Grigorieva, Elena Gennadievna; **Klyachin, Vladimir Aleksandrovich**

The study of the statistical characteristics of the text based on the graph model of the linguistic corpus. (Russian. English summary) Zbl 1442.68242

Izv. Sarat. Univ. (N.S.), Ser. Mat. Mekh. Inform. 20, No. 1, 116-126 (2020).

Summary: The article is devoted to the study of the statistical characteristics of the text, which are calculated on the basis of the graph model of the text from the linguistic corpus. The introduction describes the relevance of the statistical analysis of the texts and some of the tasks solved using such an analysis. The graph model of the text proposed in the article is constructed as a graph in the vertices of which the words of the text are located, and the edges of the graph reflect the fact that two words fall into any part of the text, for example, in – a sentence. For the vertices and edges of the graph, the article introduces the concept of weight as a value from some additive semigroup. Formulas for calculating a graph and its weights are proved for text concatenation. Based on the proposed model, calculations are implemented in the Python programming language. For an experimental study of statistical characteristics, 24 values are distinguished, which are expressed in terms of the weights of the vertices, edges of the graph, as well as other characteristics of the graph, for example, the degrees of its vertices. It should be noted that the purpose of numerical experiments is to squeak in the characteristics of the text, with which you can determine whether the text is man-made or randomly generated. The article proposes one of the possible such algorithms, which generates random text using some other text created by man as a template. In this case, the sequence of parts of speech in an auxiliary text alternation is preserved in the random text. It turns out that the required conditions are satisfied by the median value of the ratio of the text graph edge weight value to the number of sentences in the text.

MSC:

68T50 Natural language processing

Software:

word2vec

Full Text: DOI MNR

References:

[1] Kipyatkova I. S., Karpov A. A., “Automatic processing and statistic analysis of the news text corpus for a language model of a Russian language speech recognition system”, Information and Control Systems, 2010, no. 4 (47), 2-8 (in Russian)

[2] Kolmogorova A. V., Kalinin A. A., Malikova A. V., “Linguistic principles and computational linguistics methods for the purposes of sentiment analysis of Russian texts”, Actual problems of philology and pedagogical linguistics, 2018, no. 1 (29), 139-148 (in Russian)

[3] Voronina I. E., Kretov A. A., Popova I. V., “Algorithms of semantic proximity assessment based on the lexical environment of the key words in a text”, Proceedings of Voronezh State University. Ser. Systems analysis and information technologies, 2010, no. 1, 148-153 (in Russian)

[4] Berman N. D., Levenets A. V., Sergeeva L. A., “Statistical analysis of textual information”, Information Technologies of the XXI Century, Collection of Scientific Papers, Izdatel’stvo Tikhookeanskgo gosudarstvennogo universiteta, Khabarovsk, 2016, 282-286 (in Russian)

[5] Donina O. V, “The application of data mining methods in linguistics”, Proceedings of Voronezh State University. Ser. Systems analysis and information technologies, 2017, no. 1, 154-160 (in Russian)

[6] Mikolov T., Chen K., Corrado G., Dean J., Efficient Estimation of Word Representations in Vector Space, arXiv:

[7] Raigorodskii A. M., “Random Graphs”, Mathematics in Problems, Izdatel’stvo Moskovskogo tsentra nepreryvnogo matematicheskogo obrazovaniya, M., 2009, 312-315 (in Russian)

[8] Erdős P., Rányi A., “On random graphs I”, Publ. Math. Debrecen, 6 (1959), 290-297 · Zbl 0092.15705

[9] Newman M. E. J., Strogatz S. H., Watts D. J., “Random graphs with arbitrary degree distribution and their applications”, Phys. Rev. E, 64 (2001), 26-118

[10] Pavlov Yu. L., Cheplyukova I. A., “Random graphs of Internet type and the generalised allocation scheme”, Discrete Mathematics and Applications, 18:5 (2008), 447-463 · Zbl 1171.05419

[11] Pavlov Yu. L., "On the limit distributions of the vertex degrees of conditional Internet graphs", Discrete Mathematics and Applications, 19:4 (2009), 349-359 · Zbl 1237.05194