Request for computer programming assistance in corpus linguistics

Please contact:
Geoffrey Pinchbeck, Assistant Professor,
ALDS, SLALS, Faculty of Arts and Social Sciences
251 Patterson Hall
geoff.pinchbeck@carleton.ca

The terms for this work will be discussed in person

Project 1: Restructuring and cleaning of corpora of TV/Movie captions/sub-titles

I have large corpora in several languages of tv/movie subtitles from the open source website opensubtitles.
These corpora need to be cleaned up and restructured.

Currently they are in this format:

- .xml formatted text files which are within a folder that is named with the IMDB ID code of the TV-show or movie that it contains.

NOTE: I am currently interested in the movie-tv scripts for the following languages: English, French, Mandarin Chinese, Spanish, Dutch, and possibly others. Therefore, all the work below would have to allow for the text files to be processed in unicode (utf-8) format or something compatible.

- Within each folder there might be multiple versions of sub-titles/captions of the same movie or tv-show, due to different versions of the title or due to multiple submissions by contributors.
- the folders for each title are within another folder that is named for the year that the tv-show or movie came out
- the year folders are within another directory, which is in another directory.

I need to perform the following steps. I have code which will perform the steps, and I have had the code running properly in the past, but I am currently unable to get this code to work, probably because I have forgotten how to do this properly... I would like to make the following workflow smooth, and I need to learn how to perform it myself as new and updated corpora in different languages become available. All processes will need to work with tools that will run on my Mac: terminal or some kind of Python platform.

1) (code currently in javascript) recursively find all of the .xml files in the nested folders and spell check all the files. Files with more than .025% spelling mistakes will be deleted.

2) (code currently in javascript) recursively find all of the .xml files in the nested folders and remove any remaining duplicate files within the movie/tv-show title folder. the result would leave only one text file per folder.

3) (code currently in python) convert all of the .xml files to .txt, while removing all of the xml tags. The result should be text files that contain only the script of the movies/tv-shows.

4) (code currently in python) use the IMDB API to add IMDB metadata to a new database of the subtitle text files. The metadata would include the title of each movie/tv-show, the country where it was made, the original language it was made in, any parental guidance information that is provided (e.g. G, 14+, 18+, R, etc. There are different rating systems for movies and tv, and they are different for each country). etc. The database should be in a format that I can create queries on my mac. I do not have access to expensive database management applications. Mac has sqllite, but I am open to any other solutions, particularly if I don't have to learn a new language.

5) optional (no code yet): simple code to make specific queries to the database in step #4 above.