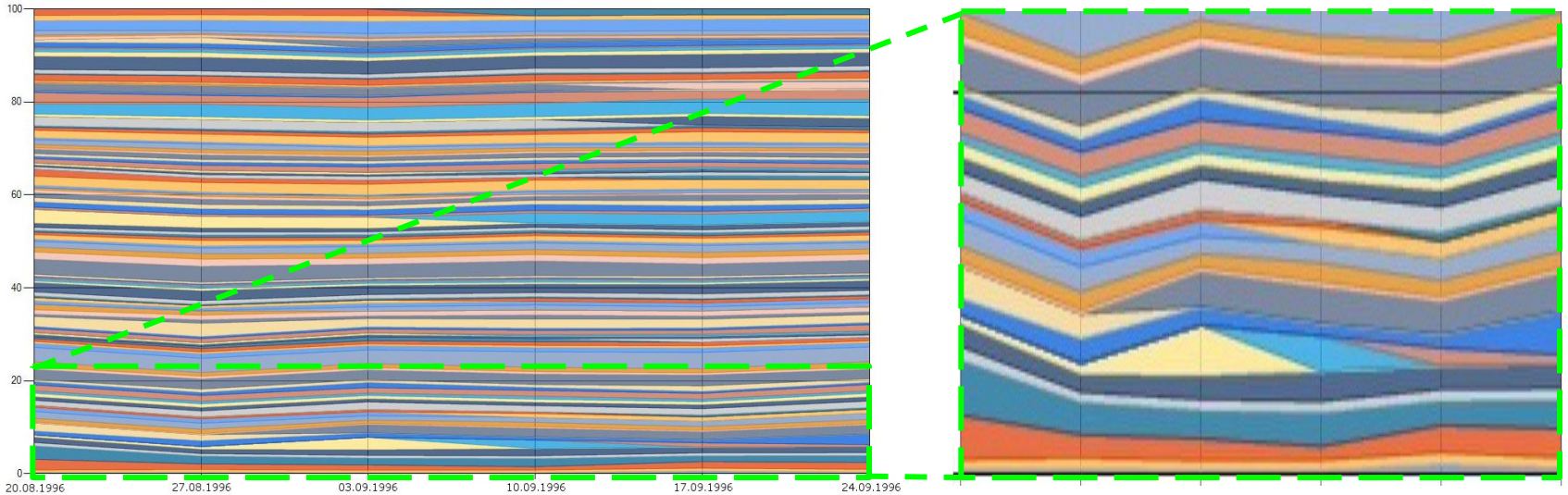
Dynamic LDA - Top Articles for Iraq Topic

- Article Headlines of top documents at the different time spans

week 1 - 3	week 4	week 5 - 6
Perry cites two incidents in Iraq no-fly zone.	Iraq fires at U.S. jets, U.S. bombers move closer.	U.S. boosts Kuwait defence by deploying Patriots.
U.S. warns it will protect pilots over Iraq.	U.S. gets Kuwaiti approval for troops deployment.	U.S. ground troops set to fly to Gulf.
Defiant Saddam urges his warplanes to resist U.S.	Kuwait agrees new troop U.S. deployment.	U.S. carrier enters Gulf, troops land in Kuwait.
Saddam urges his warplanes and gunners to resist.	Iraq says fired missiles at US and allied planes.	U.S. sends last of 3,000 ground troops to Gulf.
U.S. launches new attack on Iraq - officials.	Iraq fires at U.S. jets, U.S. bombers move closer.	U.S. declines to rule out Iraq strikes.

Dynamic LDA - Emerging Topics Topic Rivers

- Topic river for August and September 1996 of RCV1 corpus
- Threshold of 0.01 produces 20 turning points where topics emerge into new ones



Non-Negative Matrix Factorization

- Factorize large matrix into two submatrices

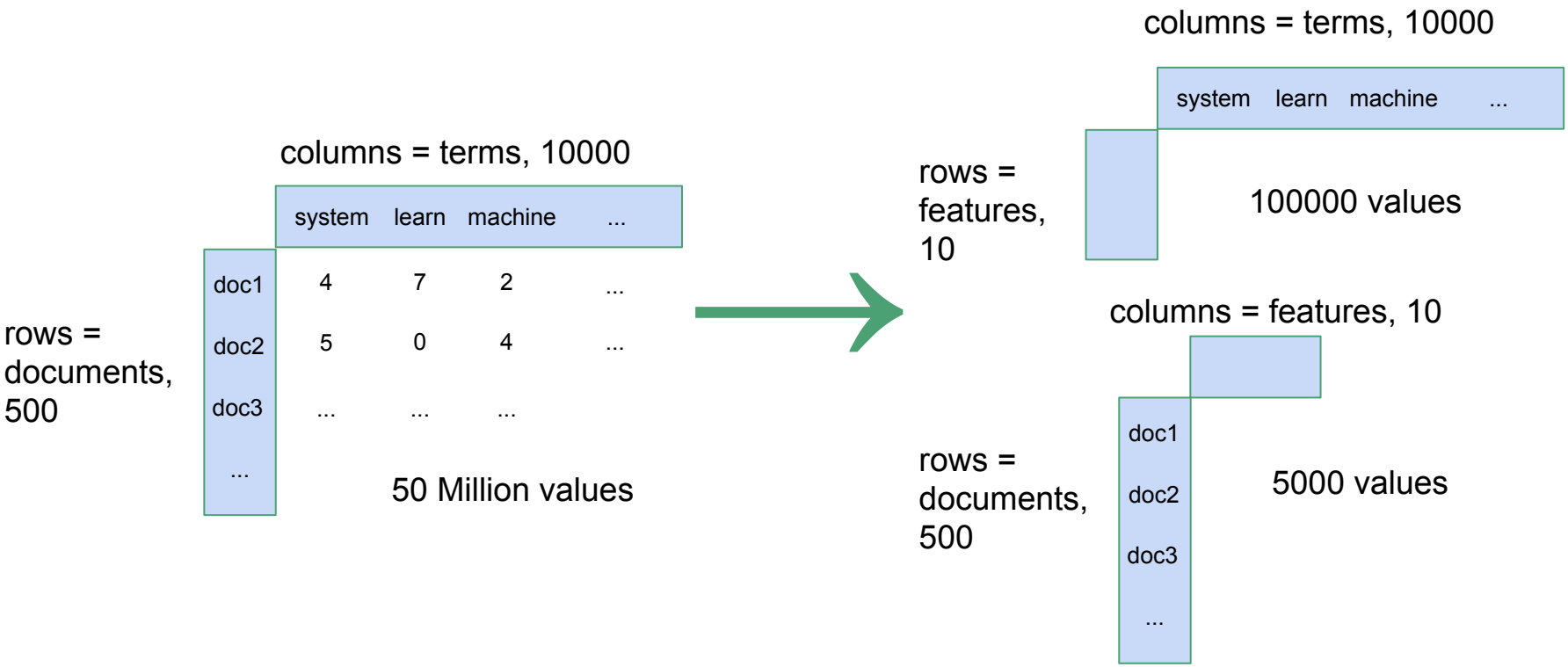
  ➜ Significant decrease in dimension

- Requirement: only non-negative matrix values

  ➜ given with term frequencies

- Submatrices form approximation of original
- Reduction reveals features ➜ in this case, topics

Non-Negative Matrix Factorization



NMF - Preprocessing

Initial vocabulary	109.259
1 - Porters stemmer	84.769
2 - Stopword removal	84.618
3 - americanization	84.578
4 - remove low occurrences	12.524

NMF - Implementation

- Java Library: Linear Algebra and Machine Learning (LAML)
- Preprocess terms in files, create vocabulary
- Input: Document-Term matrix
- TF-IDF values instead of raw occurrence count:
  - TF = Term Frequency: how often the term appears in the document
  - IDF = Inverse Document Frequency: in how many documents the term appears

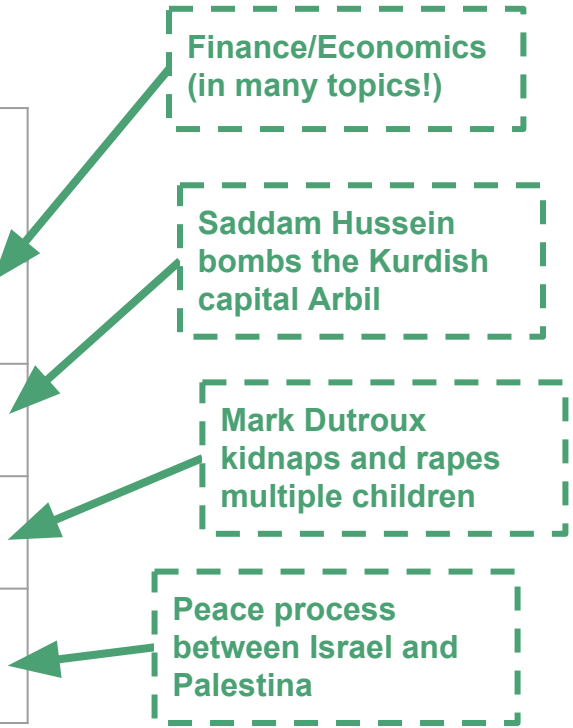
$$TFIDF = TF * \log(N / DF)$$

N = number of documents,
DF = document frequency

NMF - Implementation

Extracted topics for 1 week: **29.08. - 04.09.1996**

Topic Rank	Topic Terms						
1	indexed (8.81)	stock (5.23)	shares (5.22)	market (4.02)	dollars (3.76)	trading (3.23)	dow (1.83)
2	iraq (10.24)	iraqis (9.51)	saddam (5.50)	kurdish (5.23)	arbil (4.65)	baghdad (4.21)	attack (3.52)
9	dutroux (10.55)	children (4.00)	police (3.96)	sex (2.71)	bodies (2.36)	paedophile (2.23)	belgians (2.18)
13	palestinians (14.28)	netanyahu (11.78)	arafat (11.10)	israeli (10.75)	israel (8.97)	peace (5.30)	plo (4.71)



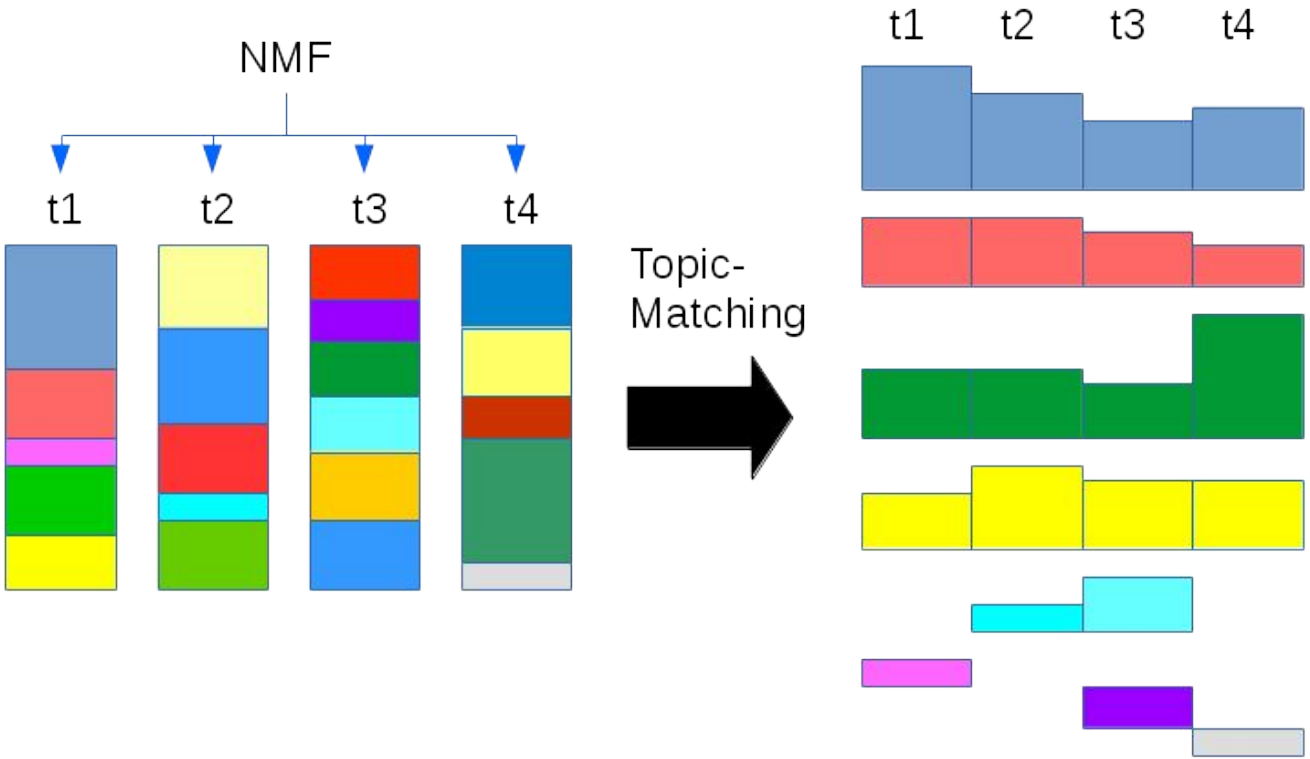
NMF - Over Time

- Apply NMF on segments of the data

  ➔ on all weeks of a year, so 52 NMF runs

  ➔ ~50MB per serialized XML file
- Build Topic Waves: for each time step, add each new topic to the wave that matches best (threshold)
- If no wave matches, new wave starts at this time step
- If a wave has no matching topic for a time step, it vanishes

Topic Matching



Topic Matching (Metrics)

Metrics:

- Number of exact matches for 20 best terms
- **Weighted sum of matching terms**
- Cosine distance

Topic 1: greece (0.022) euros (0.014) zones (0.012) ecb (0.010) tsipras (0.010)
Topic 2: greece (0.024) tsipras (0.016) euros (0.014) eu (0.013) athens (0.013)

Matching?

(0.022+0.024)+(0.014+0.014)+(0.010+0.016) > **threshold** ?

Topic Wave Example: Ukrainian Crisis (2015)

Topic Wave Rank 16 +++ Weight: 1.31% +++ Top 5 Terms: ukraine, russia, russians, sanctions, moscow

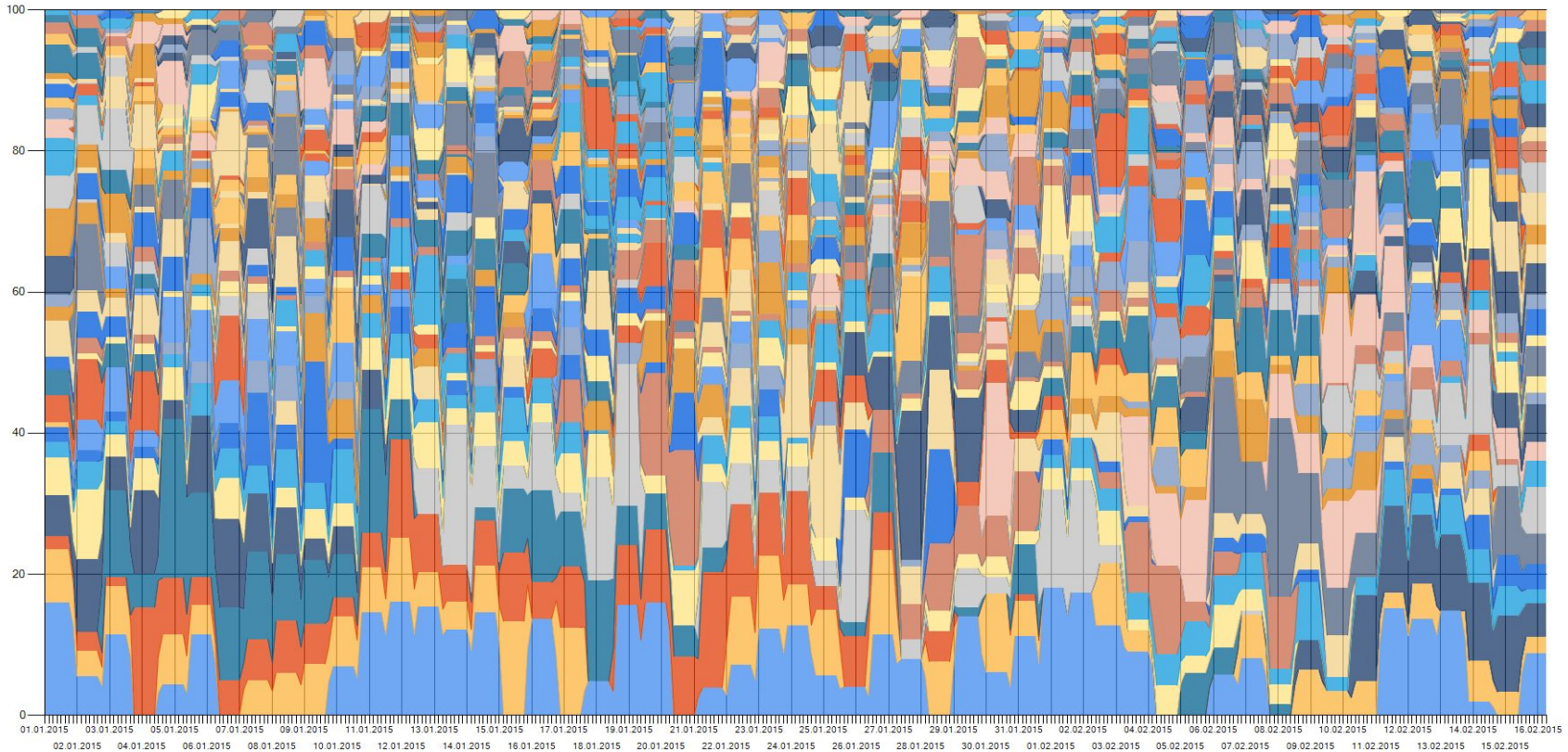
Time Step	Best-ranked Terms				
January 15th	russia	ukraine	eu	moscow	**sanctions**
February 26th	**nemtsov**	putin	opposition	boris	**killed**
March 19th	ukraine	russia	**kolo-moisky**	moscow	kiev
July 2th	**nato**	russia	putinism	strategy	**military**
November 19th	**crimea**	ukraine	russia	**power**	**annex-ation**

Worked under Yeltsin, critic of Putin, assassinated on February 27th

Ukrainian billionaire Kolomoisky loses power by Poroschenko

NATO gets involved - conflict escalates

Topics for the Year 2015



Topics for the Year 2015

Topic Rank	Best-ranked Terms					
#2	games	goals	wingers	play	periods	4.4%
#3	gmt	federal	banks	diaries	reserves	4.2%
#15	islamic	goto	pilots	jordan	jordanians	1.4%
#16	ukraine	russia	russians	sanctions	moscow	1.3%
#18	euros	greece	ecb	zones	germanic	1.2%

Comparison: LDA vs. NMF Approach

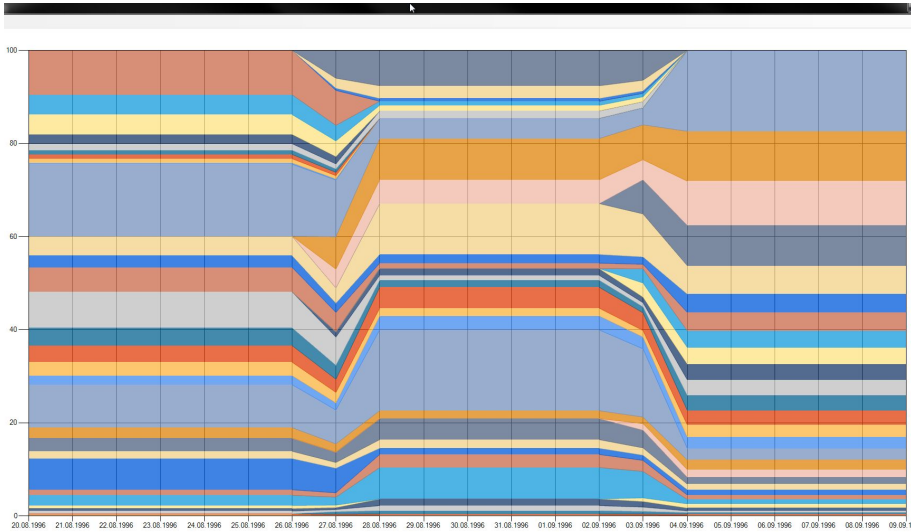
- Dynamic Topic Model Implementation ↔ NMF applied on segments, then joined via similarity metric
- LDA is much slower, vocabulary cannot be reduced as much as for NMF
- Topics in LDA change only little ↔ NMF topics are split up in many small topics
- Both algorithms are able to develop comprehensive topics, reflecting the major real-life events.

Visualization - Formatting the Data

- some named, indexed topics
- naming is based on word-prevalence for a topic
- define startdate and time-interval in days
- arbitrary number of consecutive time-intervals, f.e. 365 intervals á 1 day
- each topic has a value per day and a list of document names

Visualization - Draft

- Implement as GUI application
- Event-driven design (interactive)
- Display as a stacked graph
- Create graph using Charting-library
- Implement tooltips and zooming
- Show documents belonging to each topic



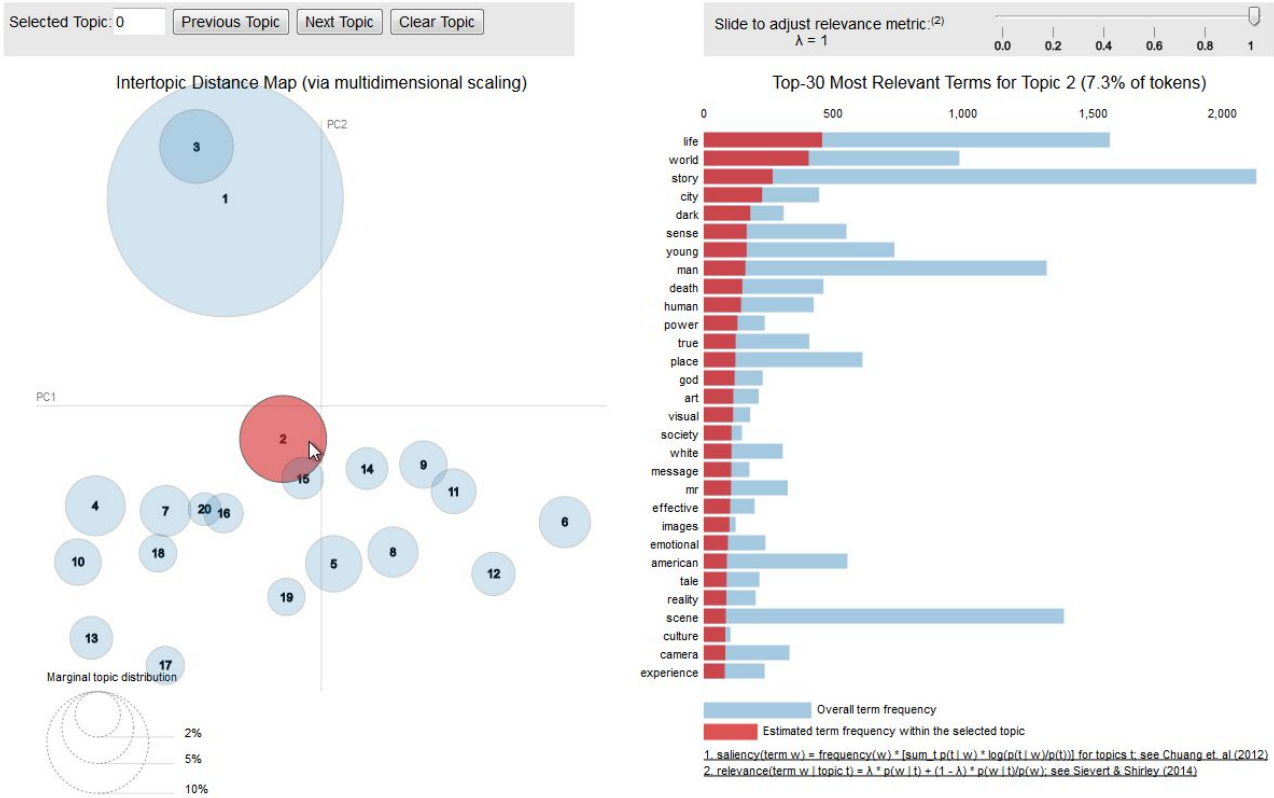
Visualization - Topic Insight

- gives an insight to topic flow and development
- hardly any information about the topic itself is shown
- pyLDAvis
  - designed to help users interpret the topics
  - interactive web-based visualization
  - Jupyter notebook

Visualization - Topic Insight

- Intertopic Distance Map
  - how topics relate to each other
  - close distance / intersection = partially the same or closely related
- Information of Topic Distribution
  - 30 most salient term if no topic is selected
  - 30 most frequent terms for the topic if there is one selected
  - overall term frequency and term frequency within the topic is displayed

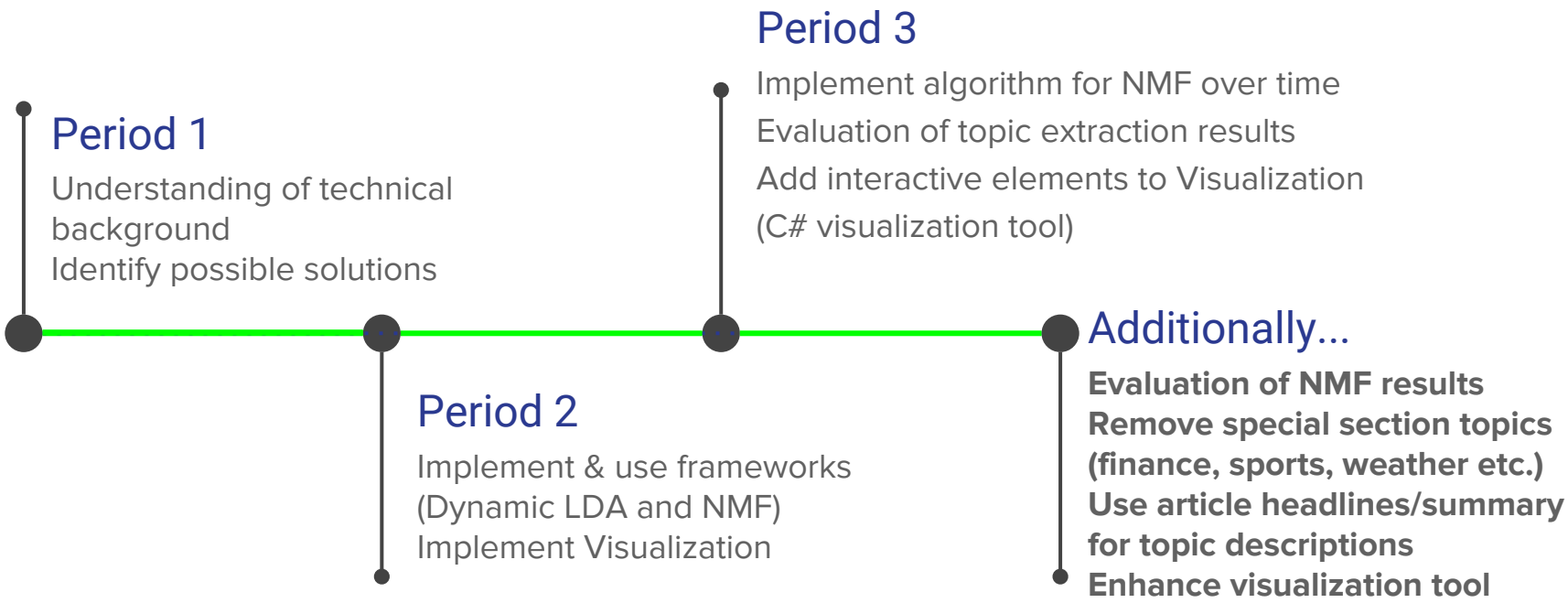
Visualization - Topics Insight



Conclusion

- Two approaches to detect evolving and dynamically changing topics: Dynamic Topic Model, and NMF over time
- Both approaches developed comprehensive topic models for the major events of the corresponding time period
- Connections between topics, relevant terms and documents can be analyzed with the visualization tool
- Topics may be investigated with the additional library pyLDAvis

Outline - Future Work



Any questions?