LitCovid track Multi-label topic classification for COVID-19 literature annotation

README

Task description

The LitCovid track calls for a community effort to address automated topic annotation for COVID-19 literature. Topic annotation in LitCovid is a multi-label document classification task that assigns one or more labels to each article. There are 7 topic labels used in LitCovid: Treatment, Diagnosis, Prevention, Mechanism, Transmission, Epidemic Forecasting, and Case Report. These topics have been demonstrated to be effective for information retrieval and have also been used in many downstream applications related to COVID-19. However, annotating these topics manually has been a significant curation bottleneck. Increasing the accuracy of automated topic prediction in COVID-19-related literature would be a timely improvement beneficial to curators and researchers worldwide.

References

- Chen, Q., Allot, A. and Lu, Z., 2020. Keep up with the latest coronavirus research. Nature, 579(7798), pp.193-193.
- Chen, Q., Allot, A. and Lu, Z., 2021. LitCovid: an open database of COVID-19 literature. Nucleic Acids Research, 49(D1), pp. D1534-D1540.

Datasets

The training and development datasets contain 24,960 and 6,239 articles from LitCovid, respectively. The topics of the articles in both datasets have been manually reviewed.

The training and development datasets are provided in csv format, with the following fields retrieved from PubMed/LitCovid:

- pmid: PubMed Identifier
- journal: journal name
- title: article title
- abstract: article abstract
- keywords: author-provided keywords
- pub_type: article type, e.g., journal article
- authors: author names
- doi: Digital Object Identifier
- label: annotated topics, i.e., **the output**
    - each article can be assigned one or more labels (Treatment, Diagnosis, Prevention, Mechanism, Transmission, Epidemic Forecasting, and Case Report)
    - each label is separated by a semicolon, e.g., 'Diagnosis;Treatment' means that the article is assigned both the label Diagnosis and the label Treatment

The test data will be provided in the same format, except the topic labels should be predicted by the participants. Submissions will be evaluated using both label-based and instance-based metrics that are commonly applied for multi-label classification. Evaluation scripts will be provided. The details of the submission will be available later.

Other formats are also available via https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/, which provides other information, such as biological entity annotations.

Contact

Please contact qingyu.chen@nih.gov with the subject heading "BioCreative Track 5 LitCovid questions" if you have any questions.

Status updates and FAQs

We will also provide updates and FAQs via https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/.