Demidovich I., Shynkarenko V., Kuropiatnyk O., Kirichenko O. Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task. International Scientific and Technical Conference on Computer Sciences and Information Technologies. Vol. 2 : 16th IEEE International Conference on Computer Science and Information Technologies (CSIT 2021), Lviv, 22–25 September 2021. P. 48–51. DOI: 10.1109/CSIT52700.2021.9648829.

Demidovich Inna

Dnipro National University of Railway Transport
Department of Computer Information Technology
Dnipro,Ukraine; ORCID 0000-0002-3644-184X

Shynkarenko Viktor

Dr. tech. sciences, professor, head of department "Computer Information Technologies", Dnipro National University of Railway Transport named after Academician V. Lazaryan, Lazaryan, 2, Dnipro, Ukraine, 49010, tel. +38 (056) 373 15 35, e-mail shinkarenko_vi@ua.fm, ORCID 0000-0001-8738-7225

Kuropiatnyk Olena

Dnipro National University of Railway Transport
Department of Computer Information Technology
Dnipro,Ukraine; ORCID 0000-0003-2286-884x

Kirichenko Oleksandr

Dnipro National University of Railway Transport
Department of Computer Information Technology
Dnipro,Ukraine

Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task

Abstract:
The previously developed method establishes the natural language texts authorship based on frequency analysis, supplemented by indicators of text complexity and recurrent analysis. The authorship indication problem is reduced to the pattern recognition classical theory. To account for the different individual indicators information content, their weights are taken into account. They are determined according to the maximum number of the correctly established texts authorship from the training sample using a genetic algorithm. This method is used to study the effectiveness of the author's style representation that is based on different types of words processing: two types of words stems and 4-grams. To obtain stems, the adapted Porter stemmer is used and creating a dictionary of the foundations of the Ukrainian language original method is applied, respectively. Taking into account the calculated indicators weights, the reliability of establishing the text authorship in the control sample reached 85-91%.

Keywords: natural language texts, authorship attribution, Porter stemmer, genetic algorithm, recurrent analysis, statistical analysis, text classification, dictionary, pattern recognition

References:

1. R. Iyer and C. Rosé, "A Machine Learning Framework for Authorship Identification FromTexts", ArXiv abs/1912. 10204, 2019.
2. R.A. Hardcastle, "CUSUM: a credible method for the determination of authorship?", Science & Justice: Journal of the Forensic Science Society, vol. 37, no. 2, pp. 129-138, 1997.
3. J. Rygl and A. Horák, "Authorship Attribution: Comparison of Single-Layer and Double-Layer Machine Learning" in Text Speech and Dialogue. TSD 2012. Lecture Notes in Computer Science, Berlin, Heidelberg:Springer, vol. 7499, pp. 282-289, 2012.
4. M. Lupei, A. Mitsa, V. Repariuk and V. Sharkan, "Identification of authorship of Ukrainian-language texts of journalistic style using neural networks", Eastern-European Journal of Enterprise Technologies, vol. 1, no. 2, pp. 30-36, 2020.
5. V. Lytvyn et al., "Development of the Quantitative Method for Automated Text Content Authorship Attribution Based on the Statistical Analysis of N-grams Distribution", Eastern-European Journal of Enterprise Technologies, vol. 6, no. 2, pp. 28-51, 2019.
6. V. Moshkina, I. Andreeva and N. Yarushkinaa, "Solving the problem of determining the author of text data using a combined assessment", CEUR Workshop Proceedings 2782, pp. 112-118, 2020.
7. I. I. Drozdova and A. D. Obuhova, "Opredelenie avtorstva teksta po chastotnyim harakteristikam", (determining the authorship of the text by frequency characteristics) in: Tehnicheskie nauki v Rossii i za rubezhom: materialyi VII Mezhdunarodnoy nauchnoy konferentsii, pp. 18-21, 2017.
8. Yunita Sari, Andreas Vlachos and Mark Stevenson, "Continuous N-gram Representations for Authorship Attribution", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, pp. 267-273, 2017.
9. I. Markov, J. Baptista and O. Pichardo-Lagunas, "Authorship Attribution in Portuguese Using Character N-grams", Acta Polytechnica Hungarica, vol. 14, no. 3, pp. 59-78.
10. D. L. Hoover, "Frequent word sequences and statistical stylistics", Literary and Linguistic Computing, vol. 17, no. 2, pp. 157-180, 2002.
11. G. O. Sidorov, "Automatic Authorship Attribution Using Syllables as Classification Features", Rhema journa11, pp. 62-81, 2018.
12. H. Gómez-Adorno, JP. Posadas-Durán and G. Sidorov, "Document embeddings learned on various types of n-grams for cross-topic authorship attribution", Computing 100, pp. 741-756, 2018.
13. O. Marchenko, A. Anisimov, A. Nykonenko, T. Rossada and E. Melnikov, "Authorship attribution system", Artificial Intelligence, vol. 2, pp. 77-85, 2016.
14. G. Wimmer, G. Altmann, L. Hrebicek, S. Ondrejovic and S. Wimmerová, "Uvod do analyzy textov", Univerzita Komenského v Bratislave, 2003.
15. V.I. Shynkarenko and I.M. Demidovich, "Determination of the attributes of authorship of natural texts", Artific. Intell., vol. 3, pp. 27-35, 2018.
16. V.I. Shynkarenko and I.M. Demidovich, "Authorship Determination of Natural Language Texts by Several Classes of Indicators with Customizable Weights", Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021), vol. I, pp. 832-844, April 22–23, 2021.
17. Great electronic dictionary of Ukrainian language (VESUM), [online] Available: https://github.com/brown-ukldict_uk.
18. T. V. Golub and M. Yu. Tyagunova, "Method of steaming Ukrainian-language texts for classification of documents based on Porter's algorithm", Scientific works of Donetsk National Technical University. Series: Informatics cybernetics and computer engineering, vol. 1, no. 24, pp. 59-63, 2017.

19. A. Hlybovets and V. Tochytsky, Algorithm of tokenization and steming for texts in Ukrainian, 2017.
20. J. P. Zbilut and C. L. Webber, "Embeddings and delays as derived from quantification of recurrence plots", Physics Letters A, vol. 171, no. 3-4, pp. 199-203, 1992.
21. Yu. V. Rohushyna, "Ispolzovanie kriteriev otsenki udobochitaemosti teksta dlia poiska informatii sootvetstvuiushchei realnym potrebnostiam polzovatelia (the usage of criteria for evaluating the readability of the text to find information that meets the real needs of the user.)", Problemy prohramiuvannia 3, pp. 76-88, 2007.
22. P. S. Sisodia, V. Tiwari and A. Kumar, "A Comparative Analysis of Remote Sensing Image Classification Techniques", International Conference on Advances in Computing Communications and Informatics (ICACCI), pp. 1418-1421, 2014.
23. V. Shynkarenko and O. Kuropiatnyk, "Constructive Model of the Natural Language", Acta Cybernetica, vol. 23, no. 4, pp. 995-1015, 2018.