LINIS INTERNET STUDIES LAB

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

Topic Modeling in Online Communication Research: New Possibilities and Challenges

Sergei Koltcov, Olessia Koltsova

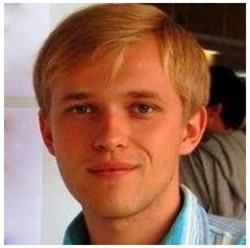
Laboratory for Internet Studies

Olessia Koltsova, sociologist

Director

Sergei Koltcov,

Physicist, IT director



Sergei Nikolenko,

Mathematician, Senior researcher

Vladimir Pyrlik,

Economist, senior researcher

Svetlana Bodrunova,

Media scholar, senior researcher

Vladimir Filippov,

Software developer

Svetlana Alexeyeva,

Computational linguist, junior researcher

Galina Selivanova,

PhD student in political science, research intern

Nora Kirkizh,

MA student in sociology, research intern

Yury Rykov,

PhD student in sociology, research intern

Outline of presentation

- Topic modeling: what is it?
- Approaches in topic modeling: General view;
- Results of simulation: Word – Topics and document – Topics distributions;
- Problem of stochastic matrix decomposition;
- Evaluation of topic model sparsity
- Problem of topic model stability
- Ways of stabilization of topic model.
- Future work.

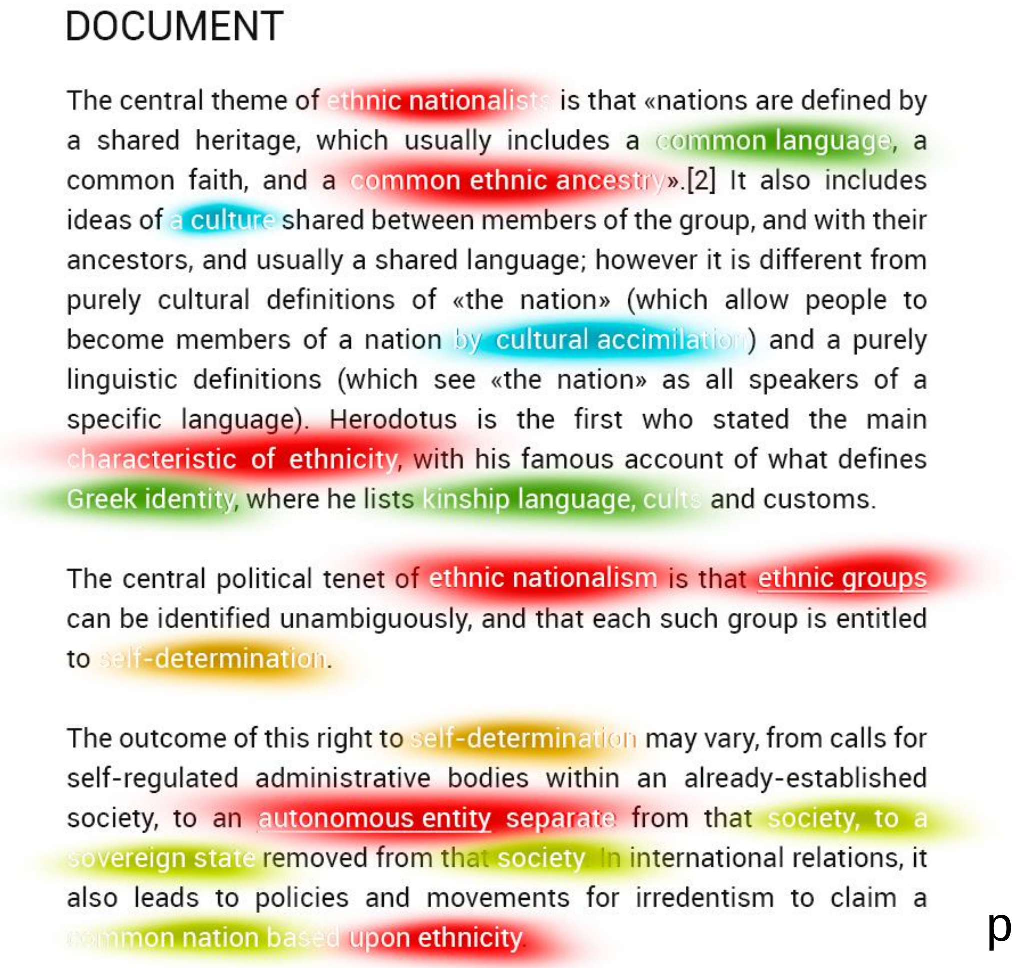
Topic modeling: description and challenges

- Topic modeling (TM) potentially can describe what topics occur in a large text collection, how big they are and how they are distributed over individual texts.

- BUT:

- TM is unstable: different runs yield different results;
- TM describes well only a minority of texts;
- It works poorly on short texts;
- Quality metrics for TM are underdeveloped because of lack of ground truth;
- As a result: hard to choose between solutions, e.g. with different topic numbers and other parameters.

Topic Modeling – what is it?



DOCUMENT

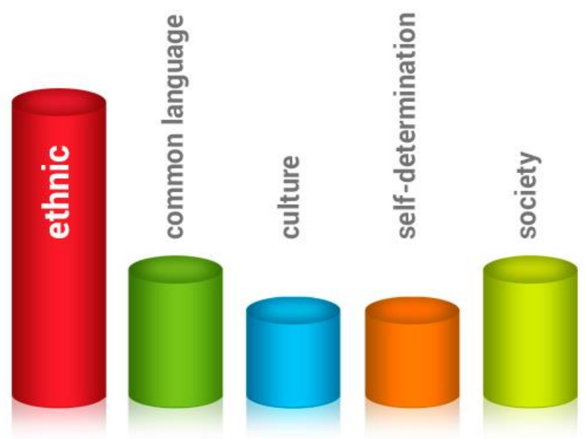
The central theme of ethnic nationalists is that «nations are defined by a shared heritage, which usually includes a common language, a common faith, and a common ethnic ancestry».[2] It also includes ideas of a culture shared between members of the group, and with their ancestors, and usually a shared language; however it is different from purely cultural definitions of «the nation» (which allow people to become members of a nation by cultural accimilation) and a purely linguistic definitions (which see «the nation» as all speakers of a specific language). Herodotus is the first who stated the main characteristic of ethnicity, with his famous account of what defines Greek identity, where he lists kinship language, cults and customs.

The central political tenet of ethnic nationalism is that ethnic groups can be identified unambiguously, and that each such group is entitled to self-determination.

The outcome of this right to self-determination may vary, from calls for self-regulated administrative bodies within an already-established society, to an autonomous entity separate from that society, to a sovereign state removed from that society. In international relations, it also leads to policies and movements for irredentism to claim a common nation based upon ethnicity.

D is a collection of documents
W is the set of all words in all documents.
T is a finite set of topics in dataset
Topic modeling is a procedure, where hidden distributions, presented by matrixes Φ_{wt} and θ_{td} are restored during simulation.

Topic distribution in $\boldsymbol{\Phi_{wt}}$



$$p(w|d) = \sum_{t=1}^{T} p(w|t)p(t|d) = \sum_{t=1}^{T} \Phi_{wt}\theta_{td}$$

Approaches in topic modeling: General view

1. Probabilistic latent semantic analysis (pLSA). Based on the idea that reconstructing Φ_{wt} and θ_{td} can be done from finding maximum of total log-likelihood:

$$L(\Phi,\theta)=\sum\nolimits_{d\in D,w\in d}\sum n_{wd}\,ln\sum_{t\in T}\Phi_{wt}\theta_{td}\rightarrow max$$

Procedure of maximization based on expectation-maximization (EM) algorithm under constraints: $\Phi_{wt}>1$, $\theta_{td}>1$ and $\sum_{t\in T}p(w|t)=1$.

2. Latent Dirichlet allocation (LDA): is a Bayesian version of pLSA: it assumes that multinomial distributions θ_{td} and Φ_{wt} are generated from prior Dirichlet distributions, one with parameter α (for the θ_{td} distributions) and one with parameter β (for the Φ_{wt} distributions).

Variation approximations of LDA – pure mathematician model

$$p(\theta,z,w|\alpha,\beta)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta)$$

$$\theta_{td}=\frac{n_{td}+\alpha_t}{n_d+\alpha_0}$$

$$\phi_{td}=\frac{n_{wt}+\beta_w}{n_t+\alpha_0}$$

Where n_{td} – number documents in topic t.

Where n_{wt} – number words in topic t.

INTERNET STUDIES LAB

Approaches in topic modeling: General view

3. Latent Dirichlet Allocation(Gibbs sampling) – based on idea from physics (Potts model)

$$P(z_i = j \mid w_i = m, z_{-i}, w_{-i}) \approx \frac{C^{WT}_{m,j} + \beta}{\sum_{m'} C^{WT}_{m',j} + V\beta} \cdot \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + \alpha T}$$

$C^{WT}_{m,j}$ - Matrix; cells: number of times a word was assigned to topic t,

$C^{DT}_{d,j}$ - Matrix; cells: number of times a word in document d is assigned to topic t.

$\sum_{m'} C^{WT}_{m',j} = n_t$ - Vector; cells: number of words assigned to topic t,

$C^{DT}_{d,j'} = n_d$ - Length of document d in words

Results of simulation:

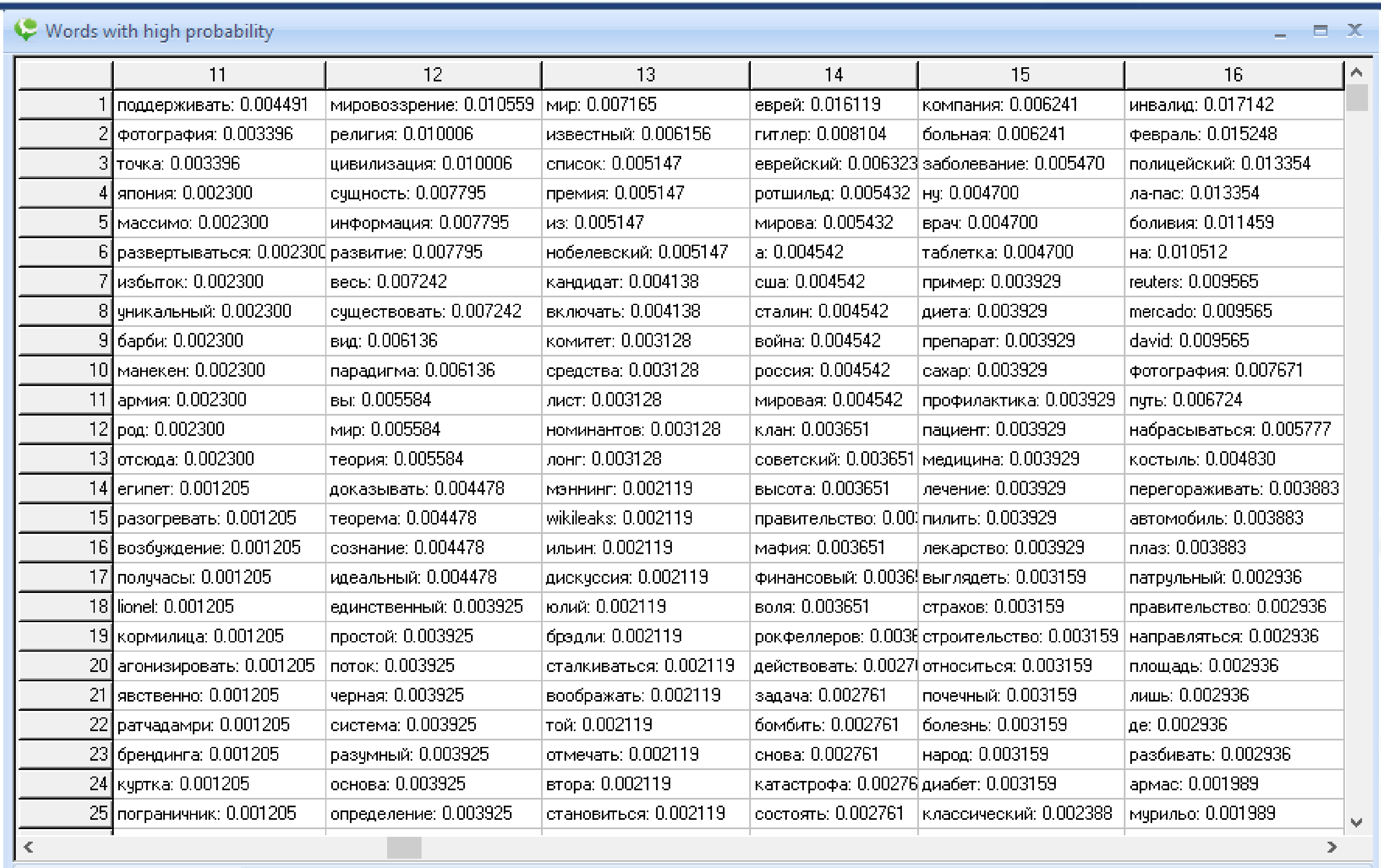
1. Matrix of words distribution in topics.

$$\theta_{dj} = \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + T\alpha}$$

2. Matrix of document distribution in topics.

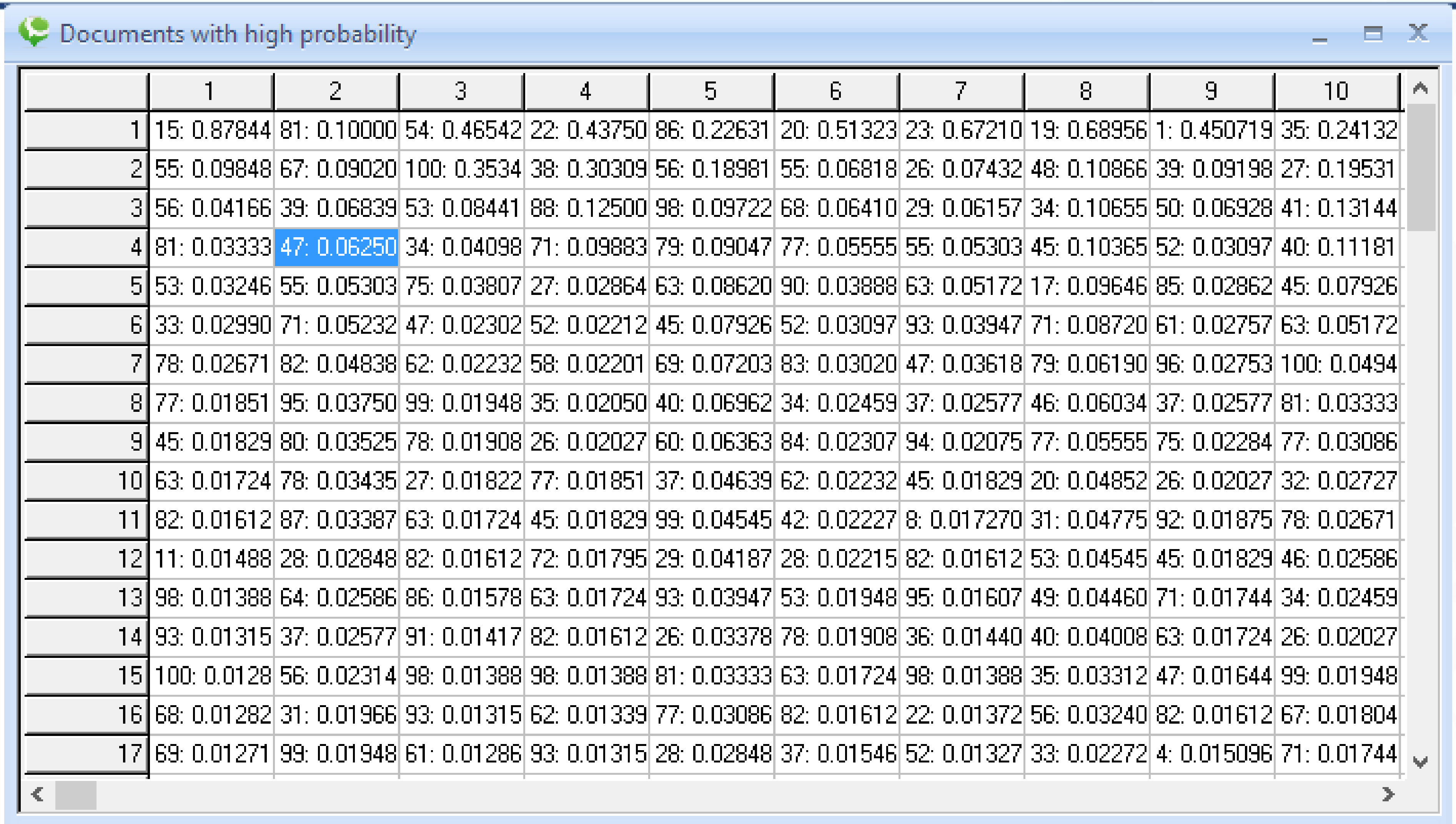
$$\phi_{m,j} = \frac{C^{WT}_{m,j} + \beta}{\sum_{m'} C^{WT}_{m',j} + V\beta}$$

Results of simulation: Word - topic distribution (Matrix Φ)



	11	12	13	14	15	16
1	поддерживать: 0.004491	мировоззрение: 0.010559	мир: 0.007165	еврей: 0.016119	компания: 0.006241	инвалид: 0.017142
2	фотография: 0.003396	религия: 0.010006	известный: 0.006156	гитлер: 0.008104	больная: 0.006241	февраль: 0.015248
3	точка: 0.003396	цивилизация: 0.010006	список: 0.005147	еврейский: 0.006323	заболевание: 0.005470	полицейский: 0.013354
4	япония: 0.002300	сущность: 0.007795	премия: 0.005147	ротшильд: 0.005432	ну: 0.004700	ла-пас: 0.013354
5	массимо: 0.002300	информация: 0.007795	из: 0.005147	мирова: 0.005432	врач: 0.004700	боливия: 0.011459
6	развертываться: 0.00230	развитие: 0.007795	нобелевский: 0.005147	а: 0.004542	таблетка: 0.004700	на: 0.010512
7	избыток: 0.002300	весь: 0.007242	кандидат: 0.004138	сша: 0.004542	пример: 0.003929	reuters: 0.009565
8	уникальный: 0.002300	существовать: 0.007242	включать: 0.004138	сталин: 0.004542	диета: 0.003929	mercado: 0.009565
9	барби: 0.002300	вид: 0.006136	комитет: 0.003128	война: 0.004542	препарат: 0.003929	david: 0.009565
10	манекен: 0.002300	парадигма: 0.006136	средства: 0.003128	россия: 0.004542	сахар: 0.003929	фотография: 0.007671
11	армия: 0.002300	вы: 0.005584	лист: 0.003128	мировая: 0.004542	профилактика: 0.003929	путь: 0.006724
12	род: 0.002300	мир: 0.005584	номинантов: 0.003128	клан: 0.003651	пациент: 0.003929	набрасываться: 0.005777
13	отсюда: 0.002300	теория: 0.005584	лонг: 0.003128	советский: 0.003651	медицина: 0.003929	костыль: 0.004830
14	египет: 0.001205	доказывать: 0.004478	мэннинг: 0.002119	высота: 0.003651	лечение: 0.003929	перегораживать: 0.003883
15	разогревать: 0.001205	теорема: 0.004478	wikileaks: 0.002119	правительство: 0.00	пилить: 0.003929	автомобиль: 0.003883
16	возбуждение: 0.001205	сознание: 0.004478	ильин: 0.002119	мафия: 0.003651	лекарство: 0.003929	плаз: 0.003883
17	получасы: 0.001205	идеальный: 0.004478	дискуссия: 0.002119	финансовый: 0.0036	выглядеть: 0.003159	патрульный: 0.002936
18	lionel: 0.001205	единственный: 0.003925	юлий: 0.002119	воля: 0.003651	страхов: 0.003159	правительство: 0.002936
19	кормилица: 0.001205	простой: 0.003925	брэдли: 0.002119	рокфеллеров: 0.003	строительство: 0.003159	направляться: 0.002936
20	агонизировать: 0.001205	поток: 0.003925	сталкиваться: 0.002119	действовать: 0.0027	относиться: 0.003159	площадь: 0.002936
21	явственно: 0.001205	черная: 0.003925	воображать: 0.002119	задача: 0.002761	почечный: 0.003159	лишь: 0.002936
22	ратчадамри: 0.001205	система: 0.003925	той: 0.002119	бомбить: 0.002761	болезнь: 0.003159	де: 0.002936
23	брендинга: 0.001205	разумный: 0.003925	отмечать: 0.002119	снова: 0.002761	народ: 0.003159	разбивать: 0.002936
24	куртка: 0.001205	основа: 0.003925	втора: 0.002119	катастрофа: 0.00276	диабет: 0.003159	армас: 0.001989
25	пограничник: 0.001205	определение: 0.003925	становиться: 0.002119	состоять: 0.002761	классический: 0.002388	мурильо: 0.001989

Results of simulation: Document - topic distribution (Matrix Θ)



	1	2	3	4	5	6	7	8	9	10
1	15: 0.87844	81: 0.10000	54: 0.46542	22: 0.43750	86: 0.22631	20: 0.51323	23: 0.67210	19: 0.68956	1: 0.450719	35: 0.24132
2	55: 0.09848	67: 0.09020	100: 0.3534	38: 0.30309	56: 0.18981	55: 0.06818	26: 0.07432	48: 0.10866	39: 0.09198	27: 0.19531
3	56: 0.04166	39: 0.06839	53: 0.08441	88: 0.12500	98: 0.09722	68: 0.06410	29: 0.06157	34: 0.10655	50: 0.06928	41: 0.13144
4	81: 0.03333	47: 0.06250	34: 0.04098	71: 0.09883	79: 0.09047	77: 0.05555	55: 0.05303	45: 0.10365	52: 0.03097	40: 0.11181
5	53: 0.03246	55: 0.05303	75: 0.03807	27: 0.02864	63: 0.08620	90: 0.03888	63: 0.05172	17: 0.09646	85: 0.02862	45: 0.07926
6	33: 0.02990	71: 0.05232	47: 0.02302	52: 0.02212	45: 0.07926	52: 0.03097	93: 0.03947	71: 0.08720	61: 0.02757	63: 0.05172
7	78: 0.02671	82: 0.04838	62: 0.02232	58: 0.02201	69: 0.07203	83: 0.03020	47: 0.03618	79: 0.06190	96: 0.02753	100: 0.0494
8	77: 0.01851	95: 0.03750	99: 0.01948	35: 0.02050	40: 0.06962	34: 0.02459	37: 0.02577	46: 0.06034	37: 0.02577	81: 0.03333
9	45: 0.01829	80: 0.03525	78: 0.01908	26: 0.02027	60: 0.06363	84: 0.02307	94: 0.02075	77: 0.05555	75: 0.02284	77: 0.03086
10	63: 0.01724	78: 0.03435	27: 0.01822	77: 0.01851	37: 0.04639	62: 0.02232	45: 0.01829	20: 0.04852	26: 0.02027	32: 0.02727
11	82: 0.01612	87: 0.03387	63: 0.01724	45: 0.01829	99: 0.04545	42: 0.02227	8: 0.017270	31: 0.04775	92: 0.01875	78: 0.02671
12	11: 0.01488	28: 0.02848	82: 0.01612	72: 0.01795	29: 0.04187	28: 0.02215	82: 0.01612	53: 0.04545	45: 0.01829	46: 0.02586
13	98: 0.01388	64: 0.02586	86: 0.01578	63: 0.01724	93: 0.03947	53: 0.01948	95: 0.01607	49: 0.04460	71: 0.01744	34: 0.02459
14	93: 0.01315	37: 0.02577	91: 0.01417	82: 0.01612	26: 0.03378	78: 0.01908	36: 0.01440	40: 0.04008	63: 0.01724	26: 0.02027
15	100: 0.0128	56: 0.02314	98: 0.01388	98: 0.01388	81: 0.03333	63: 0.01724	98: 0.01388	35: 0.03312	47: 0.01644	99: 0.01948
16	68: 0.01282	31: 0.01966	93: 0.01315	62: 0.01339	77: 0.03086	82: 0.01612	22: 0.01372	56: 0.03240	82: 0.01612	67: 0.01804
17	69: 0.01271	99: 0.01948	61: 0.01286	93: 0.01315	28: 0.02848	37: 0.01546	52: 0.01327	33: 0.02272	4: 0.015096	71: 0.01744

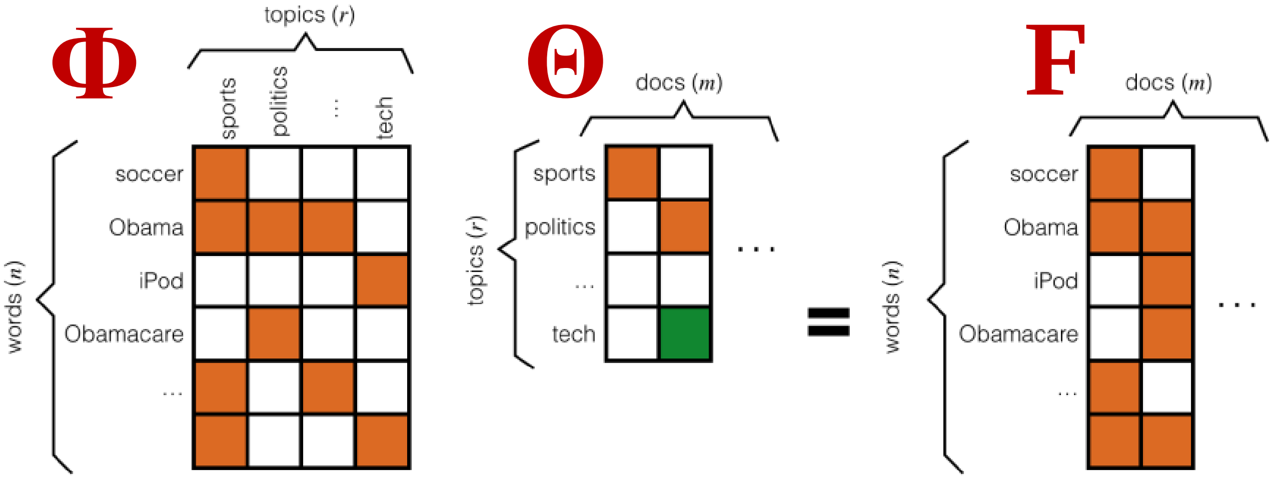
Stochastic matrix decomposition

$$F[documents \times words] = \Theta[documents \times topics] \cdot \Phi[topics \times words]$$

Matrix F represents a dataset. Our dataset can be expressed in terms of two low dimension matrices. Process of sampling is the process of approximation of matrix F by two matrices Φ and Θ. **But**:

$$F = \Theta \cdot \Phi = (\Theta \cdot R) \cdot (R^{-1}\Phi) = \Theta' \cdot \Phi'$$

Matrix F can be approximated by different combinations of matrices (but with the same dimensions



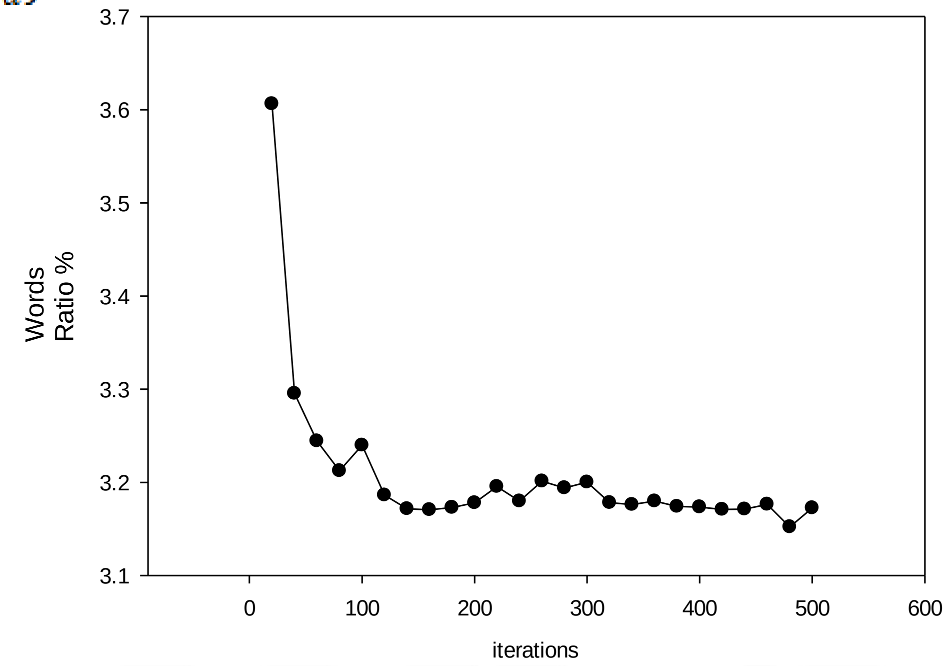
Evaluation of sparsity of topic models

LDA inference algorithm guarantees that the iterative process converges to a certain value of **perplexity** with some noise, which means that the number of words and documents used in modeling also converge to a certain value. Actually perplexity is the inverse of the geometric mean per-word likelihood.

$$peplexity = \exp\{\sum_{d,k=1}^{M,K} p(w)_{dk}) \,/\, \sum_{d=1}^{M} N_d\}$$

Words ratio

Word ratio as the parameter that characterizes the ratio of the total number of words with probability greater than **1/V** over all documents, where **V** is dictionary length.



Evaluating LDA stability with Kullback – Leibler divergence and topic stability

The Kullback - Leibler divergence is a widely accepted distance measure between two probability distributions. It can be calculated according to the following formula.

$$K = 0.5 \sum_{k}^{W} \Phi_k^1 \log\left(\frac{\Phi_k^1}{\Phi_k^2}\right) + 0.5 \sum_{k}^{W} \Phi_k^2 \log\left(\frac{\Phi_k^2}{\Phi_k^1}\right)$$

IF K=0, then two topics are identical. IF K=Max value then the value shows dissimilarity of topics.

However, directly computing KL divergence to measure similarity between two topics in a topic modeling result does not lead to a good result since the KL value is dominated by the long tail of low probability words that do not define the topic in any qualitative way and are mostly random.

KL-based similarity metric

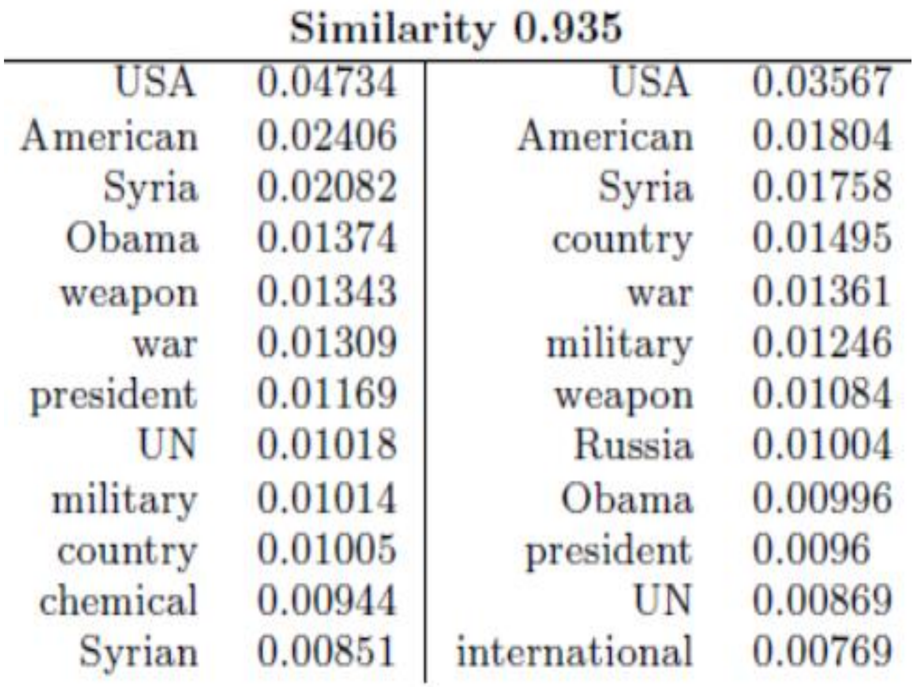
$$Kn = \left(1 - \frac{K}{Max}\right) * 100$$

IF Kn=100%, then two topics are identical. IF K=0 then that topics are totally different.

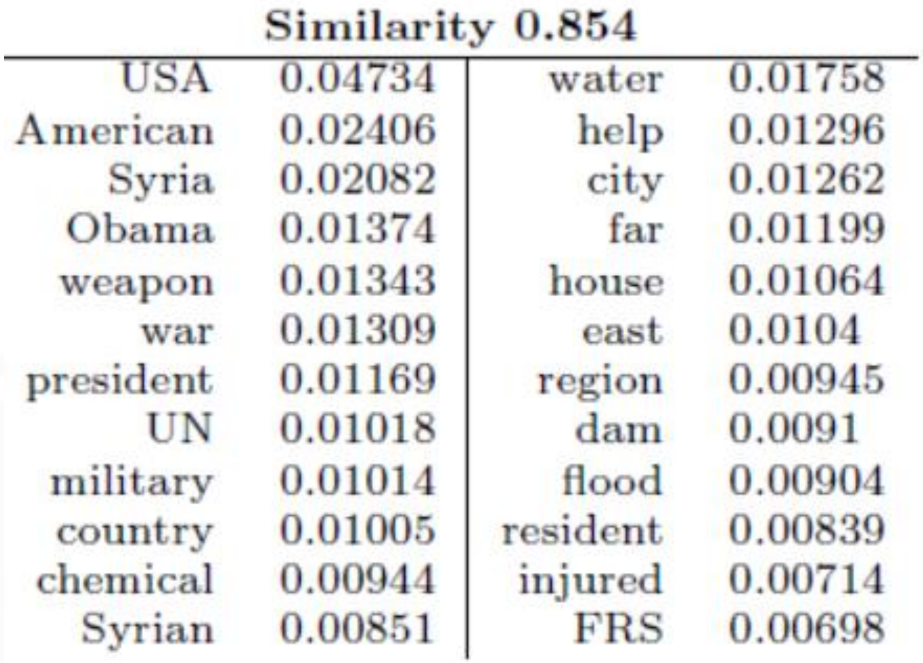
STABILITY EVALUATION

Level 90 - 93% (and more) means that first 50 words are almost identical.

Level about 85%: topics are completely different.



Similarity 0.935			
USA	0.04734	USA	0.03567
American	0.02406	American	0.01804
Syria	0.02082	Syria	0.01758
Obama	0.01374	country	0.01495
weapon	0.01343	war	0.01361
war	0.01309	military	0.01246
president	0.01169	weapon	0.01084
UN	0.01018	Russia	0.01004
military	0.01014	Obama	0.00996
country	0.01005	president	0.0096
chemical	0.00944	UN	0.00869
Syrian	0.00851	international	0.00769

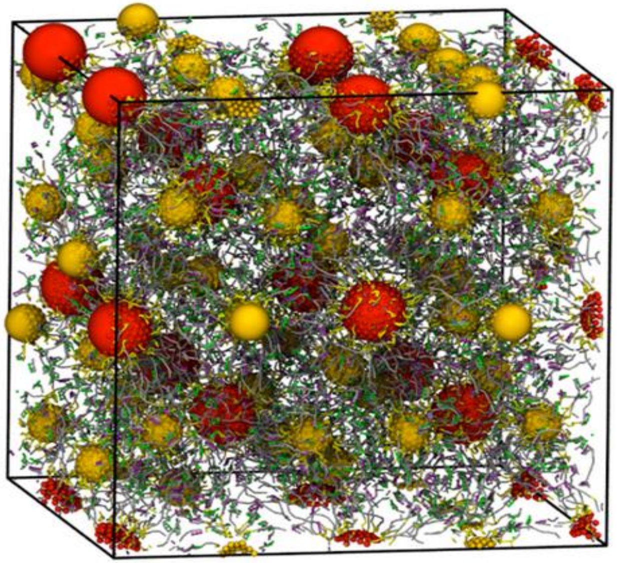


Similarity 0.854			
USA	0.04734	water	0.01758
American	0.02406	help	0.01296
Syria	0.02082	city	0.01262
Obama	0.01374	far	0.01199
weapon	0.01343	house	0.01064
war	0.01309	east	0.0104
president	0.01169	region	0.00945
UN	0.01018	dam	0.0091
military	0.01014	flood	0.00904
country	0.01005	resident	0.00839
chemical	0.00944	injured	0.00714
Syrian	0.00851	FRS	0.00698

Regularization of topic models

In optimization theory, problems with unstable solutions are called ill-posed, and a general approach to solving these problems is given by Tikhonov regularization [38]. In terms of the model definition, regularization can be viewed as extending the prior information which lets one reduce the set of solutions. Regularization is done either by introducing constraints on $\theta_{d,t}$ and $\phi_{m,t}$ matrices or by changing procedure o sampling.

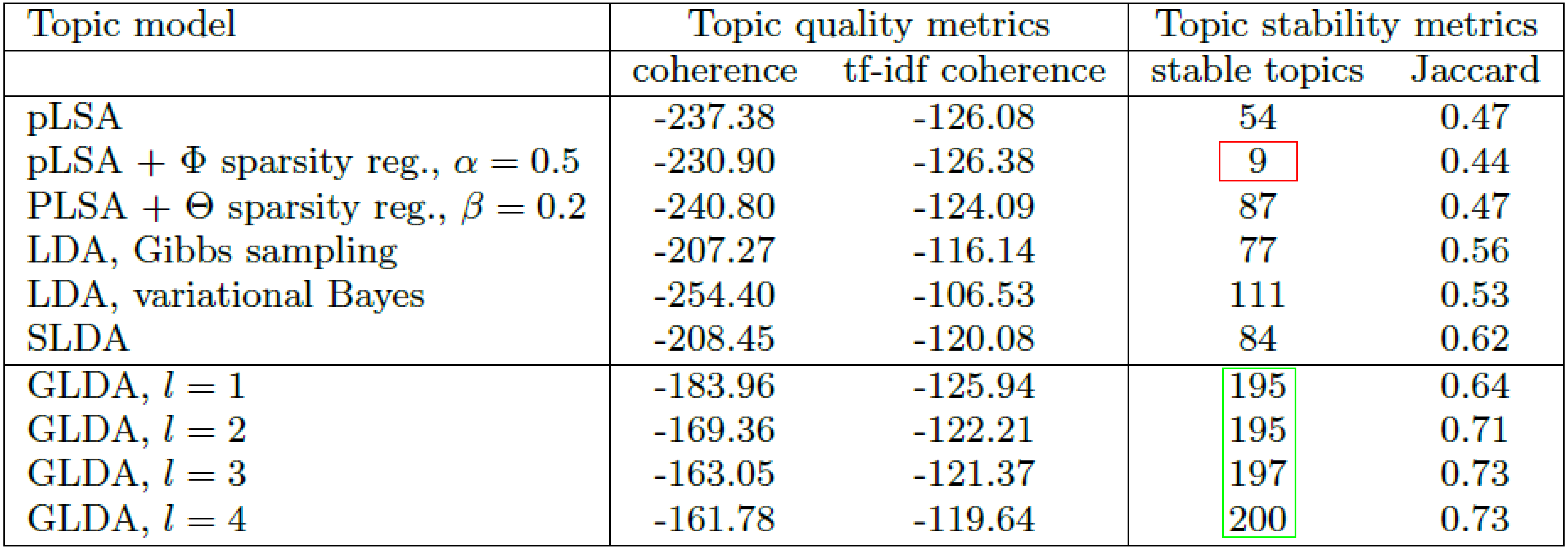
Example of regularization: Semi-Supervised Latent Dirichlet Allocation (Gibbs sampling)



If we have initial distribution of words (anchor words) over topics, then we are able to fix or glue words to topics. Therefore, when the algorithm faces an anchor word during sampling, it does not change the connection between the topic and the word. But the other words are sampled according to the standard procedure.

The SLDA modeling behaves as a process of crystallization, where anchor words are centers of crystals. The words that often co-occur with anchor words stick together during simulation and form the body of topics.

Regularization of topic models: Results



Topic model	Topic quality metrics		Topic stability metrics	
	coherence	tf-idf coherence	stable topics	Jaccard
pLSA	-237.38	-126.08	54	0.47
pLSA + Φ sparsity reg., $\alpha = 0.5$	-230.90	-126.38	9	0.44
PLSA + Θ sparsity reg., $\beta = 0.2$	-240.80	-124.09	87	0.47
LDA, Gibbs sampling	-207.27	-116.14	77	0.56
LDA, variational Bayes	-254.40	-106.53	111	0.53
SLDA	-208.45	-120.08	84	0.62
GLDA, $l = 1$	-183.96	-125.94	195	0.64
GLDA, $l = 2$	-169.36	-122.21	195	0.71
GLDA, $l = 3$	-163.05	-121.37	197	0.73
GLDA, $l = 4$	-161.78	-119.64	200	0.73

RESULT: (1) regularization can significantly improve stability, for example GLDA, but (2) regularization can almost kill stability, for example pLSA with Φ regularization.

(LDA model can be regarded as regularized version of pLSA, where regularization is adding information that distributions are Dirichlet functions).

CONCLUSION AND FURTHER WORK

- TM is convenient for big data.
- But it has shortcomings -> can be overcome with regularization.
- Some regularizations may decrease model quality.
- Improvement is important for web science and digital humanities that seek not only interpretable topics, but entire solutions to make reliable conclusions about the topical structure of text collections.

- Therefore, the problem of topic number is one of the central.
- Needed: analysis of topic models' behavior as a function of topic number.
- Probably based on physical approaches from condensed physics.

ACKNOWLEDGEMENTS

This work was supported by the HSE Academic Fund Program and Basic Research Program of the National Research University Higher School of Economics, in 2015 - 2016.

HIGHER • SCHOOL • OF • ECONOMICS

NATIONAL RESEARCH
UNIVERSITY

Thank you
for your attention!

Room 216, building 2, 55 Sedova St., St.Petersburg, Russia
Laboratory for Internet Studies
www.linis.hse.ru