International Conference on Computational Social Science 2015

Topic Modeling Stability and Granulated LDA

Sergei Koltcov, Olessia Koltsova, Sergey Nikolenko.
http://linis.hse.ru/en/

WHY TOPIC MODELING STABILITY IS IMPORTANT FOR SOCIAL SCIENCE?

•Large text corpora (e.g. user generated content) are increasingly becoming an object of social science research.

•Social scientists assume that collections of texts are devoted to a number of topics in a certain proportion that reflect real interest of the text authors to those topics.

•**Social scientists also expect that mathematicians can provide them an instrument** that can detect these topics and their proportions.

•If an instrument provides "false" – e.g. biased or unstable – results, social scientists can not mine topics and draw any meaningful conclusions.

RESEARCH GOAL

In this work, we propose GLDA - a new, more stable modification of LDA, the most well-known topic modeling algorithm.

In this presentation:

- We explain the idea of GLDA based on word co-occurance regularization.
- We report experimental results with three algorithms including GLDA.
- We explain a new metrics of stability - a modified version of Kullback – Leibler divergence that was used in the experiments along with traditional Jaccard coefficient.

Dataset

101,481 posts from the Russian LiveJournal.
172,939,000 tokens (unique words).

LDA Joint Probability

Terms: **D** – space of documents, **W** – space of unique words, **Z** – space of topics. Topics are hidden parameters, which have to be found in simulation. **Joint Probability** can be calculated according to the following equations:

$$p(w \mid z, \beta) = \int p(w \mid z, \Phi)\, p(\Phi \mid \beta)\, d\Phi \qquad p(z \mid \alpha) = \int p(z \mid \Theta)\, p(\Theta \mid \alpha)\, d\Theta$$

Θ, Φ : words – topics and document – topics matrices. Those can be found out by different algorithms.

Mean field variational inference

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max$$

Collapsed Gibbs sampling

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

In terms of matrix factorization LDA solution is:

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

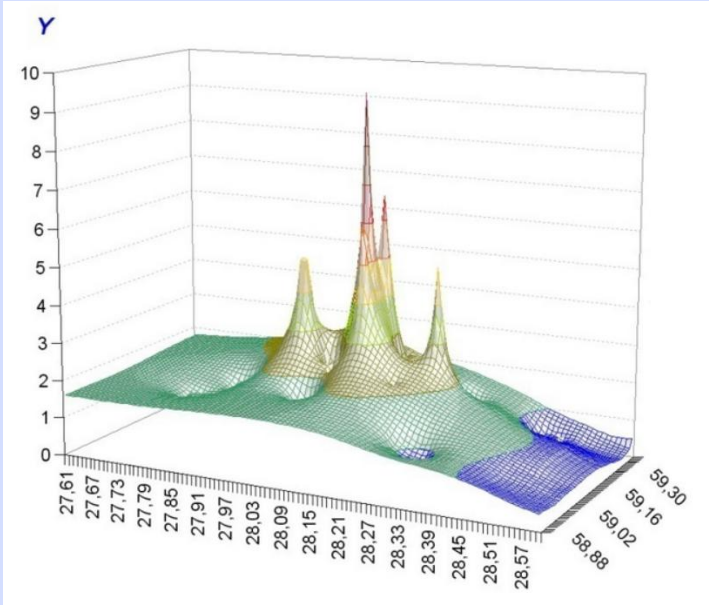
Mathematical vision of LDA

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

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 which means LDA has many solutions, e.i. many local minima..



Partly, the problem can be solved by using method of regularization whose main idea is in using additional information.

Regularization in topic modeling is a procedure that introduces reasonable restrictions on matrices.

Latent Dirichlet Allocation (Gibbs sampling)

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

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

$C^{DT}_{d,j}$ - Matrix; cells: number of times a word **w** 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 on topics. 2. Matrix of documents distribution on topics.

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

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

LDA is a pLSA model with coefficients of regularization (α,β).

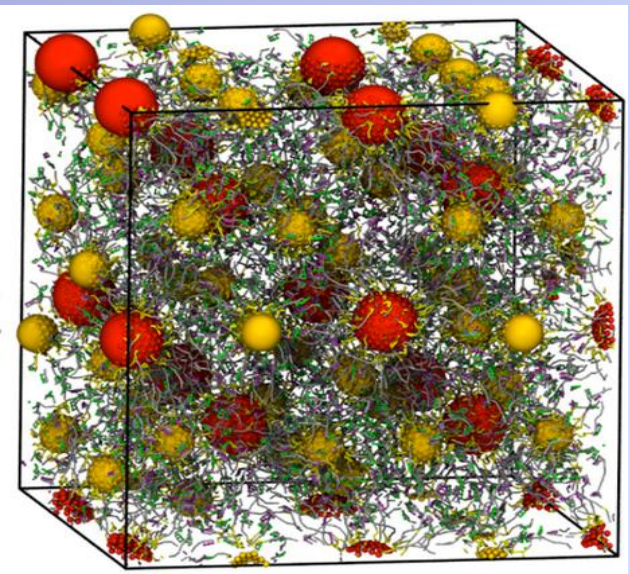
Semi-Supervised Latent Dirichlet Allocation (Gibbs sampling)

Next level of regularization is based on the following idea. 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.

$$p(\mathrm{z}, \mathrm{w}, \alpha, \beta) \propto \begin{cases} z = t \\ q(\mathrm{z}, \mathrm{w}, \alpha, \beta) \end{cases}$$

→ Initial anchor words distribution

→ Standard Gibbs sampling



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.

Therefore the fixed lists of anchor words assigned to topics lead to stabilization of topic modeling.

But SLDA works good if the list of words is known beforehand which is not always the case in social science.

GRANULATED LDA (based on Gibbs sampling)

Granulated LDA based on idea that each document from a collection can be regarded as a granulated surface – a notion we borrow from physics. Here a granule is a set of words located near each other. Therefore we can arrange sampling by granules.

All words within one granule belong to one topic. Therefore scanning documents and assigning granules to topics, the algorithm favors the words located near each other.

DOCUMENT

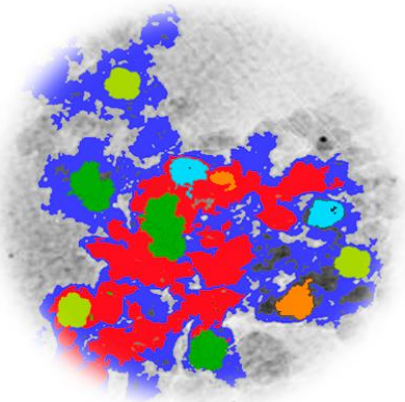
The central theme of ethnic nationalism is that «nations are defined by a shared heritage, which usually includes a common language, a common faith, and a common ethnic ancestry».[2] It also includes ideas of a culture shared between members of the group, and with their ancestors, and usually a shared 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.

KEYWORDS

ethnic	7
common language	3
culture	2
self-determination	2
society	3



granulated surface

GRANULATED LDA: Algorithm

Entrance: collection of documents D, number of topics |T|,
number of iterations, size of granules L;
Initialization: φ(w,t), θ(t,d) for all documents and topics d ∈ D, w ∈ W , t ∈ T;

Run external cycle along all documents (i)
Run internal cycle. Length of cycle is number words in documents i.
1. Generation of random number k. Max value of k is number of words in documents i.
2.Choosing word **k** from document **i**.
3. Calculating topic number **t** for word **k**.
4. Defining words that are around word k in document i.
5. Assigning topic **t** to all words which are within granule L.
End of internal cycle.
Updating the following matrices: $C^{WT}_{m,j}$ $C^{DT}_{d,j}$ $\sum_{m'} C^{WT}_{m',j} = n_t$
End of external cycle
Calculation of matrixes φ(w,t), θ(t,d) based on $\theta_{dj} = \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + T\alpha}$ $\phi_{m,j} = \frac{C^{WT}_{m,j} + \beta}{\sum_{m'} C^{WT}_{m',j} + V\beta}$

Visualization φ(w,t), θ(t,d).

Evaluating LDA quality with Kullback–Leibler divergence and Jaccard coefficient

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

$$Kn = (1 - \frac{K}{Max}) \cdot 100\% \quad \text{where} \quad Kn = 0.5 \sum_{k=1}^{W} \Omega_k^1 \log(\frac{\Omega_k^1}{\Omega_k^2}) + 0.5 \sum_{k=1}^{W} \Omega_k^2 \log(\frac{\Omega_k^2}{\Omega_k^1})$$

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

Jaccard coefficient: Jc=a/(a+b-k).

where **a** – one topic, **b** – another topic, **k** – number of words, which are identical in both topics.

Jc = 1, if two topic are identically, if **Jc = 0** then topic completely different.

Topic similarity thresholds

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

Words in topics are ordered according to probability.

Results of simulations

Dataset

1. 101481 posts from the Russian LiveJournal.
2. 172939 000 tokens (unique words).

Number of runs: **5 times**; number of topics: 200

Topic model (Gibbs sampling modification)	Number of stable topics According to Kullback-Leibler	Kullback–Leibler Divergence threshold (dictionary length= 1000 words)	Jaccard coefficient (100 most probable words in topic)
LDA	84	≈90%	≈0.37
SLDA	135	≈92.43%	≈0.16
GLDA (window ±1)	138	≈96%	≈0.6

CONCLUSION

1. We have proposed Granulated LDA as a regularized version of LDA.

2. For the purpose of detecting LDA stability, we have proposed a new topic similarity measure based on Kullback-Leibler divergence.

3. Based on this measure we have investigated stability of 3 modifications of LDA and have shown that Granulated LDA is more stable then other modifications.

4. Therefore, to be able to draw reliable sociological conclusions we recommend researchers either to run topic modeling several times, then extract stable topics that reappear across multiple runs or use Granulated LDA.

**The research is funded by the Basic Research Program
Of National Research University
Higher School of Economics**

INTERNET STUDIES LAB

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
SAINT PETERSBURG

THANK YOU!