Title: Measuring readability of technical texts

Author: Anna Kriukova

Faculty of Mathematics and Physics: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Cinková Silvie, Ph.D., Institute of Formal and Applied Linguistics

Abstract: This research explores various approaches to measuring readability of technical texts. The data I work with is provided by Hyperskill, an online educational platform dedicated mostly to Computer Science, where I did my internship. In the first part of my research, I examine classical readability formulas and try to find correlations between their values and the user statistics available for the texts. The results show that there are no high correlations, thus, the standard formulas are not suitable for the task. The second part of the research is dedicated to experiments with machine learning algorithms. Firstly, I use four sets of features to predict the average rating, completion time, and completion rate of a step. Then, I introduce a rule-based algorithm to split the texts into well- and poorly-written ones, which relies on students' comments. However, binary classification trained on this division shows low results and is not used in the final pipeline. The system suggested as the outcome of my work employs the user statistics' prediction for new texts and a rule-based comment-focused algorithm for the published texts. As a result, texts that the system identifies as "bad" are reported to the Hyperskill content team for improvement.

Keywords: readability technical texts data analytics corpus linguistics comprehensibility machine learning