

Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce Anna Kriukova
Název práce Measuring readability of technical texts
Rok odevzdání 2022
Studijní program Informatika **Studijní obor** Matematická lingvistika

Autor posudku Barbora Vidová Hladká **Role** oponent
Pracoviště ÚFAL MFF UK

Text posudku:

In her Master thesis, Anna Kriukova explores the data that come from the on-line learning platform Hyperskill to provide the Hyperskill content team with tools for improving their learning materials. Namely, (1) she analyzes the readability of Hyperskill's theoretical pieces (texts) and (2) she performs machine learning experiments to predict learner data (e.g., completion rates on a task) that can contribute to readability assessment. While the goal (1) was specified in the guidelines of thesis preparation, the goal (2) was added beyond the scope of the original guidelines.

In summary, the student performed a basic statistical analysis of the data, focused on two basic readability measures and their correlation with learner data, and performed several machine learning experiments.

The thesis presents a solid and useful work. It is well structured into four chapters and several appendices. Both the structure of the thesis and the student's style allow readers fluent reading. Each chapter begins with a short and clear introduction and ends with a summary. Therefore readers know very well starting points and what was done. I really appreciate the review of related work that is reflected by a rich bibliography counting almost 90 items. Below are my detailed comments addressing the content of the thesis.

Section 2.1

- The description of the dataset you were working with would be clearer if it were illustrated by real examples from the Hyperskill platform. Furthermore, it would be useful to do mapping between the terminology used in the thesis and in the platform.
- If I understand well, a *theory step* is a text representing a theoretical part of the Hyperskill learning process and each *topic* is associated with a collection of texts (and tasks not covered in the thesis).

Section 2.2

- The data extraction procedure resulted in 1,321 texts, right? What is the number of topics in this data set?

Section 2.4

- I would pay some attention to handling NaN values. You should considerate whether to remove them completely or replace them with a specific value (e.g. mean, median).

Section 2.5

- Do the 1,120 (out of 1,321) texts have at least one comment? What is the total number of comments? Based on Figure 2.5, I would say almost 25 thousand comments.
- What is the distribution of the number of comments in the dataset of 1,321 texts?
- The caption of Figure 2.5 should be The distribution of the number of comments across various areas.

Section 4.3

- Are all the texts associated with a topic written by one author (or a team of authors?)
- What is a setting of the experiments Version 1-5? You do sentiment analysis of the comments. Then you classify the topic readability. Why do not you classify the text readability?
- Why did you extract vocabulary from 50 comments only?

Section 4.4

- The Version 1-5 results are imbalanced with a great majority of good topics. Why did you select the same number of good and bad topics for manual annotation? What is the number of comments manually annotated?
- What is the range of the target attribute Average like of the text? $Y = \{-2, -1, 0, 1, 2\}$?
- I propose to run experiments with all the data as well, i.e., not to filter out the texts with less than 20 comments.
- Please, provide the formula for Mean Average Error.
- On which proportion of the data you run a 5-fold cross validation? In traditional machine learning, data is split into three parts - training, development test and evaluation test. The training and development subsets are used for training a best model that is tested on the eval subset at the very end.

General comments

- Type more readable labels of axes in plots
- Cite tables, figures, sections, chapters properly, i.e., Table 3 or Figure 2

The student proves that she is capable to conduct research on her own. I recommend the thesis to be defended.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 19. 08. 2022

Podpis: