Kreĭnes, M. G.; **Kreĭnes, Elena M.**

Matrix text models. Text corpora models. (Russian. English summary) Zbl 1444.68267
Mat. Model. 32, No. 2, 37-57 (2020).

Summary: The models of text corpora, formed on the basis of the matrix model of texts in natural languages, are presented. As methods to form models of collections we consider the techniques of computational identification of the thematic structure of the collections. We suggest to use the models for searching for thematically similar text collections and thematic categorization of texts based on text models and text collections. The differences of the proposed models of text collections from the common approaches to their analysis and modeling are analyzed.

MSC:

68T50 Natural language processing
68P20 Information storage and retrieval of data
91F20 Linguistics

Keywords:

natural language texts; text corpora; text corpora models; topic models; text models; text information retrieval

Software:

word2vec

Full Text: DOI MNR

References:

[1] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, Addison-Wesley, Boston, 2010, 542 pp.

[2] W. Wu, H. Xiong, Sh. Shekhar, Clustering and Information Retrieval, Network Theory, Applications, 11, Springer, N. Y., 2004, 338 pp.

[3] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, N. R. Shadbolt, "Automatic ontology-based knowledge extraction from Web documents", IEEE Intelligent Systems, 18:1 (2003), 14-21

[4] N. V. Lukashevich, Tezaurusy v zadachah informatsionnogo poiska, MGU, M., 2011, 512 pp.

[5] T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch (eds.), Handbook of Latent Semantic Analysis, Psychology Press, Hove, 2013, 544 pp.

[6] D. M. Blei, "Probabilistic topic models", Communicat. of the ACM, 55:4 (2012), 77-84

[7] K. V. Vorontsov, "Additive Regularization for Topic Models of Text Collections", Doklady Mathemaics, 89:3 (2014), 301-304 · Zbl 1358.68242

[8] M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, "From Word Embeddings To Document Distances", Proc. of the 32nd Int. Conf. on Machine Learning (Lille, France, 2015), JMLR: W&CP, 37, 2015, 957-966

[9] M. G. Kreines, E. M. Kreines. Matrix text models, "Text models and similarity of text contents", MM&CS, 2020 · Zbl 1109.58305

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality", Advances in neural information processing systems, 2013, 3111-3119

[11] I. S. Misuno, D. A. Rachkovskii, S. V. Slipchenko, "Vektornye i raspredelennye predstavleniia, otrazhaushchie mery semanticheskoi sviazi slov", Matemathchni mashini i sistemi, 3 (2005), 50-66

[12] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, "A neural probabilistic language model", Journal of Machine Learning Research, 3 (2003), 1137-1155 · Zbl 1061.68157

[13] Q. Le, T. Mikolov, "Distributed representations of sentences and documents", Proc. of the 31-st Int. Conf. on Machine Learning (Beijing, China, 2014), JMLR: W&CP, 32, 1188-1196, arXiv:

[14] M. G. Kreines, A. A. Afonin, "Klasterizatsiia tekstovykh kollektsii: pomoshch pri soderzha-telnom poiske i analiticheskii instrument", Internet-portaly: soderzhanie i tekhnologii, 4, FGU GNII ITT "Informika", eds. A.N. Tikhonov (pred.) i dr.,

Prosveshenie, M., 2007, 510-537

[15] M. G. Kreines, “Modeli tekstov i tekstovyh kolliktsii dlia poiska i analyza informatsii”, Trudy MFTI, 3 (2017), 132-142

[16] M. G. Kreines, E. M. Kreines, “The control model for the selection of reference collections providing the impartial assessment of the quality of scientific and technological pub-lications by using bibliometric and scientometric indicators”, J. of Comp. and Systems Sci. Intern., 55:5, 750-766 · Zbl 1384.93018

[17] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, “Optimizing semantic coherence in topic models”, Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing (Edinburgh, Scotland, UK, July 27-31, 2011), 262-272

[18] D. Newman, J. H. Lau, K. Grieser, T. Baldwin, “Automatic evaluation of topic coherence”, Human Language Technologies, The 2010 Annual Conf. of the North American Chapter of the ACL (Los Angeles, California, 2010), 100-108

[19] D. Newman, Y. Noh, E. Talley, S. Karimi, T. Baldwin, “Evaluating topic models for digital libraries”, Proc. of the 10th ann.Joint Conf. on Digital libraries, JCDL’10, ACM, New York, NY, USA, 2010, 215-224

[20] K. V. Vorontsov, A. A. Potapenko, Additivnaia reguliarizatsiia tematicheskih modelei, 2014, 22 pp.

[21] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D. M. Blei, “Reading tea leaves: How humans interpret topic models”, NIPS 2009, 288-296

[22] M.G. Kreines, E.M. Kreines, “Control model for the alignment of the quality assessment of scientific documents based on the analysis of content-related context”, J. of Computer and Systems Sciences International, 55:6, 938-947