X5GON Database Sanitisation Project

Now that X5GON is constantly indexing new records to the X5GON database, care has to be taken to make sure that the data collected by the X5GON index is correct and sensible to the best of our knowledge. This project is driven towards identifying the issues with the current X5GON database and mitigating these issues.

As the first phase of project, three main components are being focused on:

1. Automatic language detection of materials
2. Developing deduplication techniques for educational materials
3. Flagging unusable materials

Automatic language detection

Most materials that are indexed by X5GON get the language as metadata from the repository. Most of the time, these records are either missing or wrong. We need an automatic fast method to detect the language/s in the material.

Deduplication

Another major issue with X5GON is duplicate materials. As OER resources can be indexed in multiple repositories. Therefore, there are multiple OERs that have near-identical contents but have very different meta-data (urls, creation dates etc.) that makes it even harder to use simple meta-data based rules themselves to resolve duplicate materials. These duplicate materials heavily impact the user experience when it comes to search and recommendation tasks as these ML tools heavily depend on similarity search. Ideally, we would like to detect groups of duplicate materials so that the duplicates can be flagged and then one of the duplicates to be visible outside the index. There also lies questions on how to select the representative version among the duplicates (eg. license clarity, creation date, source etc.)

Flagging unsuitable materials

Another major problem that is evident in X5GON is the presence of materials that are unsuitable for consumption. From the observations in the pilots, a few isolated cases of “unsuitable materials” fall into categories such as exam question sheets, reference lists with no context, a score sheet, very short documents with incomplete information, part of a non-complete learning material etc. We need to figure out how to isolate these cases automatically and come up with methods to clean them from the database.

Task Breakdown

The following subtasks are anticipated to be done in order to solve the above problems.

- Language detection
    - Do background research
    - Identify a set of potential solutions with a dataset/method for testing
    - Design and run benchmarks to compare the solutions
    - Develop the code to incorporate the solution to the pipeline
- Flagging duplicates
    - Do background research on the different methods to solve this problem
    - Run exploratory analysis to understand the dataset
    - Identify a few basic methods that can be used to identify duplicates
    - Identify and develop an experiment to compare them and run benchmarks
    - Develop the code to incorporate the solution to the pipeline
- Flaggin unnecessary material
    - Run exploratory analysis to understand the dataset and isolate likely unusable materials
    - Identify a few tools and methods to detect unsuitable materials and to flag them
    - Develop the solution to incorporate these methods in the pipeline

Deliverables:

- Report outlining all the experiments undertaken
    - Full documentation of the background research
    - Description of the candidate solutions identified, experimental details and the results.
    - Observations and conclusions
- Fully functional source-code
    - All source code developed and committed to a git-repository
    - Well documented code (PEP8 standard python code)
    - A Flask API that incorporates the tools
- Production deployment:
    - Cleaned version of the existing database
    - Deploy the API endpoints to the production pipeline

Duration

2-3 months from the time the Database dump is received by the team

Out of Scope

There are numerous other issues that need to be addressed in X5GON database. To name a few,

- Licence detection
- Hyperlink resolution (redirects etc...)

These tasks are out of scope in this stage of the project.