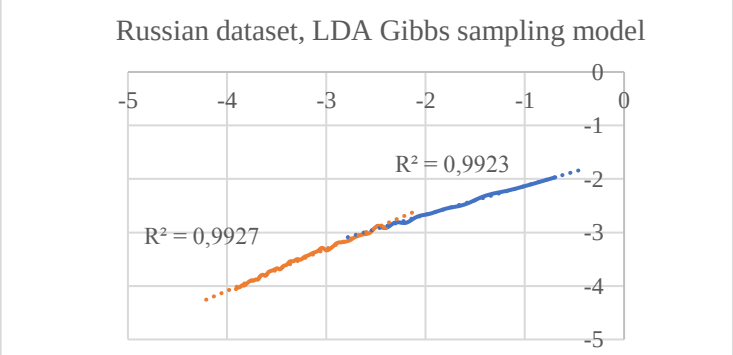
Fractal Approach for Determining the Optimal Number of Topics.

Vera Ignatenko[1], Sergei Koltcov[1], Zeyd Boukhers[2] and Steffen Staab[2]

[1]National Research University Higher School of Economics, St. Petersburg, Russia

[2]Institute for Web Science and Technologies, University of Koblenz-Landau, Koblenz, Germany

Modern information systems generate a huge number of texts such as news, blogs and comments. Analysis of big data is impossible without the construction of formalized mathematical models based on statistical physics. One of such a model is topic modelling (TM) based on Potts model [1]. This model assumes that each textual document can be considered as a one-dimensional grid and each word of a document as a node. A node can be in one of T states. Correspondingly, a collection of words (nodes) referring to one of the states can be considered a topic. The probability of a word in a document is described by the following expression: $p(w|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \phi_{wt}\theta_{td}$, where $p(w|t) = \phi_{wt}$ is the distribution of words by topics, $p(t|d) = \theta_{td}$ is the distribution of documents by topics [1], T is the number of topics, W is the number of unique words. The main parameter of TM algorithm is the number of topics, which is manually defined. The study of topic model behavior as a function of topic number is extremely actual and can be realized using fractal formalism. Topic solution under fixed number of topics is a matrix ϕ_{wt}, where the number of cells is T*W, T is the number of columns of the matrix, W is the number of rows. The size of each cell is $\varepsilon = 1/(WT)$. Each cell of the matrix contains the probability P_{ij} of belonging of a word w_i to a topic T_j. The probability density function has the following form [2]: $\rho_i = \frac{n_i}{WT}$, n_i is the number of cells in topic solution containing high probabilities. This value is a function of a topic number, and it varies from 1 to some number $\rho_i(\varepsilon) < 1$ in the process of topic modelling. The density $\rho_i(\varepsilon)$ depends on the size of cells and degree $D(\varepsilon)$: $\rho(\varepsilon) \cong \varepsilon^{-D(\varepsilon)}$. Parameters of textual collections used in this work: 1. Dataset in Russian language: 18026 unique words, 8630 documents. 2. Dataset in English language: 50948 unique words, 15404 documents. The choice of data sets depends on the availability of the number of topics embedded in texts. In this work the following topic models were used: 1. PLSA, 2. LDA Gibbs sampling and 3. BigARTM. The figure below shows an example of TM fractal behavior of the dataset in Russian. The intersection of trend lines corresponds to a point, equal to the number of topics in the given dataset and corresponds to a point of information phase transition.



Fractal analysis of topic model behaviour allows us to show that self-similar fractal clusters exist in large textual collections. The forming of clusters occurs precisely in the transition regions. Linear regions do not lead to changes in fractals, therefore, it is sufficient to find transition regions for the study of textual collections. Accordingly, the problem of the analysing the evolution of topic models can be simplified to the problem of searching transition regions in topic models.

1. Griffiths T, Steyvers M. Finding Scientific Topics // Proceedings of the National Academy of Sciences. 2004. Vol. 101 (Suppl. 1). P. 5228–5335.
2. Koltcov S. N., A thermodynamic approach to selecting a number of clusters based on topic modeling, Technical Physics Letters, 43(6), P. 584-586.