Kreĭnes, M. G.; **Kreĭnes, Elena M.**

Matrix text models. Text models and similarity of text contents. (Russian. English summary)
Zbl 1444.68266
Mat. Model. 32, No. 1, 31-49 (2020).

Summary: We present a matrix model of texts on natural languages and a model of quantitative assessment of similarity of text contents. An application of the model to search for the texts with similar content is considered. We discuss the difference of the proposed matrix models and commonly used approaches to analyze and model natural language texts.

MSC:

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

Keywords:

natural language texts; similarity of text contents; similarity assessment; text models; text information retrieval

Software:

word2vec

Full Text: DOI MNR

References:

[1] A. Ia. Shaikevich, V. M. Andrushchenko, N. A. Rebetskaia, Distributivno-statisticheskii analiz iazyka russkoi prozy 1850-1870 gg, v. 1, Iazyki slavianskoi kultury, M., 2013, 499 pp.

[2] 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

[3] Q. Le, T. Mikolov, "Distributed representations of sentences and documents" (Beijing, China, 2014), JMLR: W&CP, 32:2, Proceedings of the 31-st International Conference on Machine Learning, 1188-1196, arXiv:

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

[5] V. A. Uspenskii, "Predvarenie dlia chitatelei "Novogo literaturnogo obozrenia" k semioticheskim poslaniiam Andreia Nikolaevicha Kolmogorova", Novoe literaturnoe obozrenie, 1997, no. 24, 123-215

[6] K. V. Anisimovich, K. Yu. Druzhkin, K. A. Zuev, F. R. Minlos, M. A. Petrova, V. P. Selegei, "Syntactic and semantic parser based on ABBYY compreno linguistic technologies", Computer Linguistics and Intellectual Technologies, Proceedings of XVIII International conference "Dialog 2012", 2012, 91-103

[7] J. Fan, A. Kalyanpur, D. C. Gondek, D. A. Ferrucci, "Automatic knowledge extraction from documents", IBM J. RES. \& DEV, 56:3/4 (2012), 5, 10 pp. · Zbl 1241.62070

[8] E. V. Rahilina, Lingvistika konstruktsii, Azbukovnik, M., 2010, 583 pp.

[9] O. P. Kuznetsov, V. S. Suhoverov, L. B. Shipilina, "Ontologia kak sistematizatsiia nauchnyh znanii: struktyra, semantika, zadachi", Trudy konferentsii "Tehnicheskie i programmnie sredstva sistem upravleniia, kontrolia i izmereniia", IPU RAN, M., 2010, 762-773

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

[11] 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

[12] N. Loukachevitch, B. Dobrov, "The Sociopolitical Thesaurus as a resource for automatic document processing in Russian", Terminology, 21:2, Special issue "Terminology across languages and domains" (2015), 238-263

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

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

[15] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing \& Management, 24:5

(1988), 513-523

[16] B. Trstenjak, S. Mikac, D. Donko, “KNN with TF-IDF based framework for text categorization”, Procedia Engineering, 69 (2014), 1356-1364

[17] H. C. Wu, R. W.P. Luk, K. F. Wong, K. L. Kwok, “Interpreting TF-IDF term weights as making relevance decisions”, ACM Transactions on Information Systems, 26:3 (2008), 1-37

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

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

[20] 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

[21] A. N. Kolmogorov, Teoriia informatsii i teoriia algoritmov, Nauka, M., 1987, 304 pp.

[22] Y. Bengio, H. Schwenk, J. S. Senècal, F. Morin, J. L. Gauvain, “Neural probabilistic language models”, Innovations in Machine Learning, Springer, N.-Y., 2006, 137-186

[23] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank”, Conference on Empirical Methods in Natural Language Processing, 2013, 1631-1642

[24] K. K. Nicodemusa, B. Elvevåg, P. W. Foltzd, M. Rosensteind, C. Diaz-Asperf, D. R. Weinberger, “Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach”, Language, Computers and Cognitive Neuroscience, 55 (2014), 182-191

[25] J. Grimmer, “A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases”, Polit. Anal., 18:1 (2010), 1-35

[26] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, F. Menczer, “Predicting the political alignment of twitter users”, Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Intern. Confer. on Social Computing (SocialCom), IEEE, 2011

[27] W. Zhu, Ch. Chen, R. B. Allen, “Analyzing the propagation of influence and concept evolution in enterprise social networks through centrality and Latent Semantic Analysis”, Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, 5012, 2008, 1090-1098

[28] G. Salton, A. Wong, C. S. Yang, “A vector space model for automatic indexing”, Communications of the ACM CACM, 18:11 (1975), 613-620 · Zbl 0313.68082

[29] Zh. Yiu, J. Rong, Zh. Zhi-Hua, “Understanding bag-of-words model: A statistical framework”, International J. Machine Learning and Cybernetics, 1:1-4 (2010), 43-52

[30] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, 2016, 5 pp., arXiv:

[31] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, 2016, 7 pp., arXiv:

[32] Ch. Aswani Kumar, S. Srinivas, “On the performance of latent semantic indexing-based information retrieval”, J. of Comp. and Inform. Technol. - CIT, 17:3 (2009), 259-264

[33] M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, “From Word Embeddings To Document Distances” (Lille, France, 2015), JMLR: W&CP, 37, Proceedings of the 32 nd International Conference on Machine Learning, 957-966

[34] G. Huang, Ch. Guo, M. J. Kusner, Y. Sun, K. Q. Weinberger, F. Sha, “Supervised Word Mover”s Distance”, 30th Conference on Neural Information Processing Systems (NIPS 2016) (Barcelona, Spain, 2016), 9 pp.

[35] M. G. Kreines, A. A. Afonin (Patentoobladateli), Patent na poleznuiu model 60751 “Sistema formirovaniia lingvisticheskih dannyh dlia poiska i analiza tekstovyh documentov”, 2007

[36] M. G. Kreines, A. A. Afonin (Patentoobladateli), Patent na poleznuiu model 62263 “Sistema formirovaniia semanticheskih dannyh dlia poiska i analiza tekstovyh documentov”, 2007

[37] M. G. Kreines, “Informatsionnaia tehnologiia smyslovogo poiska i indeksirovaniia informatsii v elektronnyh bibliotekah: kluchi ot texta”, Nauchnyi servis v seti Internet, 1999, 214-218, MGU, M.

[38] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jegou, T. Mikolov, FastText.zip: Compressing text classification models, 2016, 13 pp., arXiv:

[39] H. P. Luhn, “A statistical approach to mechanized encoding and searching of literary information”, IBM Journal of research and development, 1:4 (1957), 309-317

[40] C. D. Manning, P. Raghavan, H. Schutze, “Scoring, term weighting, and the vector space model”, Introduction to Information Retrieval, Ch. 6, Cambridge University Press, Cambridge, 2008, 100-123

[41] S. E. Robertson, S. Walker, M. Beaulieu, “Experimentation as a way of life: Okapi at TREC”, Information Processing \& Management, 36 (2000), 95-108

[42] J. H. Lee et al., “Automatic generic document summarization based on non-negative matrix factorization”, Information Processing \& Management, 45:1 (2009), 20-34

[43] N. V. Timofeev-Resovskii, Vospominaniia, Vagrius, M., 2008, 397 pp.

[44] M. G. Kreines, “Intellectual Information Technologies and Scientific Electronic Publishing: Changing World and Changing Model”, Elpub 2002 Technology Interactions, Proceedings of the 6-th International ICCC/IFIP Conference on Electronic Publishing, Verlag fur Wissenschaft und Forschung, Berlin, 2002, 135-142

[45] A. A. Petrov, M. G. Kreines, A. A. Afonin, “Semanticheskii poisk nestrukturirovannoi tekstovoi informatsii na estestvennyh iazikah v zadachah organizatsii ekspertizy pri realizatsii nauchno-technicheskih program”, Informatizatsiia obtazovaniia i nauki, 18:2 (2013), 54-67

[46] A. A. Petrov, M. G. Kreines, A. A. Afonin, “Vychislitelnie modeli semantiki tekstovyh istochnikov informatsii dlya informatsionnogo obespecheniia nauchno-technicheskoi ekspertizy”, Matematicheskoe modelirovanie, 28:6 (2016), 33-52

[47] A. Singhal, “Modern Information Retrieval: A Brief Overview”, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24:4 (2001), 35-43

[48] B. Larsen, C. Aone, “Fast and effective text mining using linear-time document clustering”, Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, 16-22

[49] G. Salton, Automatic Text Processing, Addison-Wesley, N.-Y., 1989, 543 pp.

[50] B. Li, L. Han et al., “Distance weighted cosine similarity measure for text classification”, Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, 8206, eds. H. Yin et al., 2013, 611-618

[51] T. Saracevic, “Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective”, Library Trends, 56:4 (2008), 763-783