

Topic Modeling without Generative Probabilistic Model: An Approach and its Validation *

Mikhail G. Kreines, Elena M. Kreines

BaseTech LLC, Raketny boulevard, 11-7-164, Moscow, 129366, Russia
mkrf@yandex.ru

Abstract. We propose a novel approach to compute a model of semantic structure of natural language texts corpora based on the models of the texts. This approach differs from LSI and standard ways of topic modeling (D. Blei, 2012). We do not use matrix factorization and probabilistic distributions of the words given a priory. As the base for the corpora model construction we use the set of the original vector models of the corpora texts. The ways of empirical validation and applications of the models are considered.

Keywords: natural languages, texts, text collections, models, Information Retrieval

* The work is partially financially supported by the Ministry of Education and Science, Russian Federation (Contract N. 14.579.21.0090 from 27.11.2014, project identifier RFMEFI57914X0090).