V0.2

Some of the formulas in posts.xml and comments.xml did not have a “math-container” tag and id. The issue was in the formula extraction phase, those latex formulas between two $ sign, where not captured. In this version, they have the “math-container” tag with a unique id.

V0.3

In the previous version, there were some xml validation errors and in the new version those errors are fixed. In the GitHub repo, the new code to read the xml files partially and completely is added.

V1.0

In the previous version, the formula extraction had some issues. For instance, some of them were missed. In this version that is fixed. Note that the Post file is in both zip and xml format for easier download.

V1.1 (May 9, 2020)

After releasing the dataset, there was a mistake in annotating the formulas, some of the formulas were misannotated. This issue was found by Moritz Schubotz. Here is an example provided by him:

<<span class="math-container" id="13734904">p</span>>

In this version, this issue is fixed. Only the comment and post files are changed.

V1.2 (May 22, 2020)

Correcting the issue mentioned by Moritz, we ran our code on the new Archive collection, which had posts from 2020 as well. Our collection should only have posts from 2010 to 2018. We have removed those posts in this version. This issue was mentioned by Yin Ki Ng.

V1.3 (Feb 28, 2022)

In this version, the formula ids are correctly annotated in the Post and Comment XML file. If the formula is not annotated, we mark them in the TSV index files for the formulas. This issue was mentioned by Yin Ki Ng from the MathDowser team in CLEF 2021.