12-14-2020

Models and algorithms for complex analysis of large corpuses of Russian poetic texts

Olga Kozhemyakina

Vladimir B. Barakhnin

Models and algorithms for complex analysis of large corpuses of Russian poetic texts

Kozhemyakina, Olga Yu., Federal Research Center for Information and Computational Technologies, Novosibirsk Russia, olgakozhemyakina@mail.ru

Barakhnin, Vladimir B., Federal Research Center for Information and Computational Technologies, Novosibirsk Russia, bar@ict.nsc.ru

Abstract

We propose the algorithm of automated definition of the genre type and semantic characteristics of poetic texts in Russian. We formulated the approaches to the construction of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts, based on the definition of interdependence of the type of genre and stylistic colouring of the text. On the basis of these approaches the principles of formation of the training samples for the algorithms for the definition of styles and genre types were analysed. The computational experiments were conducted using the corpus of texts of A. S. Pushkin's Lyceum lyrics to select the most accurate algorithm for classifying poetic texts, including using the most well-known techniques for ensembling basic algorithms in composition, such as weighted voting, boosting and stacking, and single words, bigrams and trigrams were used as characteristic features of poems.

Keywords: Analysis of Russian poetic texts, Classifier of genre and style, Weighted voting, Boosting, Stacking

One of the most actual philological tasks at present moment is the analysis of poetic texts in order to identify the various characteristics of the literature work (the information about the verse, the rhythmics of the end of poems, a detailed description of the used rhymes, etc.). All this is necessary in the process of study of the author's work. One of the most important characteristics of a poetic text are its style and genre. Currently, the specialists in this field are forced to work with the classification almost manually, the information about the reasons of the provided predictions is not available to the user, although this information is very important in further analysis of the results. These problems can be solved by creating a software system that allows to load the categorized data, and to show a detailed report on the results of the classification on the output display. The implementation of the web application, for the control of the algorithms for automatic determination of styles and genre types of poetic texts, is a purpose of this study. It allows the experts-philologists to download the necessary data in a convenient way, to choose an automatic classifier and to analyze valuably the result of the classification, based on its justification obtained by the LIME algorithm and its implementation in the ELI5 library. This decision is not tied to any specific categories of classification, what makes it universal and easily expandable for different range of classification problems. This application allows to minimize the time of training the interaction of an expert-philologist with the system. This is done by providing the expert with appropriate advice on the style or genre of the poetic text. The approaches to the construction of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts, based on the definition of interdependence of the type of genre and stylistic colouring of the text, are formulated. On the basis of these approaches the principles of formation of the training samples for the algorithms for the definition of styles and genre types were analyzed. Computational experiments were conducted using the corpus of texts of A. S. Pushkin's Lyceum lyrics to select the most accurate algorithm for classifying poetic texts, including using the most well-known techniques for ensembling basic algorithms in composition, such as weighted voting, boosting and stacking, and single words, bigrams and trigrams were used as characteristic features of poems. The algorithms considered have shown their efficiency (based on the criterion of maximizing minimum accuracy, a multi-layer perceptron should be used,

and trigrams should be used as lexical characteristics of poems) and can be used to automate the complex analysis of Russian poetic texts, significantly facilitating the expert's work in determining their styles and genres by providing appropriate recommendations

Acknowledgments. The study was carried with the support of the Russian Science Foundation (project No. 19-18-00466).