Contribution ID: **39** Type: **Poster**

Ensuring Reproducibility in OntoClue –Vector-Based Document Similarity for Biomedical Literature – Using Docker

Wednesday 26 February 2025 19:40 (20 minutes)

In the realm of biomedical research, the ability to accurately assess document-to-document similarity is crucial for efficiently navigating vast amounts of literature. OntoClue is a comprehensive framework designed to evaluate and implement a variety of vector-based approaches to enhance document-to-document recommendations based on similarity, using the RELISH corpus as reference. RELISH is an expert curated biomedical literature database comprising PubMed IDs (PMIDs) and document-to-document relevance assessments categorized as "relevant," "partial," or "irrelevant." The dataset includes titles and abstracts of associated articles, which are preprocessed to remove stop words and structural words, convert text to lowercase, and tokenize the content.

OntoClue integrates various natural language processing (NLP) models, including Word2Vec, Doc2Vec, fastText, and state-of-the-art BERT-based models like SciBERT, BioBERT, and SPECTER, as well as in-house hybrid approaches that leverage annotated text through Named Entity Recognition (NER) to incorporate semantic understanding into plain text. The framework assesses document similarity using evaluation metrics that consider relevance judgment, search efficiency, and re-ranking in the context of biomedical research.

Recognizing the complexity of managing various repositories for each vector-based approach and their dependencies, OntoClue employs Docker containerization to mitigate potential conflicts and ensure seamless execution across platforms. Our methodology involves splitting the dataset into training, validation, and test sets. The training set facilitates the model training, while the validation set employs Optuna for hyperparameter optimization, using Precision@5 as the objective function. The test set is used for final evaluation, using metrics like precision@N and nDCG@N to ensure relevance and efficiency in document retrieval.

Despite rigorous testing—such as setting the random seed for model training to ensure consistent initialization, using a single worker to manage parallel processing, and configuring Optuna to run with a single job for stability—we encountered occasional inconsistencies in our results. To address this, we used Docker to standardize the Python environment and set the Python Hash seed (in the Dockerfile), which not only enhances reproducibility but also ensures that any user can replicate the results without being affected by local environmental discrepancies.

Docker-based containerization is integral to OntoClue, ensuring that code dependencies, datasets, and execution environments are fully portable and reproducible. This approach not only simplifies model training but also guarantees version control and resolves dependency conflicts, thereby enhancing ease of use and consistency in performance. Furthermore, we conduct reproducibility tests to compare results from identical runs using the same embedding models and hyperparameters. These tests require consistent hyperparameter configurations in the same order across multiple runs for the same number of iterations, and demand that Precision@N values match exactly to four decimal points. When these conditions are met, we confirm the reliability of the pipeline, reinforcing the integrity of our research.

The OntoClue Docker features a user-friendly command-line interface that allows researchers to select from 18 different embedding approaches. Upon selection, the Docker container automates essential processes, including repository cloning, dataset downloading, and class distribution selection for training. Additionally, the framework includes options for dataset integrity checks and model reproducibility tests, ensuring that the pipeline delivers consistent, reliable results.

I want to participate in the youngRSE prize

no

Primary author: RAVINDER, Rohitha (ZB MED - Information Centre for Life Sciences, Cologne, Germany)

Co-authors: Prof. REBHOLZ-SCHUHMANN, Dietrich (ZB MED - Information Centre for Life Sciences, Cologne, Germany); Dr CASTRO, Leyla Jael (ZB MED - Information Centre for Life Sciences, Cologne, Germany)

Presenter: RAVINDER, Rohitha (ZB MED - Information Centre for Life Sciences, Cologne, Germany)

Session Classification: Poster and Demo Session together with Reception

Track Classification: Research Software: AI and ML in a research context