README: ARQMath Formula Datasets

Dataset created by Behrooz Mansouri

Author: B. Mansouri and R. Zanibbi, Feb 28, 2022

The formula datasets include all formulas represented as LaTeX in the ARQMath collection, produced from Math Stack Exchange posts from 2010-2018.

Formats

The main index represents the LaTeX for each identified formula (where each formula is given *explicitly* as LaTeX in the Math Stack Exchange postings). This LaTeX index is the 'official' formula index for the ARQMath tasks. It contains:

```
28,320,920 formulas in LaTeX
```

There are also two sub-indices for the MathML representations for each LaTeX formula, created using LaTeXML (https://dlmf.nist.gov/LaTeXML):

```
1. SLT - Symbol Layout Tree: in Presentation MathML format (appearance)
2. OPT - Operator Tree: in Content MathML format (semantics)
```

In ARQMath-1, the MathML representations were incomplete, as LaTeXML was sometimes unable to convert LaTeX strings due to syntax errors, ambiguities, or other issues. In the newers collection (using an updated LaTeXML, 0.8.5, for ARQMath-2 and ARQMath-3) all formulas successfully converted to Presentation MathML were also converted to Content MathML.

For ARQMath-3, in each sub-index we have:

```
28,267,310 formulas in Presentation MathML (SLT) ~99.81%
28,267,310 formulas in Content MathML (OPT)  ~99.81%
```

For ARQMath-3, visually distinct identifiers used in evaluation have been corrected in some cases. The number of converted formulas is just slightly smaller than in ARQMath-2, but this

primarily because a number of invalid LaTeX strings were skipped this time, avoiding some errors noticed in assigning both MathML trees and visual ids to formulas.

File Sizes and Organization

Expanded sizes for each zip file:

```
1.6GB latex_representations.zip
10.0GB slt_representations.zip
9.0GB opt_representations.zip
```

Each formula index is represented using a list of tab-separated variable (TSV) files, of roughly 10-100MB each in size.

All TSV index files (for LaTeX, Presentation MathML, and Content MathML) contain 9 fields (commas shown for clarity, files separate columns with tabs (`\t`):

```
[ id, post_id, thread_id, type, comment_id, old_visual_id, visual_id, iss
ue, formula ]
```

Please note: the first four fields are identical for each individual formula in all index files (i.e., every formula has a unique id, post_id, thread_id, and post type).

Here is a detailed description of each column/field in a formula TSV file:

1. **'id'** is the unique integer identifier for the formula.
2. **'post_id'** identifies the post where the formula appears. **Note** that for formulas in the comments this id represents the post id not the comment id.
3. **'thread_id'** identifies the question thread in which the post associated with the formula appears, along with the thread in which the post appears.
4. **'type'** refers to the type of post, which can be one of four values: 'question,' 'comment,' 'answer,' or 'title.' Note that for question posts, the title field of the question ('title') and body of the question ('question') will have the *same* post and thread identifiers.
5. **'comment_id'** is the id of comment in which the formula appreared. If the formula has appreared in posts, this column is empty.
6. **'old_visual_id'** is an integer identifier that shows the visual id of a formula. Visual ids are used for clustering of results in task 2. In ARQMath-2 the organizers realized an error in some of the visual clusters. This happened error happened as some of the formulas were asssinged wrong SLT representation. We included this column in our index file, for fair comparison with the runs in CLEF 2021. We have update QREL files with new visual ids

and in ARQMath-3 **the visual ids in this column will not be used**.

7. **'new_visual_id'** this is the visual id that will be used in ARQMath-3. For the formulas that had wrong visual id, this column has the new visual id. If the formula had a valid visual id, we simply use the 'old_visual_id'. The visual ids are assigned by comparing the SLT string of formulas using Tangent-S system and if the SLT representation was not available, we used the LaTex representation by removing the spaces.
8. **'issue'** shows if there was an issue with the formula. During the previous cycles of ARQMath lab our participants have detected two main issues with the collection: 1) formula not existing in the XML files 2) formula having wrong visual id. This column indicates if there is an issue with the formula. For the formulas that do not exist in the XML files, value 'd' is indicated in this column. **Note** that these formulas should not be used in retrieval results for task 2. Formulas that had wrong visual id has value 'v' in this column. Finally, if both errors occured for a formula, we have 'dv' in this column.
9. '**formula'** is the text representation of the formula (as LaTeX, Presentation MathML (SLT), or Content MathML (OPT)). It is provided between double quotes (to allow easy parsing by spreadsheet programs and CSV/TSV reading libraries).

Formulas in the ARQMath Corpus

Formulas are represented using tags within the ARQMath Corpus, using:

```
<span id=FID class="math-container"> ... </span>
```

Where FID is a formula identifier (integer representing the individual formula).

These tags are used in Math Stack Exchange to represent math for rendering by MathJax; in creating the collection we added additional tags so that all formulas to be used in the tasks would be tagged/identifiable. Some formulas are not provided in explicit LaTeX, in which case they will not be tagged, or used for evaluation in this first instance of the competition.